

TESIS DE LA UNIVERSIDAD
DE ZARAGOZA

2021

372

Manuel Lagunas Arto

Learning Visual Appearance: Perception, Modeling and Editing.

Director/es

Dra. D^a. Masía Corcoy, Belén
Dr. D. Gutiérrez Pérez, Diego

<http://zaguan.unizar.es/collection/Tesis>

ISSN 2254-7606



Premsas de la Universidad
Universidad Zaragoza



Universidad
Zaragoza

Tesis Doctoral

LEARNING VISUAL APPEARANCE: PERCEPTION,
MODELING AND EDITING.

Autor

Manuel Lagunas Arto

Director/es

Dra. D^a. Masía Corcoy, Belén

Dr. D. Gutiérrez Pérez, Diego

UNIVERSIDAD DE ZARAGOZA
Escuela de Doctorado

2021

LEARNING VISUAL APPEARANCE:
PERCEPTION, MODELING AND EDITING

WRITTEN BY:
MANUEL LAGUNAS

SUPERVISED BY:
BELÉN MASIÁ AND DIEGO GUTIÉRREZ

TESIS DOCTORAL - INGENIERÍA DE SISTEMAS E INFORMÁTICA
GRAPHICS AND IMAGING LAB
DEPARTAMENTO DE INFORMÁTICA E INGENIERÍA DE SISTEMAS
UNIVERSIDAD DE ZARAGOZA

OCTOBER 2021

To everyone that in one way or another supported me during this journey

ABSTRACT

Visual appearance determines our understanding of an object or image, and as such it is a fundamental aspect in digital content creation. It is a general term, embracing others like material appearance, which can be defined as the visual impression we have about a material, and involves the physical interaction between light and matter, and how our visual system perceives it. However, computationally modeling the behavior of our visual system is a complex task, partially because no definite, unified theory of perception exists. Moreover, although we have developed algorithms that are able to faithfully model the interaction between light and matter, there is a disconnection between the physical parameters that those algorithms use and the perceptual parameters that the human visual system understands. This, in turn, makes manipulating such physical parameters and their interactions a cumbersome and time-consuming task, even for expert users. This thesis aims at furthering our understanding of material appearance perception, and leveraging it to improve existing algorithms for visual content generation. This is done by establishing connections between the physical parameters governing the interaction between light and matter, and high-level, intuitive parameters or attributes understood by humans. Specifically, the thesis makes contributions in three areas: proposing new computational models for measuring appearance similarity; investigating the interaction between illumination and geometry, and their effect on material appearance; and developing applications for intuitive appearance manipulation, in particular, human relighting and material appearance editing.

The first part of this thesis explores metrics to measure appearance similarity. How to accurately measure similarity between two materials, or images, is a classic problem in visual computing fields like computer graphics or computer vision. We first approach the problem of measuring appearance similarity between materials. We propose a deep learning-based framework directly trained on images paired with human judgements on material similarity, collected through user studies. In addition, we also explore the problem of icon similarity. In this case, we rely on siamese neural networks, and the subjective style and identity given by the artists play a key role in such similarity measure.

The second part takes steps towards understanding the effect of confounding factors on our perception of material appearance. Two key factors determining the final appearance of a scene are geometry and illumination. We start by investigating the joint effect of geometry and illumination in our performance recognizing materials through several crowdsourced experiments and statistical analysis. We also perform an analysis of the effect of motion blur on material perception.

In the third part, we explore intuitive applications to manipulate visual appearance. First, we address the problem of single-image full-body human relighting. We propose a new problem formulation and, based on it, design and train a generative neural network capable of relighting a scene. Last, we approach the intuitive material editing problem. We collect human judgements on the perception of high-level attributes defining material appearance, and present a generative model able to produce plausible edits just by varying such collected attribute values.

RESUMEN

La apariencia visual determina como entendemos un objeto o imagen, y, por tanto, es un aspecto fundamental en la creación de contenido digital. Es un término general, englobando otros como la apariencia de los materiales, definida como la impresión que tenemos de un material, y la cual supone una interacción física entre luz y materia, y como nuestro sistema visual es capaz de percibirla. Sin embargo, modelar computacionalmente el comportamiento de nuestro sistema visual es una tarea difícil, entre otros motivos porque no existe una teoría definitiva y unificada sobre la percepción visual humana. Además, aunque hemos desarrollado algoritmos capaces de modelar fehacientemente la interacción entre luz y materia, existe una desconexión entre los parámetros físicos que usan estos algoritmos, y los parámetros perceptuales que el sistema visual humano entiende. Esto hace que manipular estas representaciones físicas, y sus interacciones, sea una tarea tediosa y costosa, incluso para usuarios expertos. Esta tesis busca mejorar nuestra comprensión de la percepción de la apariencia de materiales y usar dicho conocimiento para mejorar los algoritmos existentes para la generación de contenido visual. Específicamente, la tesis tiene contribuciones en tres áreas: proponiendo nuevos modelos computacionales para medir la similitud de apariencia; investigando la interacción entre iluminación y geometría; y desarrollando aplicaciones intuitivas para la manipulación de apariencia, en concreto, para el re-iluminado de humanos y para editar la apariencia de materiales.

Una primera parte de la tesis explora métodos para medir la similaridad de apariencia. Ser capaces de medir cómo de similares son dos materiales, o imágenes, es un problema clásico en campos de la computación visual como visión por computador o informática gráfica. Abordamos primero el problema de similaridad en la apariencia de materiales. Proponemos un método basado en *deep learning* que combina imágenes con juicios subjetivos sobre la similitud de materiales, recogidos mediante estudios de usuario. Por otro lado, se explora el problema de la similaridad entre iconos. En este segundo caso, se hace uso de redes neuronales siamesas, y el estilo y la identidad que dan los artistas juega un papel clave en dicha medida de similaridad.

La segunda parte avanza en la comprensión de cómo los factores de confusión (*confounding factors*) afectan a nuestra percepción de la apariencia de los materiales. Dos factores de confusión claves son la geometría de los objetos y la iluminación de la escena. Comenzamos investigando el efecto de dichos factores a la hora de reconocer los materiales a través de diversos experimentos y estudios estadísticos. También investigamos el efecto del movimiento del objeto en la percepción de la apariencia de materiales.

En la tercera parte exploramos aplicaciones intuitivas para la manipulación de la apariencia visual. Primero, abordamos el problema de la re-iluminación de humanos. Proponemos una nueva formulación del problema, y basándonos en ella, se diseña y entrena un modelo basado en redes neuronales profundas para re-iluminar una escena. Por último, abordamos el problema de la edición intuitiva de materiales. Para ello, recopilamos juicios humanos sobre la percepción de diferentes atributos y presentamos un modelo, basado en redes neuronales profundas, capaz de editar materiales de forma realista simplemente variando el valor de los atributos recogidos.

MEASURABLE CONTRIBUTIONS

This thesis has led to the following results, further detailed in Section 1.5:

- Four JCR-indexed journal publications (one of them in *ACM Transactions on Graphics*) [171, 170, 172, 56].
- Two peer-reviewed conference publications [207, 173].
- Two research internships (totalling more than nine months) at *Adobe Research* in San Jose (California, USA).
- One peer-reviewed poster presented at SIGGRAPH conference [57].
- One patent application as a result of the internships at *Adobe Research* in San Jose (California, USA) [301].
- One publicly available dataset resulting from one of the journal publications.
- Four invited talks in different international institutions and workshops.

In addition to the aforementioned results, during this thesis the following tasks were also performed:

- Supervisor for one BEng final degree project and three interns.
- Reviewer for five journals and seven international conferences.

This thesis has been done within the ERC project: *CHAMELEON: Intuitive Editing of Visual Appearance from Real-world Datasets*.

ACKNOWLEDGEMENTS

There is a long list of people to say thanks throughout these four years (that felt actually like four days!). They either taught me something, pushed me to achieve my goals, or just supported me when it was needed. Thanks to all of you. I would additionally like to highlight few of them here.

I would like to start by thanking my supervisors *Belen Masia* and *Diego Gutierrez*. Thanks for all your help, guidance, and patience (especially patience) during all those years. You helped me navigating this complex (and some times frustrating) world of research, you taught me to be a better version of myself and hopefully, also, a better researcher. Thanks.

Thanks to *Elena Garces*, for introducing me into the lab, into research, being my supervisor during the Bachelor and Masters' thesis, and also during the beginning of my PhD studies. Thanks a lot also for your continuous support and encouragement during these years.

Thanks to *Xin Sun* for accepting me as your intern, for showing me the world of the industry, and for all those coffees we had while talking about how to approach the relighting project. You have been a great, friendly, and easy approachable supervisor. I could not ask for anything better.

Thanks to the members of the *Graphics & Imaging Lab* for those fruitful discussions and for giving me a hand when I asked for it. Especial thanks to *Adrian* and *Ibón* for those endless coffee conversations; *Dani*, *Ana*, *Julio*, *Marta*, for being always helpful; *Sandra* for those free uber rides in San Jose; and *Johanna*, *Neerja*, *Dario* and *Juan Raul* for bringing extra international vibes to our lab.

Thanks to all the people I met during my two internships in Adobe. *Yu*, thanks for all those lunches, ping-pong games, and discussions; *Ruben* for those 1v1 matches in Adobe's basketball court; *Jimei*, *Jianming*, and *Zhixin*, for all the help and guidance during my internships.

Thanks to past members of the *Graphics & Imaging Lab*. *Cheve* for your exceptional welcoming in San Jose; *Carlos* for those beers when you are around in Spain; *Crespo* I would write an automatic system spending 6 hours of my life to say thanks but I prefer to write it manually in 6 seconds; and *Galindo*, first of all, congrats on your new baby, then thanks a lot for all those chats, autochess games, and rides back home.

Thanks to the students I have supervised in different internships or bachelor projects. *Jorge E.*, *Jorge C.*, *Alejandro*, *Daniel*. Thanks a lot. I hope you learned something while I supervised you. I definitely learnt something new with each of you.

Thanks to my friends that helped me to disconnect from research when needed. *Andres* for always being there; *Mostro* and *Iñigo* for those talks about internships, robotics, or just for grabbing a coffee with me; my friends from the basketball team (especially *Arregui*, *Patxi*, *Moli*, and *Ruiz*); from the school and high-school (*Pablo*, *Dani*, *Raz*, *Pedro*, *Dani*); university colleagues (*Rueda*, *Sergio*, *Toni*, *Victor*, *Hector*); and probably I am missing someone. Thanks all of you.

Last, but not least, I would like to give an enormously big thanks to my family, for always being there, for always being supportive, for your encouragement, for your unconditional trust, for your love and your care. большое спасибо.

CONTENTS

I INTRODUCTION AND OVERVIEW	
1 INTRODUCTION	3
1.1 Measuring Appearance Similarity	5
1.2 Confounding Factors in Material Perception	6
1.3 Intuitive Applications for Appearance Editing	7
1.4 Goal and Overview	9
1.5 Contributions and Measurable Results	9
II A REVIEW OF THE LITERATURE	
2 PREVIOUS WORK	15
2.1 Visual Appearance Perception	15
2.2 Visual Appearance Modeling	17
2.3 Visual Appearance Editing	19
III MEASURING APPEARANCE SIMILARITY	
3 MATERIAL APPEARANCE SIMILARITY	25
3.1 Introduction	25
3.2 Materials Dataset	27
3.2.1 Why a New Materials Dataset?	27
3.2.2 Description of the Dataset	27
3.3 Collecting Appearance Similarity Information	28
3.3.1 Adaptive Sampling Scheme for the User Studies	30
3.4 Learning Perceived Similarity	31
3.4.1 Loss Function	31
3.4.2 Training Details	33
3.5 Evaluation	33
3.5.1 Comparison with Other Metrics	33
3.5.2 Ablation Study	34
3.5.3 Alternative Networks	35
3.5.4 Results by Category	35
3.5.5 Failure Cases	36
3.6 Applications	37
3.6.1 Material Suggestions	37
3.6.2 Visualizing Material Datasets	37
3.6.3 Database Clustering	38
3.6.4 Database Summarization	39
3.6.5 Gamut Mapping	39
3.7 Objective and Subjective Measures	40
3.7.1 Analysis	41
3.8 Discussion	42
4 ICONS APPEARANCE SIMILARITY	45
4.1 Introduction	45
4.2 Problem Definition	47
4.2.1 Overview	47
4.3 Collecting Data	48
4.4 Modeling Visual Appearance of Icons	49
4.4.1 The Loss Function	50
4.4.2 Training the Model	51
4.5 Evaluation	52

4.5.1	Other Architectures	53
4.5.2	Comparing Against Previous Work	54
4.6	Results and Applications	54
4.7	Conclusion and Future Work	57
IV CONFOUNDING FACTORS IN MATERIAL PERCEPTION		
5	GEOMETRY AND ILLUMINATION IN MATERIAL PERCEPTION	61
5.1	Introduction	61
5.2	Methods	63
5.2.1	Stimuli	63
5.2.2	Participants	65
5.2.3	Procedure	66
5.3	Results	67
5.3.1	Analysis of User Performance and Time	67
5.3.2	High-level Factors Driving Material Recognition	71
5.4	Discussion	75
5.5	Conclusions	79
6	MOTION IN MATERIAL PERCEPTION	81
6.1	Introduction	81
6.2	Stimuli Creation	83
6.3	First Experiment: Rating Material Attributes	84
6.3.1	Analysis	85
6.4	Second Experiment: Brightness Map Construction	87
6.4.1	Influence of brightness	87
6.4.2	Analysis	89
6.5	Discussion and Future Work	90
6.5.1	Conclusion	90
6.5.2	Future work	91
V INTUITIVE APPLICATIONS FOR APPEARANCE EDITING		
7	FULL-BODY HUMAN RELIGHTING	95
7.1	Introduction	95
7.2	Background	96
7.3	Our Image Reconstruction Formulation	98
7.3.1	Problem Formulation	98
7.4	Dataset	99
7.5	Our Model	100
7.5.1	Model Architecture	100
7.5.2	Training	102
7.5.3	Loss Functions	102
7.6	Results	103
7.6.1	Ablation Studies	106
7.7	Discussion	106
8	INTUITIVE MATERIAL EDITING	111
8.1	Introduction	111
8.2	Our Framework	112
8.2.1	Goal and Overview	112
8.2.2	Model Architecture	113
8.2.3	Loss Functions and Training Scheme	114
8.3	Training Dataset	116
8.4	Results and Evaluation	118
8.4.1	Ablation Studies	120
8.4.2	Consistency of the Edits	121

8.4.3	User Study	121
8.5	Discussion and Limitations	123
VI CONCLUSION		
9	CONCLUSION AND FUTURE WORK	129
VII APPENDIX		
A	MATERIAL APPEARANCE SIMILARITY: ADDITIONAL RESULTS	135
A.1	Additional Loss Terms	135
A.1.1	Cross-entropy Term \mathcal{L}_{CE}	135
A.1.2	Batch-mining Triplet Loss Term \mathcal{L}_{BTL}	135
A.2	Queries and Agreement with Humans	135
A.3	Material Suggestion Examples	136
B	GEOMETRY AND ILLUMINATION IN MATERIAL PERCEPTION: ADDITIONAL DETAILS	141
B.1	Additional Details on Image Statistics	141
C	MOTION IN MATERIAL PERCEPTION: STIMULI AND ADDI- TIONAL RESULTS	143
C.1	Stimuli Used in the First Experiment	143
C.2	Stimuli Used in the Second Experiment	148
C.3	Attribute Plots for the First Experiment	150
C.4	Tables for the Statistical Tests	150
D	FULL-BODY HUMAN RELIGHTING: ADDITIONAL RESULTS	155
D.1	Additional Results	155
E	INTUITIVE MATERIAL EDITING: ADDITIONAL DETAILS AND RESULTS	157
E.1	Additional Details on the Framework	157
E.2	Additional Details on the Normal Prediction	158
E.2.1	Architecture	159
E.2.2	Losses	159
E.2.3	Training	160
E.3	Additional Results	160
	BIBLIOGRAPHY	163

LIST OF FIGURES

Figure 1.1	Examples of different visual appearances that are found in nature.	3
Figure 1.2	Scenes where we compare the visual appearance. . .	5
Figure 1.3	Appearances obtained by varying confounding factors.	7
Figure 1.4	Example of an application for intuitive editing of visual appearance.	8
Figure 3.1	Automatically generated depictions of the same scene where materials' similarity varies in a controlled manner.	26
Figure 3.2	All six illuminations used in the dataset.	27
Figure 3.3	Sample images of all 15 scenes in the dataset.	29
Figure 3.4	Sample stimuli for our appearance similarity collection.	29
Figure 3.5	Scheme of the training process.	33
Figure 3.6	Examples from our dataset queries where our model agrees with the majority response.	34
Figure 3.7	Error for the different metrics derived from human responses and Representative example of the two most similar materials to a given reference.	36
Figure 3.8	Examples where humans' majority disagree with our metric	36
Figure 3.9	Examples of material suggestions using our model. .	38
Figure 3.10	Additional material suggestions results.	39
Figure 3.11	2D visualization using our feature vectors.	39
Figure 3.12	Material suggestions with our perceptual database clustering.	40
Figure 3.13	Representative clusters on the Extended MERL dataset.	40
Figure 3.14	Database summarization example.	41
Figure 3.15	Gamut mapping using our similarity metric.	41
Figure 3.16	RDMs of each network organized by reflectance . . .	42
Figure 3.17	Interplay of confounding factors showcased in different scenes.	43
Figure 3.18	Results using our metric on heterogeneous materials.	43
Figure 4.1	Example of six different collections of icons in the dataset.	46
Figure 4.2	Examples of similarity between icons.	47
Figure 4.3	Overview of our complete framework to measure icons similarity.	48
Figure 4.4	Architecture proposed to measure icons similarity. . .	50
Figure 4.5	Examples of the triplets sampled during training. . .	51
Figure 4.6	Model performance while varying the number of layers.	53
Figure 4.7	t-SNE visualization of the feature space created by our model.	55
Figure 4.8	Comparison of the shape kernels generated by our model.	56
Figure 4.9	Most similar icons given a reference returned by our method.	56
Figure 4.10	Icons set proposal for semantic keywords using our method.	57
Figure 4.11	Screenshot of the user study.	58

Figure 5.1	Two silver spheres under two different illuminations.	62
Figure 5.2	Two silver objects, with different shape, under the same illumination.	62
Figure 5.3	Graphical user interface of the online behavioral experiments.	63
Figure 5.4	Examples of the stimuli in each different online behavioral experiment.	64
Figure 5.5	Illuminations depicted in the online behavioral experiments together with their frequency spectrum.	65
Figure 5.6	Candidate illumination employed in the online behavioral experiments together with its frequency spectrum.	66
Figure 5.7	High-frequency content measure (HFC) for each illumination.	66
Figure 5.8	Accuracy in each experiment by geometry.	68
Figure 5.9	Top-5 accuracy by reference illumination.	69
Figure 5.10	Top-5 accuracy by reference and candidate illumination and geometry.	70
Figure 5.11	Average time spent in each trial.	71
Figure 5.12	Visualization of user answers for each experiment using the t-STE algorithm.	72
Figure 5.13	Two-dimensional visualization using the UMAP algorithm on the ResNet model.	74
Figure 5.14	Normalized pairwise similarity for each online behavioral experiment and the deep neural network trained for material classification.	74
Figure 5.15	Top-5 accuracy obtained by participants in the original experiment according to the displayed image size.	76
Figure 5.16	Top-5 accuracy for each reference illumination when varying the candidate illumination.	77
Figure 5.17	Example of a convolution and its effect in the resulting frequencies.	78
Figure 5.18	Users' performance, in terms of top-5 accuracy, for material recognition tasks taking into account the reflectance of the materials.	79
Figure 6.1	Influence of motion in material attributes.	82
Figure 6.2	Influence of motion in perceived luminance.	82
Figure 6.3	Workflow employed to generate the stimuli.	83
Figure 6.4	Subset of the stimuli used in the experiments.	84
Figure 6.5	Average ratings for each perceptual attribute.	88
Figure 6.6	Brightness and motion maps.	89
Figure 7.1	Relighted results given a single image as input for different illumination maps.	96
Figure 7.2	Comparison between the data generated with our framework and that of the previous work.	98
Figure 7.3	Two examples of all the scenes generated in our dataset.	100
Figure 7.4	Architecture of our UNet-like convolutional neural network.	101
Figure 7.5	Workflow of each component of our model.	102
Figure 7.6	Example result of our model for a synthetic image.	103
Figure 7.7	Comparison between our model and previous work in two examples of the validation dataset.	104
Figure 7.8	Example results of our model on real photographs.	105

Figure 7.9	Comparison of image reconstructions between our method and previous work.	107
Figure 7.10	Reconstruction results obtained on the different ablation experiments.	107
Figure 7.11	Example of the limitations of our model.	108
Figure 7.12	Relighting results for two different illuminations (ennis, and pisa) and five different input images.	109
Figure 8.1	Single photograph material editing using our framework and high-level perceptual attributes.	112
Figure 8.2	Overview of the different components of our material editing framework.	113
Figure 8.3	Representative samples of the image dataset used for training.	117
Figure 8.4	Different viewpoints in the user study samples.	117
Figure 8.5	Screenshot of the perceptual study.	118
Figure 8.6	Control images used in the validation study.	119
Figure 8.7	Representative images of our synthetic evaluation dataset.	119
Figure 8.8	Representative examples of our real images evaluation dataset.	120
Figure 8.9	Editing results by varying the perceptual attributes Metallic and Glossy.	121
Figure 8.10	Ablation studies where we trained and tested out each of the individual components of our framework.	121
Figure 8.11	Example results illustrating the consistency of our editing framework.	122
Figure 8.12	Input images and edited stimuli used in our user-study.	123
Figure 8.13	Answers collected in our validation study for both attribute Metallic and Glossy and for the two sets of images.	124
Figure 8.14	Limitations of our editing framework.	125
Figure A.1	Queries to our method (I).	137
Figure A.2	Queries to our method (II).	138
Figure A.3	Queries with our measure (I).	139
Figure A.4	Queries with our measure (II).	140
Figure C.1	Stimuli rendered with a glass material.	143
Figure C.2	Stimuli rendered with a metallic material.	144
Figure C.3	Stimuli rendered with a paint material	145
Figure C.4	Stimuli rendered with a plastic material.	146
Figure C.5	Stimuli rendered with a rubber material.	147
Figure C.6	Stimuli rendered with a stone material.	148
Figure C.7	Stimuli rendered with a plastic material.	149
Figure C.8	Stimuli rendered with a rubber material.	150
Figure C.9	Participants' ratings for each attribute (I).	151
Figure C.10	Participants' ratings for each attribute (II).	154
Figure D.1	Additional relighted results on photographs for two different illuminations.	156
Figure E.1	Editing results varying the metallic attribute.	161
Figure E.2	Editing results varying the glossy attribute.	162

LIST OF TABLES

Table 3.1	Accuracy and perplexity of our model compared against human performance.	35
Table 3.2	Accuracy and perplexity for other alternative loss functions.	37
Table 3.3	Metrics per material category.	38
Table 4.1	Comparison of the precision and perplexity of different models and methods.	54
Table 6.1	Results from Friedman rank test.	86
Table 6.2	Results from the Nemenyi post hoc test.	87
Table 7.1	Quantitative results of our model for synthetic images and real photographs	104
Table 7.2	Quantitative comparisons using real photographs against previous work.	106
Table 8.1	Pearson correlation coefficients between the expected attribute and participants' answers.	123
Table C.1	All p -values for the Chi-squared test and Nemenyi post hoc test for each attribute and material category	153

Part I

INTRODUCTION AND OVERVIEW

INTRODUCTION

"Somewhere, something incredible is waiting to be known."
Carl Sagan, 1934-1996

Visual data plays a key role in the way we understand and perceive the world around us: As humans, about 90% of our input information comes from sight. Just in a glimpse, we can gather data about the environment around us, acquiring information on the illumination, or understanding the visual appearance of objects or images. Visual appearance refers to the particular look of an object or an image, but also involves how our visual system integrates and processes such information to give us an understanding of it, whether the information comes from a real-world scene that we observe, from a non-realistic illustration, or from a realistic rendering depicted in an image [127]. Visual appearance is therefore a general term, embracing others like material appearance that can be defined as "the visual impression we have of a material" [65], involving a physical interaction between light and matter, and how our visual system perceives it.



Figure 1.1: Examples of different appearances that are found in nature. The particular look of a real-world scene that we observe is built by a complex multidimensional interaction capable of creating effects of astonishing richness such as in pearls (left), northern lights (middle), or the wings of a butterfly (right).

This physical interaction between light and matter is one of the aspects determining the visual impression of a material, and is a consequence of the way the illumination is reflected and transmitted, but it also involves several other aspects like the geometry of an object, or whether it is in motion [65]. Such multidimensional physical interaction allows creating appearances of astonishing richness (e.g., pearls, northern lights, or the wings of a butterfly; see Figure 1.1), which have, throughout the years, inspired painters or photographers and, more recently, computer graphics professionals. Our society is increasingly relying on computer-generated imagery for everyday tasks; consequently, visual appearance and material appearance are a fundamental aspect, not only of how humans understand the world and communicate concepts or ideas, but also of how we create and develop digital content.

On the other hand, how our visual system integrates and processes such multidimensional physical interaction to give us the final impression about a material is not yet fully understood. Under typical viewing conditions, humans can effortlessly recognize materials and infer their key physical properties at a glance. Indeed, just by briefly looking at them we can usually

tell whether they would feel soft, cold, or if they would be wet. However, untangling the processes that happen in our visual system, and finding a direct relationship between our subjective impression and the parameters that govern the underlying physical interaction, remains an open challenge without a definite solution. This, subsequently, difficults the creation of computational models (or algorithms) of appearance that take into account human perception.

Being able to compare how similar two objects or images are, or being able to infer an object's properties from an image, are valuable and important abilities that we, as humans, perform every day for a wide range of common tasks. They are, however, hard well-known open problems in fields like computer graphics and vision. This thesis, therefore, aims at improving computational models for measuring appearance similarity by taking into account the subjective nature of human perception, exploring how the interaction between physical parameters affects material appearance perception, and easing the use of applications for appearance manipulation by developing more intuitive frameworks.

One of the first open problems this thesis looks into is appearance similarity. In our daily lives, we are constantly comparing the different visual appearances, either to understand which piece of fruit we want to buy in the supermarket, or to select the style of the digital illustration we will use to decorate the background of our laptop (see Figure 1.2). Comparing and measuring the similarity between appearances is a task that humans perform effortlessly. However, it is a challenging problem, where the influence of confounding factors, like geometry or illumination, or the subjective nature of human perception (which has often been ignored by traditional methods) play a key role. Therefore, to obtain more precise methods that measure appearance similarity, or help to obtain plausible solutions in cases where a unique mathematical model does not exist, human perception should be taken into account. This thesis (Part III) looks at both material appearance similarity and image (vector graphics icons) similarity.

Secondly, this thesis aims at taking steps towards a better understanding of material appearance perception by investigating the interaction between physical parameters, their effect on material appearance, and how they influence our perception of it (Part IV). Material appearance is defined by the complex interaction between light and matter giving objects a particular look, and also by how our visual system perceives and understands it. The physical parameters describing such interaction in a scene are heterogeneous, ranging from low-level parameters defining the appearance of a material (e.g., the amount of dispersed microscopic interference pigments within a dielectric resin used to represent pearlescent appearances), to high-level parameters like the illumination of a scene, the geometry of an object, or whether it is in motion (see Figure 1.3). Our experience suggests that humans are able to correctly recognize a wide variety of materials: If we observe the same static object made of chrome in sunlight and moonlight, we would see completely different depictions of the same object, yet our visual system is capable of telling that it is actually made of chrome by just seeing a blurred reflection of the environment with a gray-like tint, a few smudges, and maybe some scratches. How our visual system processes this information, abstracts itself from the effect of confounding factors on the appearance, and is then able to recognize the chrome object, remains not fully understood. Investigating the influence of confounding factors as well as their relationship



Figure 1.2: Examples of a scene where we would compare the visual appearance of materials, effortlessly, either to pick the freshest fruits, or the best looking flowers. Measuring the similarity between appearances is a long-standing problem in computer graphics and vision for which no definite solution exists yet.

with the physical information encoded in the data is a key aspect towards understanding how humans interpret material appearance.

The last part of this thesis explores the development of intuitive applications for appearance editing (Part V). Being able to intuitively edit material appearance is a core area in the field of computer graphics (see Figure 1.4). It is not only a fundamental aspect of digital content creation, but an inherent part of our lives as well: We rely on computer-generated imagery in many daily tasks, while many industrial processes depend on correct appearance simulations to convey the desired visual information. Therefore, the development of proper applications or algorithms for the manipulation of appearance is key for current methods and algorithms, but can also have an impact on emerging fields such as computational materials or fabrication. Nowadays, we have the storage to acquire accurate real-world information, and the power and technical knowledge to accurately represent its appearance [70, 211]. However, computational representations of appearance rely on complex physical parameters and models that are sub-optimal with respect to what humans can easily understand and process: The data is machine-friendly, but not human-friendly. Despite its current importance, the applications to edit the visual appearance remain unintuitive and inefficient, where expert knowledge is often required to understand their parameters and the interactions between them. Finding (and developing) connections between the complex physical parameters used by current applications, and the perceptual parameters that humans employ to understand material appearance, could open up a new space with great potential for intuitive applications to edit appearance.

In summary, this thesis presents contributions on three relevant areas of visual appearance: measuring appearance similarity taking subjective judgements into account (Part III), analyzing the influence of confounding factors in our perception of material appearance (Part IV), and proposing intuitive applications for relighting or material editing (Part V).

1.1 MEASURING APPEARANCE SIMILARITY

The first part of the thesis deals with establishing metrics to measure appearance similarity. This is a long-standing problem in fields within visual computing, like computer vision or computer graphics. In particular, this thesis focuses on measuring material appearance and icons' appearance similarity. Since the subjective nature of perception should be taken into account, these concepts differ from original notion of image similarity, which

can be defined as the difference between intensity patterns in two images. Material appearance similarity, instead, defines how similar two materials are, and it could be measured directly from images or from their physical representation (the BRDF); however, the latter would not take into account the influence of confounding factors. Icons' similarity, on the other hand, could be influenced by the different styles or visual identities given by the artists during the creative process. Traditional methods to measure similarity have relied on hand-crafted feature vectors [95], metrics directly working in BRDF space [223, 91], or directly in image space [224]. However, those do not take into account the influence of human perception, or the potential effect of confounding factors in the final appearance.

We propose computing visual appearance similarity in terms that correlate with the human notion of similarity. Unfortunately, mapping raw data or optical parameters to perceptual parameters is severely ill-defined, and therefore a challenging process. We leverage the large availability of visual appearance data representing the physical parameters that govern the interaction between light and matter, online tools that allow running large-scale user studies to capture the subjective impression defining appearance, and deep learning models that are able to identify the features that better correlate with the human notion of appearance similarity. Part III of this thesis explores first how to combine subjective human judgements on material appearance with deep learning models, and develop a similarity measure from material appearance working directly on image space. Then, we investigate the problem of measuring similarity between icons by leveraging online datasets and siamese neural networks to learn our measure.

First, we tackle the complex problem of creating a robust material appearance similarity measure that correlates with human judgements. We design and render a dataset of images representing a wide variety of real-world materials, shapes, and illuminations. We use such dataset to launch crowdsourced experiments from which we gather perceptual data on material appearance similarity, in the form of subjective judgements. We leverage this data, together with the corresponding images in the rendered dataset, to learn a deep learning-based measure capable of correlating with the human notion of material similarity. We employ a set of custom loss functions suitable for the aforementioned problem, and validate our model against previous work. Last, we propose several applications enabled by our metric, such as: material suggestions, database visualization, clustering and summarization, and gamut mapping.

We also present a method to measure similarity in the non-photorealistic domain of iconography, where the style or visual identity given by the artists play a key role in the message they convey. First, we collect a database of icons from online sources where each icon image is paired together with semantic labels that are given by artists. We train a siamese neural network to measure icons' appearance similarity according to the given semantic labels. We further validate the model by launching a crowdsourced experiment where we collect groundtruth subjective judgements from humans. Last, we use our method to develop useful applications such as icon retrieval by similarity, or icon set proposals.

1.2 CONFOUNDING FACTORS IN MATERIAL PERCEPTION

When we observe a photograph of a chrome sphere, our visual system is capable of gathering information from several variables such as the geometry

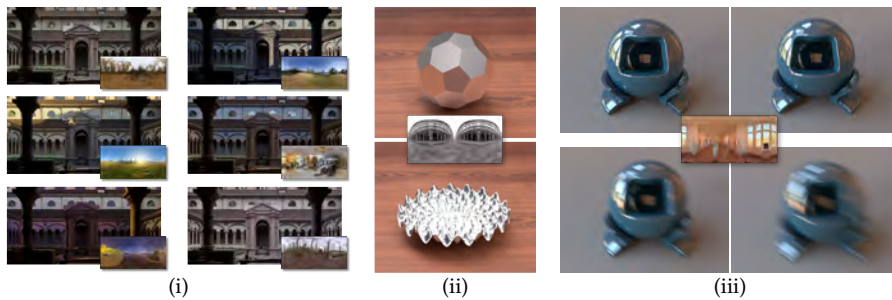


Figure 1.3: Different appearances obtained by varying confounding factors: (i) The same scene under different illuminations, (ii) two different geometries with the same material and illumination [172], (iii) renderings of an object with the same geometry, material, and illumination, where the object has different degrees of motion. In all three examples, the insets show the illumination used to render the scene.

or the illumination, and tell that the material of such sphere is actually chrome. Something similar happens if we modify the object’s geometry to be a dragon statue, if we move it to a place with a different illumination, or if we add a certain degree of motion to it. Our visual system, even though the image that has arrived to it completely differs from the original chrome sphere, is capable of abstracting itself from the influence of confounding factors in the final look, and inferring that these objects are all made of the same chrome material. Identifying the physical causes of the patterns and features that our visual system creates to understand what we observe is widely considered to be the central challenge of vision [86, 89]. In Part IV of this thesis, we investigate how confounding factors — such as geometry, illumination, or motion (see Figure 1.3) — affect human perception. While previous work has studied the influence of illumination [86, 128] and geometry [320, 118], we first propose a systematic, in-depth analysis of their joint influence. Then, we additionally study the influence of motion (in isolation), and how it affects a set of attributes describing material appearance.

First, we perform a comprehensive analysis of how the interplay between geometry, illumination, and their spatial frequencies affects human performance on material recognition tasks. From this analysis we observe that simple image statistics do not correlate with human perception [90, 10]. Therefore, we perform a high-level comparison between human answers and highly non-linear statistics such as deep neural networks [295, 89], finding preliminary evidence that these highly non-linear models and humans may use similar high-level factors for material recognition tasks.

We also look at the influence of motion in our perception of material appearance, and specifically, at the influence of motion blur in the presence of linear object movement. We start by creating a database with a diverse set of stimuli covering a wide range of appearances and under different degrees of motion. We use such database to develop two experiments where participants rate different material attributes. Our analysis shows that certain attributes undergo a significant change, varying appearance perception under motion.

1.3 INTUITIVE APPLICATIONS FOR APPEARANCE EDITING

As mentioned before, we nowadays have enough power, storage, and tools to, in many cases, precisely acquire and simulate the appearance of the real



Figure 1.4: Example of intuitive editing of material appearance. The method takes as input an image with an opaque object, and it is capable of changing its appearance to a glass-looking one where lighting effects such as inter-reflections, and even caustics are computed automatically (image from the work of Gutierrez et al. [115]).

world. This opens up extra features, such as computationally reproducing objects' appearance faithfully, or using deep learning algorithms trained with computer-generated datasets. However, the data-driven paradigm shift we are witnessing also leads to limitations: There is a disconnect between the physical parameters of the captured data employed by the algorithms and human-understandable perceptual parameters, making the use of applications to model and edit appearance cumbersome and often restricted to professionals. In addition, the data captured from the real world are heterogeneous, vary between acquisition methods, and are very high-dimensional, aggravating the intuitiveness of such applications which, in turn, yields a more restricted creative process for professionals. Finding relationships between the physical parameters used by the algorithms and the perceptual parameters understandable by humans can help developing more intuitive applications where no expert knowledge is required beforehand. Traditionally, intuitive applications involved finding direct mappings between physical estimates and perceptual traits [142, 50]; or deriving low-dimensional spaces that are easy to explore and navigate in human-understandable terms [276, 246]. The emergence of deep learning has opened up additional possibilities, where we can accurately model the physical properties of large scale datasets [295], derive low-dimensional embeddings [350], or directly propose intuitive applications [274, 148]. In Part V of this thesis, we approach two well-known problems: relighting and material appearance editing. We present two frameworks that require only human-understandable parameters as the input, and are capable of producing plausible solutions for both problems. We rely on a combination of large-scale datasets representing the physical parameters, user studies capable of capturing the subjective nature of human perception, and deep learning methods capable of finding the features that connect them.

First, we tackle the relighting problem. We propose a framework to intuitively relight images with full-body humans in them, where the user just gives a target illumination together with the input image. We leverage pre-computed radiance transfer (PRT) and spherical harmonics (SH) to propose our image reconstruction formulation, which explicitly models diffuse and specular reflection. We gather and generate a dataset of full-body humans with our formulation and train a deep neural network on this data capable of intuitively relighting a photograph with a human in it.

We additionally propose a method to intuitively edit material appearance in RGB images. We rely on a synthetic dataset to collect ratings on high-level attributes such as *glossiness* or *metallic*. Using the dataset together with the human ratings, we train a generative neural network capable of faithfully editing the appearance of the material in an image just by providing a value of an attribute. We conduct an additional crowdsourced experiment to validate our method and demonstrate its applicability also in real photographs.

1.4 GOAL AND OVERVIEW

This thesis is divided in three main parts, one for each of the three topics described in the introduction.

- Part III discusses the problem of appearance similarity. In Chapter 3, we propose a deep learning framework to measure material appearance similarity. We learn our model directly from human judgements by using a custom loss function. We validate our model, show that it outperforms previous work, and how it correlates with perceptual judgements. In Chapter 4, we propose a framework to measure icons appearance similarity learned directly from noisy labels that is later tested against human judgements.
- Part IV studies the influence of confounding factors in our perception of material appearance. In Chapter 5, we propose a comprehensive study to understand the joint role of geometry and illumination in our performance recognizing materials. We do an analysis on the frequency domain, explore the correlation with simple image statistics, and find a possible connection between high-level factors and low-dimensional spaces created by deep neural networks. In Chapter 6, we continue exploring the influence of confounding factors by launching several user studies where we analyze how motion blur affects our perception of certain material attributes.
- Part V proposes intuitive applications to manipulate the appearance of a scene just from single images. Chapter 7 presents a framework for full-body human relighting. The framework relies on a synthetic dataset and a deep neural network. We show that our improved image formulation leads to better relighted reconstructions. In Chapter 8, we present a framework for intuitive material editing using just a single image, and a value representing a change in appearance for a material attribute (e.g., *glossiness* or *metallic*). We rely on several user studies to collect human judgements regarding such material attributes, and we use this data to train a generative adversarial network capable of performing material editing.

While the author of this thesis is the leading author in many of the presented works, they have been done in collaboration with different colleagues. To favor readability, the work described is contextualized at the beginning of each chapter and, when needed, the contribution of the author of this thesis is explicitly described.

1.5 CONTRIBUTIONS AND MEASURABLE RESULTS

PUBLICATIONS In the following we state the publications which support the contributions of this thesis. Most of the work presented in this thesis has

been already published. In particular, in four journals (including one paper in ACM Transactions on Graphics, and also presented at SIGGRAPH) indexed in the Journal Citations Report (JCR), and two peer-reviewed international conferences:

- Measuring Appearance Similarity (Part III, Chapters 3 and 4):
 - The main work on material appearance similarity (Chapter 3) was accepted and presented at SIGGRAPH 2019, and published in ACM Transactions on Graphics [171]. This journal has an impact factor of 5.084, and its position in the JCR index is 8th out of 108 (Q1) in the category Computer Science, Software Engineering (data from 2019).
 - Further analysis on the role of objective and subjective measures in material similarity learning (Section 3.7) has been published as a peer-reviewed poster in SIGGRAPH 2020 [57].
 - One paper performing a preliminary study on deep learning for appearance similarity in icons (Chapter 4) has been published in Multimedia Tools and Applications in 2019 [170]. This journal has an impact factor of 2.313, and its position in the JCR index is 34th out of 108 (Q2) in the category Computer Science, Software Engineering (data from 2019).
- Confounding Factors in Material Perception (Part IV, Chapters 5 and 6):
 - The main work proposes a thorough and rigorous study on the joint role of geometry and illumination on material recognition (Chapter 5). It was accepted to the Journal of Vision (JoV) in 2021 [172]. This journal has an impact factor of 2.240, and its position in the JCR index is 39th out of 62 (Q3) in the category Ophthalmology (data from 2020).
 - An additional study on the effects of motion in our perception of material attributes (Chapter 6) was published and presented in the Symposium on Applied Perception (SAP) in 2019 [207].
- Intuitive Applications for Appearance Editing (Part V, Chapters 7 and 8):
 - The main work proposes a deep-learning based system for full-body human relighting (Chapter 7). It was presented at the Eurographics Symposium on Rendering (EGSR) in 2021 [173]. This work was done during the internships at *Adobe Research*.
 - One paper where we devise an in-the-wild framework for intuitive material editing with high-level attributes (Chapter 8). This work was referred to the Computer Graphics Forum (CGF) journal [56]. This journal has an impact factor of 2.078, and its position in the JCR index is 49th out of 108 (Q2) in the category Computer Science, Software Engineering (data from 2020).

RESEARCH INTERNSHIPS Two research internships, totaling more than nine months, were carried out during this PhD:

- June 2019 – October 2019 (four months): Research Intern at *Adobe Research*, San Jose (California). Supervised by Dr. Xin Sun. As a result of this internship, a patent application was filed [301].

- June 2020 – December 2020 (five months): Research intern at *Adobe Research*, San Jose (California). Supervised by Dr. Xin Sun. This internship led to the publication of one of the articles of this thesis [173] (Chapter 7).

SUPERVISED STUDENTS During the development of this thesis the author has supervised the following students:

- 2020: Jorge Esteban. BEng final degree project: *Analysis of a deep learning-based framework for automatic eye movement classification*. Grade: 8.7/10. Title in Spanish: *Estudio de un sistema de clasificación automática de movimientos oculares basado en técnicas de aprendizaje profundo*
- 2020: Daniel Subias. Internship: *Exploration of NeRF-Based Methods for Material Editing*.
- 2020: Jorge Condor. Internship: *An In-the-wild Surface Normal Estimation Framework*.
- 2018: Alejandro Lanaspá. Internship: *Style Transfer for Texture Synthesis*.

RESEARCH PROJECTS This thesis has been carried out within the following research project:

- *CHAMELEON: Intuitive editing of visual appearance from real-world datasets*. European Research Council (ERC). Grant agreement No 682080. PI: Diego Gutierrez.

Part II

A REVIEW OF THE LITERATURE

This chapter performs a review of the literature concerning visual appearance and material appearance. We start by reviewing the literature regarding visual appearance perception (Section 2.1). We present previous work studying the perception of high-level attributes describing material appearance, discuss if our visual system works as an inverse optics or as a matching statistics approach, introduce the problem of material recognition, and review the influence of confounding factors in our perception of material appearance. Then, we move onto how to model visual appearance and material appearance (Section 2.2). We introduce datasets capturing real-world material appearances, review low-dimensional spaces for material appearance modeling, discuss the appearance of deep learning frameworks, present different approaches to model icons' appearance, and introduce methods to measure style and shape similarity. In the last part, we review the computational methods employed to edit visual appearance (Section 2.3). We address the problem of material appearance editing, discuss image-based approaches, introduce inverse rendering and image-based rendering methods for appearance manipulation, and, finally, present the problem of human relighting.

2.1 VISUAL APPEARANCE PERCEPTION

The exact way in which our visual system infers and understands the appearances of the image that arrives to it is yet to be discovered [86, 85, 307, 3]. While we have learned how to faithfully represent materials, through their physical interaction between light and matter [249], material perception is a cognitive process [241] whose underlying intricacies are not fully understood yet [9, 88]. There have been many works aiming to understand the perceptual properties of materials [9, 88, 82, 205]; a comprehensive review can be found in the work of Thompson and colleagues [307]. Finding a direct mapping between perceptual estimates and physical parameters is a hard task, involving many dimensions that are not necessarily correlated. This section reviews the literature regarding human perception of visual appearance and material appearance.

MATERIAL PROPERTIES To understand the perceptual properties of a given material is a hard task given the complexity of the information that arrives at our visual system. A large body of work has been devoted to understand the visual cues that we use to infer isolated appearance properties such as glossiness [40, 331, 182, 75, 305, 319], translucency [106, 335, 104], softness [268, 39], or color [29]; while others aimed at understanding the perceptual cues used by artists when depicting materials in realistic paintings [62, 58]. We, humans, are able to infer the physical properties of objects without the need of touch [90, 87, 205, 1, 304], just by briefly looking at them [280, 281].

MATERIAL PERCEPTION: INVERSE OPTICS OR MATCHING STATISTICS One of the goals in vision science research is to untangle the processes that happen on our visual system in order to understand their roles and know what information they carry. There is an ongoing discussion on whether our visual system is solving an inverse optics problem [150, 250], or if it matches the statistics of the input to our visual system in order to understand the world that surrounds us [2, 218, 308]. Recent studies dismiss

the inverse optics approach and claim that it is unlikely that our brain estimates the parameters of the reflectance of a surface, when, for instance, we want to measure glossiness [82, 100]. Instead, they suggest that our visual system joins low- and mid-level statistics to make judgements about surface properties [4]. On this hypothesis, Motoyoshi et al. [218] suggest that the human visual system could be using a measure of histogram symmetry to distinguish glossy surfaces. Other works have explored image statistics in the frequency domain [119, 266], for instance, to characterize material properties [103], or to discriminate textures [146, 265]. However, it is argued whether our visual system actually derives any aspects of material perception from simple statistics [10, 159, 235]. Recent work by Fleming and Storrs [89] proposes that to infer the properties of the scene, our visual system is doing an efficient and accurate encoding of the proximal stimulus (image input to our visual system). Thus, highly non-linear models, such as deep neural networks, may better explain human perception.

MATERIAL RECOGNITION Recognizing materials and inferring their key features by sight is invaluable for many tasks. Our experience suggests that humans are able to correctly predict a wide variety of rough material categories like textiles, stones or metals [82, 99, 84, 186, 3]. Humans are also capable of identifying the materials in a photograph just by briefly looking at them [280, 281] or of inferring their physical properties without the need to touch them [90, 87, 205, 1, 209]. This ability is built from experience, by actually confirming visual impressions with other senses. However, even though a lot of work is devoted to understanding how to recognize materials, how to quantify the individual contribution of each visual and non-visual cue in our perception of materials remains an unsolved open problem [9, 82, 88, 205]. For a comprehensive study on early material recognition systems and latest advances, we refer to the reader to the work of Fleming [83]. In this thesis (Chapter 5), we give a step towards understanding how humans recognize materials by studying the joint influence of geometry and illumination in our perception. We launch several user studies proposing a comprehensive analysis of the results, explore how image statistics correlate with human answers, and find a preliminary relationship between high-level factors used by humans and low-dimensional spaces created by a deep neural network.

INFLUENCE OF CONFOUNDING FACTORS Material perception is a complex process that involves a large number of distinct dimensions [279, 230] that, sometimes, are impossible to physically measure [135]. It depends not only on the intrinsic properties of the material, but also on external factors. Humans are capable of estimating the reflectance properties of a surface [28, 228, 68] even when there is no information about its illumination [67, 85], yet we perform better under illuminations that match real-world statistics [86]. We are not good perceiving the shape of an object in isolation [238, 264], and if the curvature of a geometry is carefully tweaked, we can modify our perception of the material it is made of [320], or of the illumination in the scene [240]. Usually, in order to accurately perceive the materials that surround us, humans rely on a combination of cues that include shape [320, 118, 267, 227, 216], illumination [347, 19, 32, 310, 128, 165], motion [207, 63, 117, 312, 221, 262, 208], touch [334, 309], or the interaction between some of them [172, 43, 275, 236]. This thesis continues taking steps towards understanding the influence of confounding factors in the perception

of material appearance. We start by exploring the joint influence of geometry and illumination in our performance recognizing materials (Chapter 5); then, we analyze the influence of motion in our perception of a set of material attributes through several user studies (Chapter 6).

2.2 VISUAL APPEARANCE MODELING

Humans are visual creatures, as such, visual appearance and material appearance are a core aspect, not only on how we understand the world, but also on how we create digital content. Current pipelines to model appearance are heterogeneous, and the data are machine-friendly but not human-friendly. This creates a disconnection between the physical parameters used by those pipelines, and the perceptual parameters that humans understand [3, 2, 81]. This chapter reviews the literature regarding visual appearance modeling, discuss current datasets capturing real-world material appearance, present approaches to circumvent the disconnection between physical parameters and perceptual parameters, and review the literature to model appearance similarity in iconography.

MATERIAL DATASETS One core aspect to manipulate material appearance is to have accurate computational representations in the form of datasets that cover a wide spectrum of real-world appearances. Early image-based material datasets include CURet [51], KTH-TIPS [120], or FMD [280]. OpenSurfaces [22] includes over 20,000 real-world images, with surface properties annotated via crowdsourcing. This dataset has served as a baseline to others, such as the Materials in Context Database (MINC) [23], an order of magnitude larger; SynBRDF [160], with 5,000 rendered materials randomly sampled from OpenSurfaces; or MaxBRDF dataset [323], which includes synthetic anisotropic materials. Also, there have been efforts on designing common languages for the description of materials such as MDL (Material Definition Language) [155] which can be rendered with the OptiX ray tracing engine [242]. Databases with *measured* materials include MERL [211] for isotropic materials, UTIA [79] for anisotropic ones, the Objects under Natural Illumination Database [194], which includes calibrated HDR information, or the database by Dupuy and Jakob which measures spectral reflectance [70]. For most of the proposed methods and frameworks in this thesis we have chosen, as a starting point, the MERL dataset since it contains a wider variety of isotropic materials, directly measured from the real-world, and it is still being successfully used in many applications such as gamut mapping [299], material editing [276, 297, 275], BRDF parameterization [292], or photometric light source estimation [199]. In this thesis (Chapter 3), we also contribute to that end by publicly releasing a dataset with wide variety of realistic scenes under different illuminations, geometries, and materials.

LOW-DIMENSIONAL SPACES A large body of work has been devoted to analyzing the relationships between the physical parameters defining material appearance, and deriving low-dimensional perceptual embeddings [211, 331, 226, 276, 292, 78, 50]. These embeddings can be further used to derive material similarity metrics, which are useful to determine if two materials convey the same appearance, and thus benefit a large number of applications. Those include metrics computed either directly over measured BRDFs [91, 223], in image space [248, 224, 299], or in re-parametrizations of BRDF spaces based on perceptual traits [246, 276]. Besides, low-dimensional spaces modeling

material appearance are useful for particular applications such as BTF filtering [142], understanding the motion of liquids [149], representing cloth [8], material editing [276, 219], or manipulation of textural information [184].

DEEP-LEARNING BASED METHODS Recent works have suggested that material perception might be driven by complex non linear statistics, similar to the ones extracted by neural networks [89, 295, 57]. Image patches have shown to contain enough information for material recognition [273], and several works have leveraged this to derive learning techniques for material recognition tasks. Recent work has shown the extraordinary ability of deep features to match human perception in the assessment of perceptual similarity between two images [350]. Bell et al. [23] introduce a CNN-based approach for local material recognition using a large annotated database, while Schwartz and Nishino explicitly introduce global contextual cues [272]. Georgoulis et al. [102] use both an object’s image and its geometry to create a full reflectance map, which is later used as an input to a four-class coarse classifier (metal, paint, plastic or fabric). Deep neural networks have been also successfully used for material editing in inverse rendering problems [190, 212, 256, 180]. In this thesis, we draw inspiration from the successes of the works mentioned above, and rely on deep learning frameworks for measuring material similarity while correlating with human judgements (Chapter 3), using siamese neural networks to measure icons similarity (Chapter 4), creating a framework for intuitive editing of appearance from a single-image in the problem of human relighting (Chapter 7), and developing an intuitive method for image-based material editing (Chapter 8).

MODELING ICONS APPEARANCE The visual appearance of icons plays a key role in how we interact with graphical user interfaces in our everyday life. The style and visual identity that is given by the artists when creating them, plays a key role in the message they convey [317, 318, 132, 6]. However, such message may not be easy to computationally model, in part, due to the interaction with human perception. Previous works have focused on generating semantically relevant icons to improve visualizations [277]. In particular, Setlur and Mackinlay [278] develop a method for mapping categorical data to icons. They found out that users prefer stylistically similar icons within a set, as opposed to automatic sets that might differ in look-and-feel. Lewis et al. [185] studied how the perception of icons is affected by spatial layouts, and present a shape grammar to generate visually distinctive icons. More recently, the work of Liu et al. [193] proposes a semi-automatic method to create icons from images according to a given style, while the work of Bernstein and Li [26] describes a technique to make icons scale independent. Computationally measuring the similarity between icons is an interesting and complex problem with the potential of developing more intuitive applications that would help professional workflows and novice users alike. We investigate this problem in Chapter 4, where we propose an image-based measure of icons similarity by taking into account both, their style and their visual identity, and using deep learning methods.

MEASURING STYLE AND SHAPE SIMILARITY Style similarity metrics have been recently proposed for fonts [232], infographics [263], 3D models [200, 192], or interior designs [21]. The work of Garces et al. [94] uses a hand-made feature vector to measure style similarity for clip arts. However, since the feature descriptors were manually selected for that particular task,

they do not account for additional high-level properties. In a follow-up work, they find that shape is a property that people take into account when comparing clip arts [95], however, it is not measured in their existing style metric. Deep learning-based methods have recently also been used to fit icons into a particular interface [353] or to learn personal styles [189]. To measure shape similarity is a long-standing problem in computer graphics. Bober [30] shows how to represent and match shape representations under the MPEG-7 standard [285]. Osada et al. [239] propose several silhouette-based descriptors that can be used for 2D and 3D shape retrieval. Other shape descriptors have been proposed, including Hu-moments [133], shape context [24], the use of Zernike moments [158], a pyramid of descriptors [167], or Fourier descriptors [346]. Kleiman et al. [164] focused on 3D shape similarity, using part-based models, while other works compare shapes using single closed contours [176, 12]. In contrast to the works that rely on a feature-based representation of the data, kernel methods aim to obtain directly the similarity matrix for a fixed set of objects, thus, such approaches do not generalize to objects outside the chosen set [110, 283]. The work of Laursen et al. [74] proposes an embedding of a small fixed set of icons optimized for comprehensibility and identifiability properties. Demiralp et al. [59] re-order icon sets to maximize perceptual discriminability. Also, kernel methods have been used to measure similarity and to propose content-based retrieval methods [72, 332, 66, 333]. The framework presented in Chapter 4 proposes to automatically learn a distance metric measuring style and visual identity in icons by training a deep siamese neural network.

2.3 VISUAL APPEARANCE EDITING

Editing the visual appearance is a complex task since there is a disconnection between physical parameters used to model appearance and the parameters that our perception manipulates. We provide here a brief cross-section of different material editing approaches, inverse rendering and image-based rendering approaches, and discuss the human relighting problem.

BRDF EDITING Several perceptually-based frameworks have been proposed to provide users with more intuitive controls over parametric appearance models [76, 247, 14]. Non-parametric models such as measured BRDFs are harder to edit. One potential solution would be to fit the non-parametric BRDFs to parametric models [298, 11, 27], use inverse shading trees [177], rely on polynomial bases [25], or directly employing deep-learning techniques [356]. Other authors have proposed links between human perception and editing of non-parametric BRDFs through a set of intuitive perceptual traits [219, 276, 211]. A comprehensive review of appearance editing methods can be found in the work of Schmidt et al. [269]. Recently, the emergence of NeRF [215] allowing to capture and model a 3D scene from several photographs taken at known locations, has enabled frameworks with unprecedented levels of realism, where follow-up works also allow the materials to be edited [293, 348, 352]. However, these methods only provide a new material definition that can later be used in a 3D scene, but do not allow to modify the material directly in an existing image.

IMAGE-BASED EDITING Image-based material editing techniques allow the user to directly alter the pixels in an image without manipulating an underlying BRDF nor requiring to re-render a scene. The work of Khan

et al. [156] exploits the fact that human vision is tolerant to many physical inaccuracies to propose a material editing framework requiring a single HDR image as input. Such approach was later extended to include global illumination [115] or weathering effects [337]. Other methods are based on frequency-domain analyses [33], visual goals [225], or use a light field as input [20, 141]. Since geometry and illumination also play a key role in the final appearance of the material, several works focused on explicitly decomposing the image into material, illumination and geometry information [17, 116, 341, 97], allowing to manipulate each of these properties independently. This thesis presents a new method to manipulate material appearance intuitively and in the wild, just from photographs (discussed in Chapter 8), where we rely on a combination of human judgements and a novel generative architecture allowing to produce realistic material edits.

INVERSE RENDERING METHODS The inverse rendering problem, requires to obtain the shape, material, and illumination from a single image. This is a highly ill-posed problem, with infinite solutions, classically solved assuming that some information is known beforehand. *Shape from shading* [252, 136] is one of the earliest methods, estimating shape from shading under a known illumination. Other methods estimate shape relying on simple illumination models such as directional, point, or area light sources [47, 233, 197], or environmental lighting encoded into spherical harmonics [145]. Reflectance and illumination can be estimated from a known convex shape [42], a shape with occluding contours [196], or just an approximated geometry [157]. A similar line of research has focused on *intrinsic images* [16, 338, 96, 97, 329], which aims to decompose a scene into its shading and albedo components [175]. Recent techniques leverage deep learning to predict illumination [130, 129, 98, 172], estimate specular reflectance and illumination [101, 195], devise material reflectance metrics [57, 171], or perform intrinsic image decomposition [204, 188, 17]. Other line of work, relies on complex hardware setups [48, 114, 351] to achieve this goals, however such setups are not widely available. In this thesis, we approach the problem of human relighting in Chapter 7, where we draw inspiration from inverse rendering and intrinsic images. Our framework estimates albedo and shading from a single input image, however, we additionally decompose shading into shape and illumination by developing a framework inspired by precomputed radiance transfer (PRT) [253, 181, 291]. Our decomposition also takes into account diffuse and specular material reflectance, thus producing more realistic relighted results.

IMAGE-BASED RENDERING A classic application of image-based rendering (IBR) [284] allows to take several pictures of a subject from the same viewpoint under different illuminations, and relight it using a weighted linear combination of those images [54, 55]. More sophisticated approaches optimize energy functions [183], work with layered decompositions [290], or employ RGB-D cameras [125]. However, those techniques require a large number of input images, as well as precise control over the lighting, making them unfeasible for single-image, in-the-wild applications. Recent work exploits the potential of implicit representations and Fourier mappings of the input to learn high-quality 3D scene representations using one multi-layer perceptron (MLP) per scene and several hundreds of images as the input [215, 349, 31], although these methods do not generalize across scenes.

The work of Wang et al. [326] addresses this by combining implicit models with IBR to generate novel views without relighting.

HUMAN RELIGHTING Human relighting refers to the problem of changing the appearance of a scene, with a human in it, by manipulating its illumination. Single-image human relighting approaches have been proposed for faces [328]: Sengupta et al. [274] show how we can relight faces using convolutional neural networks and spherical harmonics, later extended with more complex model architectures [354], or by directly fitting encoder-decoder architectures to light-stage portrait data [296, 222]. In this thesis (Chapter 7), we devise a single-image in-the-wild method to intuitively perform full-body human relighting using a UNet-like neural network [173]. Closer to the proposed method is the work of Kanamori and Endo [148], performing full-body relighting and assuming Lambertian materials. In contrast, we present a framework that lifts their assumption of materials being Lambertian by explicitly modeling the diffuse and specular reflectance in our data. We also add a residual term to the image reconstruction equation that allows to better model errors in the relighted image reconstruction.

Part III

MEASURING APPEARANCE SIMILARITY

The first half of this part introduces a framework to measure similarity in the complex domain of material appearance. The main contribution is the integration of human judgements within a deep learning framework using a custom loss function. This, in turn yields a framework whose similarity values agree with the human notion of material appearance similarity. Second half proposes a method to measure similarity in icons gathered from online databases. The main contribution is the gathered dataset together with the siamese deep learning architecture used to directly measure similarity relying on semantic labels.

This chapter introduces a model to measure the similarity in appearance between different materials, which correlates with human similarity judgements. We first create a database of 9,000 rendered images depicting objects with varying materials, shape and illumination. We then gather data on perceived similarity from crowdsourced experiments; our analysis of over 114,840 answers suggests that indeed a shared perception of appearance similarity exists. We feed this data to a deep learning architecture with a novel loss function, which learns a feature space for materials that correlates with such perceived appearance similarity. Our evaluation shows that our model outperforms existing metrics. Last, we demonstrate several applications enabled by our metric, including appearance-based search for material suggestions, database visualization, clustering and summarization, and gamut mapping.

This work has been published in *ACM Transactions on Graphics (TOG)* and presented at *SIGGRAPH 2019* [171]. While I was the leading author (under the supervision of Belén Masiá and Diego Gutiérrez); Ana Serrano, Sandra Malpica, and Elena Garces provided invaluable help with developing additional analysis and applications, on the manuscript text, and figures.

A follow up work analyzing in depth the role of objective and subjective measures in material similarity learning was later presented as a peer-reviewed poster in *ACM SIGGRAPH 2020 Posters* [57]. I was not the leading author in this work. My contribution was setting up and training the deep learning networks with the different configurations used in the analysis.

M. Lagunas, S. Malpica, A. Serrano, E. Garces, D. Gutierrez, & B. Masia
A Similarity Measure for Material Appearance
ACM Transactions on Graphics Vol. 38 (4), SIGGRAPH 2019

J. Delanoy, M. Lagunas, I. Galve, D. Gutierrez, A. Serrano, R. Fleming, & B. Masia
The Role of Objective and Subjective Measures in Material Similarity Learning
ACM SIGGRAPH 2020 Posters

3.1 INTRODUCTION

Humans are able to recognize materials, compare their appearance, or even infer many of their key properties effortlessly, just by briefly looking at them. Many works propose classification techniques, although it seems clear that labels do not suffice to capture the richness of our subjective experience with real-world materials [83]. Unfortunately, the underlying perceptual process of material recognition is complex, involving many distinct variables; such process is not yet completely understood [9, 82, 205].

Given the large number of parameters involved in our perception of materials, many works have focused on individual attributes (such as the perception of gloss [246, 331], or translucency [105]), while others have focused on particular applications like material synthesis [355], editing [276], or filtering [142]. However, the fundamentally difficult problem of establishing a *similarity measure for material appearance* remains an open problem. Material appearance can be defined as “the visual impression we have of a material” [65]; as such, it depends not only on the BRDF of the material, but also on external factors like lighting or geometry, as well as human judgement [82, 3]. This is different from the common notion of image similarity

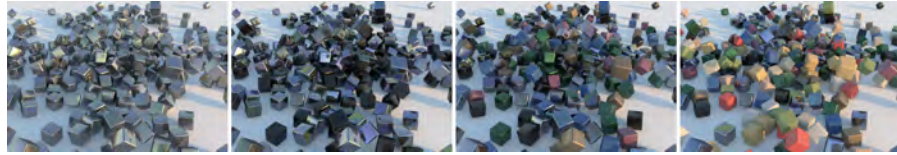


Figure 3.1: The cubes in the leftmost image have all been rendered with the same aluminium material. Our similarity measure for material appearance can be used to automatically generate alternative depictions of the same scene, where the similarity of the materials varies in a controlled manner. The next three images show results with materials randomly chosen by progressively extending the search distance from the original aluminium, from similar in appearance to farther away materials within the same dataset.

(devoted to finding detectable differences between images, e.g., [327]), or from similarity in BRDF space (which has been shown to correlate poorly with human perception, e.g., [276]). Given the ubiquitous nature of photorealistic computer-generated imagery, and emerging fields like computational materials, a similarity measure of material appearance could be valuable for many applications.

Capturing a human notion of perceptual similarity in different contexts has been an active area of research recently [94, 5, 201]. In this chapter we develop a novel image-based material appearance similarity measure derived from a learned feature space. This is challenging, given the subjective nature of the task, and the interplay of confounding factors like geometry or illumination in the final perception of appearance. Very recent work suggests that perceptual similarity may be an emergent property, and that deep learning features can be trained to learn a representation of the world that correlates with perceptual judgements [350]. Inspired by this, we rely on a combination of large amounts of images, crowdsourced data, and deep learning. In particular, we create a diverse collection of 9,000 stimuli using the measured, real-world materials in the MERL dataset [211], which covers a wide variety of isotropic appearances, and a combination of different shapes and environment maps. Using triplets of images, we gather information through Mechanical Turk, where participants are asked which of two given examples has a more similar appearance to a reference. Given our large stimuli space, we employ an adaptive sampling scheme to keep the number of triplets manageable. From this information, we learn a model of material appearance similarity using a combined loss function that enforces learning of the appearance similarity information collected from humans, and the main features that describe a material in an image; this allows us to learn the notion of material appearance similarity explained above, dependent on both the visual impression of the material, and the actual physical properties of it.

To evaluate our model, we first confirm that humans do provide reliable answers, suggesting a shared perception of material appearance similarity, and further motivating our similarity measure. We then compare the performance of our model against humans: Despite the difficulty of our goal, our model performs on par with human judgements, yielding results better aligned with human perception than current metrics. Last, we demonstrate several applications that directly benefit from our metric, such as material suggestions (see Figure 3.1), database visualization, clustering and summarization, or gamut mapping. In addition to the 9,000 rendered images, our database also includes surface normals, depth, transparency, and ambient oc-

¹All the code, data, and models are available at: webdiis.unizar.es/~mlagunas/publication/material-similarity/



Figure 3.2: All six environment maps used in the dataset and corresponding rendered spheres with the *black-phenolic* material.

clusion maps for each one, while our collected data contains 114,840 answers; we provide both, along with our pre-trained deep learning framework, in order to help future studies on the perception of material appearance¹.

3.2 MATERIALS DATASET

3.2.1 Why a New Materials Dataset?

To obtain a meaningful similarity measure of material appearance we require a large dataset with the following characteristics:

- Data for a wide variety of materials, shapes, and illumination conditions.
- Samples featuring the *same* material rendered under different illuminations and with different shapes.
- Materials represented by measured BRDFs, with reflectance data captured from real materials.
- Real-world illumination, as given by captured environment maps.
- A large number of samples, amenable to learning-based frameworks.

These characteristics enable renditions of complex, realistic appearances and will be leveraged to train our model, explained in Section 3.4. To our knowledge, none of the existing material datasets features all these characteristics.

3.2.2 Description of the Dataset

In the following, we thoroughly describe the characteristics of our dataset, including implementation details.

MATERIALS Our dataset includes all 100 materials from the MERL BRDF database [211]. This database was chosen as starting point since it includes real-world, measured reflectance functions covering a wide range of reflectance properties and types of materials, including paints, metals, fabrics, or organic materials, among others.

ILLUMINATIONS We use six natural illumination environments, since they are favored by humans in material perception tasks [86]. They include a variety of scenes, ranging from indoor scenarios to urban or natural landscapes, as high-quality HDR environment maps².

²The HDR illuminations are gathered from: <http://gl.ict.usc.edu/Data/HighResProbes/>

SCENES Our database contains 13 different 3D models, with an additional camera viewpoint for two of them, defining our 15 scenes. It includes widely used 3D models, and objects that have been specifically designed for material perception studies [118, 320]. The viewpoints have been chosen to cover a wide range of features such as varying complexity, convexity, curvature, and coverage of incoming and outgoing light directions. The 3D models feature different complexities and come from various sources including:

³Stanford 3D repository:
<http://graphics.stanford.edu/data/3Dscanrep/>

⁴TurboSquid website with 3D models:
<https://www.turbosquid.com>

- The Stanford 3D Scanning Repository³ (4 geometries and 5 views) .
- TurboSquid⁴ website (3 geometries and 4 views).
- Havran et al [118] perceptually motivated geometries (2 geometries and 2 views).
- Vangorp et al [320] blob shape (1 geometry and 1 view).
- Blender’s mascot *Suzzane* (1 geometry and 1 view).
- Utah teapot (1 geometry and 1 view).
- Standard sphere (1 geometry and 1 view).

By combining the aforementioned materials (100), illumination conditions (6), and scenes (15), we generate a total of 9,000 dataset samples using the Mitsuba physically-based renderer [140] with Jonathan Dupuy’s library⁸, which includes support for MERL BRDFs. For each one we provide: The rendered HDR image, a corresponding LDR image⁵, along with depth, surface normals, alpha channel, and ambient occlusion maps. All the images are rendered with a resolution of 896×896 pixels using path tracing with different samples per pixel according to the geometry (between 512 and 2048) to avoid visible artifacts. After rendering, we use a bilinear up-sampling on the resulting image to obtain a final resolution of 1024×1024 . While not all these maps are used in the present work, we make them available with the dataset should they be useful for future research. Figure 3.3 shows sample images for all 15 scenes.

⁵All HDR images are tone-mapped using the algorithm by Mantiuk et al. [206], with the predefined lcd office display, and color saturation and contrast enhancement set to 1.

3.3 COLLECTING APPEARANCE SIMILARITY INFORMATION

We describe here our methodology to gather crowdsourced information about the perception of material appearance.

STIMULI We use 100 different stimuli, covering all 100 materials in the dataset, rendered with the *Ennis* environment map. We choose the *Havran-2* scene, since its shape has been designed to maximize the information relevant for material appearance judgements by optimizing the coverage of incoming and outgoing light directions sampled [118]. Figure 3.4 shows some examples.

PARTICIPANTS A total of 603 participants took part in the test through the Mechanical Turk (MTurk) platform, with an average age of 32, and 46.27% female. Users were not aware of the purpose of the experiment.

PROCEDURE Our study deals with the *perception* of material appearance, which may not be possible to represent in a linear scale; this advises against ranking methods [152]. We thus gather data in the form of relative comparisons, following a 2AFC scheme; the subject is presented with a triplet

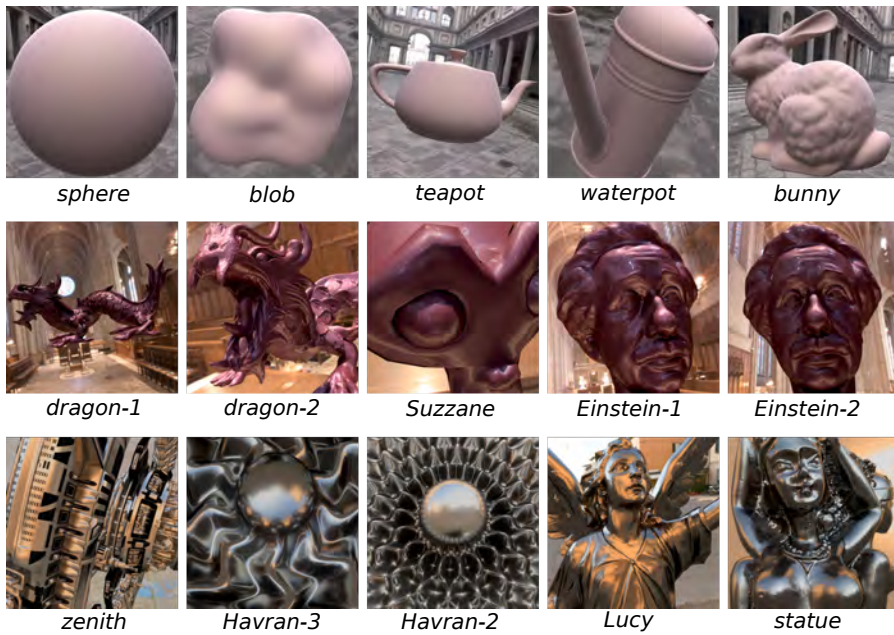


Figure 3.3: Sample images of all 15 scenes with different materials and illumination conditions. First row: *pink-felt* and *Uffizi*; second row: *violet-acrylic* and *Grace*; third row: *nickel* and *Pisa*. The 3D models *bunny*, *dragon*, *Lucy* and *statue* belong to The Stanford 3D Scanning Repository; *waterpot* (modelled by gykservy), *Suzzane* (killzone75), *Einstein* (oliverlaric), and *zenith* (Kuh-Industries) were obtained from TurboSquid.

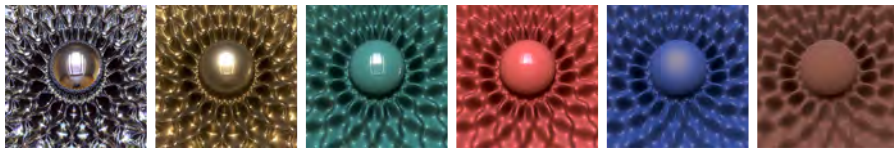


Figure 3.4: Sample stimuli for our appearance similarity collection. They correspond to the *Havran-2* scene, with materials from the MERL database, rendered with the *Ennis* environment map. In reading order: *chrome*, *gold-metallic-paint*, *specular-green-phenolic*, *maroon-plastic*, *dark-blue-paint* and *light-brown-fabric*.

made up of one *reference* material, and two *candidate* materials, and their task is to answer the question *Which of these two candidates has a more similar appearance to the reference?* by choosing one among the two candidates. This approach has several additional advantages: it is easier for humans than providing numerical distances [213, 271], while it reduces fatigue and avoids the need to reconcile different scales of similarity among subjects [151].

However, given our 100 different stimuli, a naive 2AFC test would require 495,000 comparisons, which is intractable even if not all subjects see all comparisons. To ensure robust statistics, we aim to obtain five answers for each comparison, which would mean testing a total of 2,475,000 comparisons. Instead, we use an iterative *adaptive sampling* scheme [303]: For any given reference, each following triplet is chosen to maximize the information gain, given the preceding responses (please refer to Section 3.3.1 for a more detailed description of the method).

Each test (HIT in MTurk terminology) consisted of 110 triplets. To minimize worker unreliability [330], each HIT was preceded by a short training session that included a few trial comparisons with obvious answers [258, 94]. In addition, 10 control triplets were included in each HIT, testing repeated-trial consistency within participants. We adopt a conservative approach and reject participants with two or more different answers. In the end, we obtained 114,840 valid answers, yielding a participants’ consistency of 84.7%.

As a separate test, to validate how well our collected answers generalize to other shapes and illuminations, we repeated the same comparisons, this time with symmetric and asymmetric triplets chosen randomly from our dataset, with the condition that they do not contain the *Havran-2* shape nor the *Ennis* illumination. For symmetric triplets, the three items in the triplet differ only in the material properties, while in asymmetric triplets geometry and lighting also vary. We launched 2,500 symmetric triplets, and found that participants’ majority matched the previous responses with a 84.59% rate. When we added the same number of asymmetric triplets to the test, participants’ answers held with a 80% match rate.

3.3.1 Adaptive Sampling Scheme for the User Studies

In order to adaptively sample the comparisons to be used as queries in the material appearance similarity study, we use an *adaptive sampling* algorithm proposed by Tamuz et al. [303], which takes into account previously answered queries in order to estimate the optimal pattern of triplets in the next iteration, by computing the potential information gain. We perform 25 iterations of the algorithm and in each iteration we sample 10 new pairs for every reference (not taking into account the 10 control ones), getting a total of 1000 triplets split in 10 HITs that are each answered by 5 users. After this process, the mean information gain per iteration is less than 10^{-5} , confirming the convergence of the sampling scheme. This scheme allows us to drastically reduce the number of required comparisons, while providing a good approximation to sampling the full set of triplets.

This method works iterating over two steps:

- Fit the kernel K using the answered triplets.
- Find the comparison (a, b) for each reference r that yields the largest information gain.

This adaptive sampling scheme measures how well a triplet $(r, a, b) \in T$ is modeled in terms of a probability, where T is the set of all triplets. K is defined as a kernel such that $K = XX^T$, where k_{ra} is the element of K in the row r and column a , and X is the low-dimensional embedding the optimization aims to find. The probability of a triplet (r, a, b) being well modeled is described as follows:

$$p_{rab} = \frac{k_{rr} + k_{aa} - 2k_{ra} + \mu}{(k_{rr} + k_{aa} - 2k_{ra}) + (k_{rr} + k_{bb} - 2k_{rb}) + 2\mu}, \quad (3.1)$$

where μ is usually a small value used as a regularizer (in our case $\mu = 0.05$). A higher value of p_{rab} indicates that a particular triplet is less well modeled. The algorithm tries to minimize the empirical log-loss of that probability:

$$\min_K \sum_{r,a,b \in T} \log(p_{rab}), \quad \text{subject to } \forall i \ k_{ii} = 1 \text{ and } K \succeq 0. \quad (3.2)$$

Once the optimization in Equation 3.2 converges, the kernel K can be used to obtain the materials that need more samples in order to be well modeled.

To obtain the pair of materials that yields the greatest information gain for any reference image a posterior distribution $\tau(r)$ for the answered triplets A is computed as follows:

$$\tau(r) \propto \pi(r) \prod_{(r,a,b) \in T} p_{rabr}, \quad (3.3)$$

where $\pi(r)$ is a prior distribution over the reference images, defined as a uniform distribution. The probability that users will rate a as more similar to any candidate r than to b can be expressed as:

$$p \propto \int_r \frac{k_{rr} + k_{aa} - 2k_{ra}}{(k_{rr} + k_{aa} - 2k_{ra}) + (k_{rr} + k_{bb} - 2k_{rb})} \tau(r) dr, \quad (3.4)$$

if the user rates a more similar to r then, it has a posterior distribution τ_a (note τ_b can be computed similarly):

$$\tau_a = \tau(r) \frac{k_{rr} + k_{aa} - 2k_{ra}}{(k_{rr} + k_{aa} - 2k_{ra}) + (k_{rr} + k_{bb} - 2k_{rb})}. \quad (3.5)$$

Finally, for every possible triplet the information gain can be computed as the difference between the information of the answered triplets and the possible triplets that can be sampled such that:

$$IG = H(\tau) - pH(\tau_a) - (1-p)H(\tau_b), \quad (3.6)$$

where H is an entropy measure $H(\tau) = -\sum \tau \log \tau$.

Since in our user-study each HIT is answered by 5 users and the model assumes single answers, we consider the opinion of the majority as the answer for each triplet.

3.4 LEARNING PERCEIVED SIMILARITY

This section describes our approach to learn perceived similarity for material appearance. Given an input image ψ , our model provides a feature vector $f(\psi)$ that transforms the input image into a feature space well aligned with human perception.

We use the ResNet architecture [122], based on its generalization capabilities and its proven performance on image-related tasks. The novelty of this architecture is a residual block meant for learning a residual mapping between the layers, instead of a direct mapping, which enables training very deep networks (hundreds of layers) with outstanding performance. For training we use image data from our materials dataset (Section 3.2), together with human data on perceived similarity (Section 3.3). We first describe our combined loss function, then our training procedure.

3.4.1 Loss Function

We train our model using a loss function consisting of two terms, equally weighted:

$$\mathcal{L} = \mathcal{L}_{TL} + \mathcal{L}_P \quad (3.7)$$

The two terms represent a perceptual triplet loss, and a similarity term, respectively. The terms aim at learning appearance similarity from the

participants' answers, while extracting the main features defining the material depicted in an image. In the following, we describe these terms and their contribution.

3.4.1.1 Triplet Loss Term \mathcal{L}_{TL}

This term allows to introduce the collected MTurk information on appearance similarity. Let $\mathcal{A} = \{(r_i, a_i, b_i)\}$ be the set of answered relative comparisons, where r is the reference image, a is the candidate image chosen by the majority of users as being more similar to r , and b the other candidate; i indexes over all the relative comparisons. Intuitively, r and a should be closer together in the learned feature space than r and b . It is not feasible to collect user answers for all possible comparisons (n different images would lead to $n \binom{n-1}{2}$ tests); however, as we have shown in Section 3.3, the collected answers for a triplet (r, a, b) involving materials m^r , m^a and m^b generalize well to other combinations of shape and illumination from our dataset involving the same set of materials. We thus define $\mathcal{A}^M = \{(m_i^r, m_i^a, m_i^b)\}$ as the set of relative comparisons with collected answers (m^a represents the material chosen by the majority of participants). We then formulate the first term as a triplet loss [45, 270, 169]:

$$\mathcal{L}_{TL} = \frac{1}{|\mathcal{B}^A|} \sum_{(r,a,b) \in \mathcal{B}^A} [\|f(r) - f(a)\|_2^2 - \|f(r) - f(b)\|_2^2 + \mu]_+ \quad (3.8)$$

where $f(\psi)$ is the feature vector of image ψ , and the set \mathcal{B}^A is defined as:

$$\mathcal{B}^A = [(r, a, b) \mid (m^r, m^a, m^b) \in \mathcal{A}^M \wedge (r, a, b) \in \mathcal{B}] \quad (3.9)$$

with \mathcal{B} the current training batch. In Equation 3.8, μ represents the margin, which accounts for how much we aim to separate the samples in the feature space.

3.4.1.2 Similarity Term \mathcal{L}_P

We introduce a second loss term that maximizes the log-likelihood of the model choosing the same material as humans. We define this probability p_{ra} (and conversely p_{rb}) as a quotient between similarity values s_{ra} and s_{rb} :

$$p_{ra} = \frac{s_{ra}}{s_{rb} + s_{ra}}, \quad p_{rb} = \frac{s_{rb}}{s_{rb} + s_{ra}} \quad (3.10)$$

These similarities are derived from the distances between r , a and b in the feature space, where a similarity value of 1 means perfect similarity and a value of 0 accounts for total dissimilarity:

$$s_{ra} = \frac{1}{1 + d_{ra}}, \quad s_{rb} = \frac{1}{1 + d_{rb}}, \quad \text{where} \quad (3.11)$$

$$d_{ra} = \|f(r) - f(a)\|_2^2, \quad d_{rb} = \|f(r) - f(b)\|_2^2 \quad (3.12)$$

With this, we can formulate the similarity term as:

$$\mathcal{L}_P = -\frac{1}{|\mathcal{B}^A|} \sum_{(r,a,b) \in \mathcal{B}^A} \log p_{ra} \quad (3.13)$$

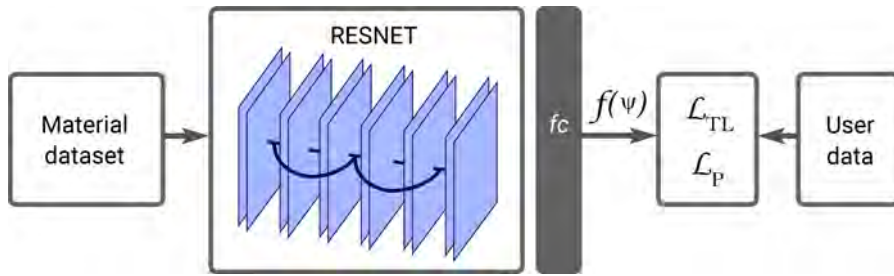


Figure 3.5: Scheme of the training process, using both image data from our material dataset, and human data of perceived similarity. We train our model so that, for an input image ψ , it yields a 128-dimensional feature vector $f(\psi)$.

3.4.2 Training Details

For training, we remove the *Havran-2* and *Havran-3* scenes from the dataset, leading to 7,800 images (13 (scenes) \times 6 (env. maps) \times 100 (materials)), augmented to 39,000 using crops, flips, and rotations. These 39,000 images, together with the collected MTurk answers, constitute our training data. We use the corrected *Adam* optimization [255, 161] with a learning rate that starts at 10^{-3} to train the network. We train for 80 epochs and the learning rate is reduced by a factor of 10 every 20 epochs. For initialization, we use the weights of the pre-trained model [122] on ImageNet [60, 260]. To adapt the network to our loss function, we remove the last layer of the model and introduce a fully-connected (*fc*) layer that outputs a 128-dimensional feature vector $f(\psi)$. We use a margin $\mu = 0.3$ for the triplet loss term \mathcal{L}_{TL} . Figure 3.5 shows a scheme of the training procedure.

3.5 EVALUATION

We evaluate our model on the set of images of the material dataset not used during training. We employ the *accuracy* metric, which represents the percentage of triplet answers correctly predicted by our model. It can be computed as *raw*, considering each of the five answers independently as the correct one, or *majority*, considering the majority opinion as correct [331, 94]. Using our MTurk data from Section 3.3, the results are 73.10% and 77.53% respectively for human observers, indicating a significant agreement across subjects. Our model performs better than human accuracy, with 73.97% and 80.69% respectively. In other words, our model predicts the majority’s perception of similarity almost 81% of the time. We include an *oracle* predictor in Table 3.1, which has access to all the human answers and returns the majority opinion; note that its raw accuracy is not 100 due to human disagreement. Figure 3.6 shows examples from our 26,000 queries where our model agrees with the majority response, while we discuss failure cases later in this section. More examples of queries and our model’s answers are included in Appendix A, Section A.2.

3.5.1 Comparison with Other Metrics

We compare the performance of our model to six different metrics used in the literature for material modeling and image similarity: The three common metrics analyzed by Fores and colleagues [91], the perceptually-based metrics by Sun et al. [299] and Pereira et al. [248], and SSIM [327], a well-known

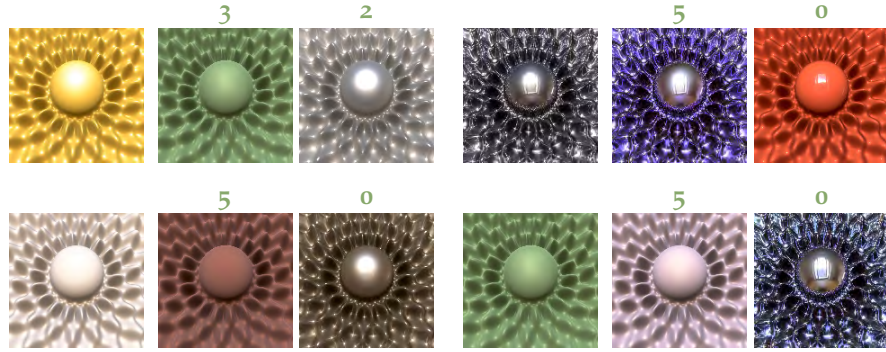


Figure 3.6: Examples from our 26,000 queries (reference, plus the two candidates) where our model agrees with the majority response (this is the case almost 81% of the time). The numbers indicate the number of votes each image received from the participants. More examples are included in Appendix A, Section A.3.

image similarity metric. We analyze again accuracy, and we additionally analyze *perplexity*, which is a standard measure of how well a probability model predicts a sample, taking into account the uncertainty in the model. Perplexity Q is given by:

$$Q = 2^{-\frac{1}{|\mathcal{A}|} \sum_{\Omega} \log_2 p_{ra}} \quad (3.14)$$

where $\Omega = (r, a) \in \mathcal{A}$, $|\mathcal{A}|$ is the number of collected answers, and p_{ra} is the probability of a being similar to r (Section 3.4.1). Perplexity gives higher weight where the model yields higher confidence; its value will be 1 for a model that gives perfect predictions, 2 for a model with total uncertainty (random), and higher than 2 for a model that gives wrong predictions. As Table 3.1 shows, our model captures the human perception of appearance similarity significantly better, as indicated by the higher accuracy and lower perplexity values. Note that perplexity cannot be computed for humans nor the oracle, since they are not probability distributions.

Additionally, we compute the mean error between distances derived from human responses and our model’s predictions, across all possible material pair combinations from the MERL dataset. To obtain the derived distances from the collected human responses, we use t-Distributed Stochastic Triplet Embedding (tSTE) [316], which builds an n -dimensional embedding that aims to correctly represent participants’ answers. We use a value of $\alpha = 5$ (degrees of freedom of the Student-t kernel), which correctly models 87.36% of the participants’ answers. We additionally compute the mean error for the six other metrics. As shown in Figure 3.7, our metric yields the smallest error. Error bars correspond to a 95% confidence interval.

3.5.2 Ablation Study

We evaluate the contribution of each term in our loss function to the overall performance via a series of ablation experiments (see Table 3.2). We first evaluate performance using only one of the two terms (\mathcal{L}_{TL} and \mathcal{L}_P) in isolation. We also analyze the result of incorporating two additional loss terms, which could in principle apply to our problem: A cross-entropy term \mathcal{L}_{CE} , and a batch-mining triplet loss term \mathcal{L}_{BTL} . The former aims at learning a soft classification task by penalizing samples which do not belong to the

EVALUATION OF OUR MODEL				
Metric	Accuracy		Perplexity	
	Raw	Majority	Raw	Majority
Humans	73.10	77.53	-	-
Oracle	83.79	100.0	-	-
RMS	61.63	64.72	3.61	3.13
RMS-cos	61.60	64.67	3.86	3.33
Cube-root	63.71	67.40	1.96	1.86
L2-lab	63.76	67.21	2.16	2.07
L4-lab	60.60	62.93	15.36	11.66
SSIM	62.35	64.74	2.02	1.94
Our model	73.97	80.69	1.74	1.55

Table 3.1: Accuracy and perplexity of our model compared to human performance, an oracle (which always returns the majority opinion), and six other metrics from the literature: RMS, RMS-cos, Cube-root [91], L2-lab [299], L4-lab [248] and SSIM [327]. For accuracy, higher values are better, while for perplexity lower are better.

same class [302], while the latter has been proposed in combination with the cross-entropy term to improve the model’s generalization capabilities and accuracy [93] (more details about these two terms can be found in the appendix). Last, we analyze performance using *only* these two terms (\mathcal{L}_{CE} and \mathcal{L}_{BTL}), without incorporating participants’ perceptual data. As Table 3.2 shows, none of these alternatives outperforms our proposed loss function. Although the single-term \mathcal{L}_P loss function yields higher accuracy, it also outputs higher perplexity values; moreover, as Figure 3.7 shows, the mean error is much higher, meaning that it does not capture the notion of similarity as well as our model.

3.5.3 Alternative Networks

We have tested two alternative architectures, VGG [287], which stacks convolutions with non-linearities; and DenseNet [134], which introduces concatenations between different layers. Both models have been trained using our loss function. As shown in Table 3.2, both yield inferior results compared to our model. DenseNet has a low number of learned parameters, insufficient to capture the data distribution, hampering convergence. VGG has a larger number of parameters; however, the residual mapping learned by the residual blocks in the architecture of our model yields the best overall performance.

3.5.4 Results by Category

We additionally divide the materials into eight categories: *acrylics*, *fabrics*, *metals*, *organics*, *paints*, *phenolics*, *plastics*, and *other*, and analyze raw and majority accuracy in each. We can see in Table 3.3 how our model is reasonably able to predict human perception also within each category. For instance,

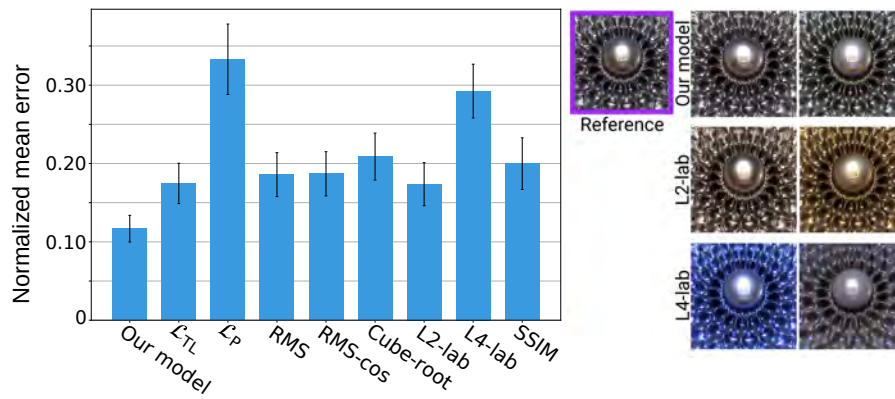


Figure 3.7: **Left:** Mean error for different metrics (each normalized by its maximum value) with respect to distances derived from human responses, across all possible pair combinations from the MERL dataset (the \mathcal{L}_{TL} and \mathcal{L}_P columns refer to the ablation studies in Table 3.2; please refer to the main text). Error bars correspond to a 95% confidence interval. **Right:** Representative example of the two most similar materials to a given reference, according to (from top to bottom): Our model, and the two perceptually-based metrics L2-lab [299], and L4-lab [248]. Our model yields less error, and captures the notion of appearance similarity better.

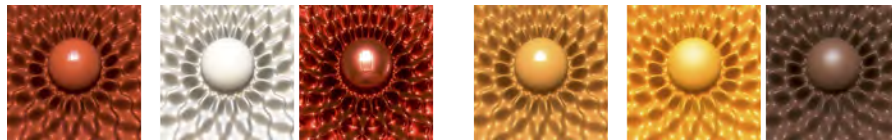


Figure 3.8: Two examples where humans' majority disagrees with our metric. For both, humans agreed that the middle stimulus is perceptually closer to the reference on the left, while our metric scores the right stimuli as more similar.

although the numbers are relatively consistent across all the categories, humans perform on average slightly worse for phenolics or acrylics, and better for fabrics; our metric mimics such behavior. The only significant difference occurs within the *organics* category, where our metric performs worse than humans. This may be due to the combination of a low number of material samples and a large variety of appearances within such category, which may hamper the learning process.

3.5.5 Failure Cases

Being on par with human accuracy means that our similarity measure disagrees with the MTurk majority 19.31% of the time. Figure 3.8 shows two examples where humans were consistent in choosing one stimuli as closer to the reference (5 votes out of 5), yet our metric predicts that the second one is more similar. In the leftmost example, the softness of shadows may have been a deciding factor for humans. In the rightmost example, humans may have been overly influenced by color, whilst our metric has factored in the presence of strong highlights. These examples are interesting since they illustrate that neither color nor reflectance are persistently the dominant factors when humans judge appearance similarity between materials.

ABLATION STUDY AND ALTERNATIVE NETWORKS					
Model	Accuracy		Perplexity		
	Raw	Majority	Raw	Majority	
\mathcal{L}_{TL}	69.32	74.12	1.89	1.73	
\mathcal{L}_P	75.22	82.31	3.16	2.13	
$\mathcal{L}_{TL} + \mathcal{L}_P + \mathcal{L}_{CE}$	71.82	77.53	1.76	1.66	
$\mathcal{L}_{TL} + \mathcal{L}_P + \mathcal{L}_{CE} + \mathcal{L}_{BTL}$	71.78	77.76	1.76	1.67	
$\mathcal{L}_{CE} + \mathcal{L}_{BTL}$	56.88	58.44	1.96	1.93	
VGG	70.70	76.40	2.25	1.89	
DenseNet	60.90	63.49	2.66	2.46	
Our model	73.97	80.69	1.74	1.55	

Table 3.2: Accuracy and perplexity for other loss functions, as well as for two alternative architectures (VGG and DenseNet).

3.6 APPLICATIONS

We illustrate here several applications directly enabled by our similarity measure.

3.6.1 Material Suggestions

Assigning materials to a complex scene is a laborious process [355, 44]. We can leverage the fact that the distances in our learned feature space correlate with human perception of similarity to provide controllable material suggestions. The artist provides the system with a reference material, and the system delivers perceptually similar (or farther away) materials in the available dataset, thus creating a controlled amount of variety without the burden of manually selecting each material. Figure 3.1 illustrates this, where the search distance is progressively extended from a chosen reference, and the materials are then assigned randomly to each cube. Suggestions need not be automatically assigned to the models in the scene, but may also serve as a palette for the artist to choose from, facilitating browsing and navigation through material databases. Figure 3.9 shows two MERL samples used as queries, along with returned suggestions from the *Extended MERL* dataset [276]. The figure shows results at close, intermediate, and far distances from the query. Additional examples can be seen in Figure 3.10, and in Appendix A.3.

3.6.2 Visualizing Material Datasets

The feature space computed by our model can be used to visualize material datasets in a meaningful way, using dimensionality reduction techniques. We illustrate this using UMAP (Uniform Manifold Approximation and Projection [214]), which helps visualization by preserving the global structure of the data. Figure 3.11 shows two results for the MERL dataset, using images not included in the training set. On the left, we can observe a clear gradient in reflectance, increasing from left to right, with color as a secondary, softer

ANALYSIS PER MATERIAL CATEGORY							
Category	Materials	Answers	Humans		Our model		Oracle
			Raw	Majority	Raw	Majority	Raw
Acrylics	4	4719	67.27	70.69	67.57	74.18	79.89
Fabrics	14	16019	79.65	83.70	83.03	90.44	87.87
Metals	26	32337	74.20	78.90	75.63	83.10	84.54
Organics	7	8370	69.28	73.08	60.46	62.43	81.28
Paints	14	15101	74.22	78.85	75.22	81.84	84.61
Phenolics	12	13025	66.49	70.53	67.62	74.36	79.72
Plastics	11	12031	70.53	74.70	69.25	74.06	82.05
Other	12	13198	74.80	79.38	78.21	86.11	84.89
Total	100	114800	73.10	77.53	73.97	80.69	83.79

Table 3.3: Statistics per category. **From left to right:** Category, number of materials in each category, number of collected answers, humans’ accuracy (raw and majority), accuracy of our model, and oracle raw accuracy.

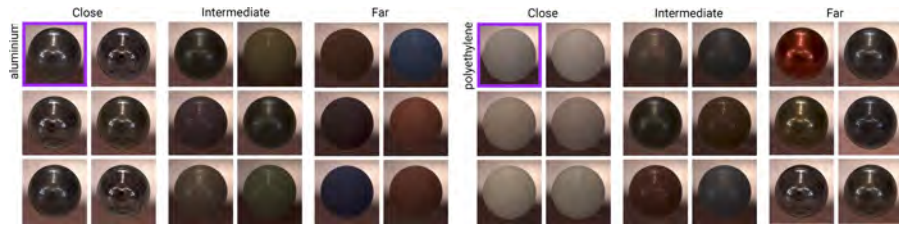


Figure 3.9: Two examples of material suggestions using our model. Queries from MERL (violet frame), and returned results for perceptually close, intermediate, and far away materials from the Extended MERL dataset.

grouping factor. The right image shows a similar visualization using only three categories: *metals*, *fabrics*, and *phenolics*.

3.6.3 Database Clustering

For unlabeled datasets like Extended MERL, our feature space allows to obtain clusters of perceptually similar materials. To further analyze the clustering enabled by our perceptual feature space, we rely on the Hopkins statistic, which estimates randomness in a data set [13]. A value of 0.5 indicates a completely random distribution, lower values suggest regularly-spaced data, and higher values (up to a maximum of 1) reveal the presence of clusters. The Hopkins statistic⁶ computed over our 128-dimensional feature vectors for the Extended MERL dataset yields a value of 0.9585, suggesting that meaningful clusters exist in our learned feature space (Figure 3.13 shows three representative clusters using the Extended MERL database). For comparison purposes, using only *metals* in MERL the Hopkins statistic drops to 0.6935, since their visual features are less varied within that category. Figure 3.12 shows an example of material suggestions leveraging our perceptual clusters in unlabeled datasets.

⁶The Hopkins statistic is an averaged value over 100 iterations since its computation involves random sampling of the elements in the dataset.

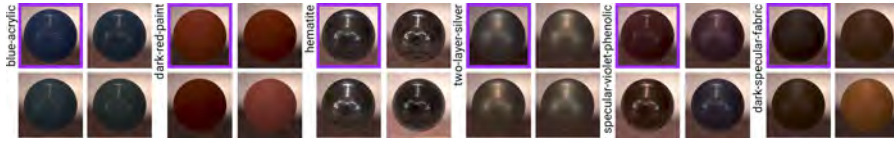


Figure 3.10: Additional material suggestion results. Queries (violet frame) and results for the closest materials in the *Extended MERL* dataset.

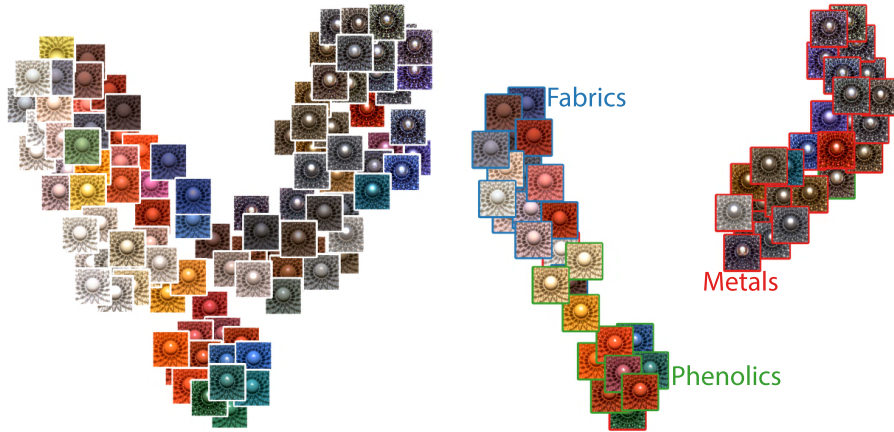


Figure 3.11: Visualization of the MERL dataset in a 2D space based on the feature vectors provided by our model, using UMAP [214]. **Left:** The entire MERL dataset. **Right:** Materials from three different categories (*metals*, *fabrics*, and *phenolics*).

3.6.4 Database Summarization

Perceptually meaningful clustering leads in turn to the possibility of database summarization. We can estimate the appropriate number of clusters using the elbow method, taking the number of clusters that explains the 95% of the variance in our feature vectors. In the 400-sample *Extended MERL* dataset, this results in seven clusters. Taking the closest material to the centroid for each one leads to a seven-sample database summarization that represents the variety of material appearances in the dataset (Figure 3.14).

3.6.5 Gamut Mapping

In general, our model can be used for tasks that involve minimizing a distance. This is the case for instance of gamut mapping, where the goal is to bring an out-of-gamut material into the available gamut of a different medium, while preserving its visual appearance; this is a common problem with current printing technology, or in the emerging field of computational materials. We illustrate the effectiveness of our technique in the former. Gamut mapping can be formulated as a minimization on image space [248, 299]. We can use our feature vector $f(\psi)$ to minimize the perceptual distance between two images as

$$\min_w \|f(o) - f(g * w)\|_2^2, \quad (3.15)$$

where o is the out-of-gamut image, and $g * w$ represents the image in the printer's gamut, defined as a linear combination of inks g [210]). Figure 3.15 shows some examples.



Figure 3.12: Material suggestions using our perceptual database clustering. The images show random materials assigned from three different clusters of varying appearance. The robot model (cKalten) was obtained from TurboSquid.

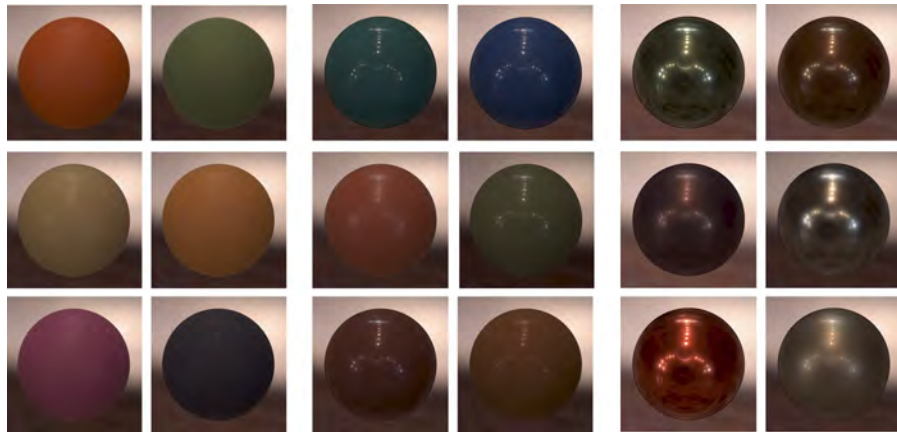


Figure 3.13: Representative samples of three clusters on the Extended MERL database. The Hopkins statistic on our feature space confirms that our similarity metric creates perceptually-meaningful clusters of materials.

3.7 OBJECTIVE AND SUBJECTIVE MEASURES

The deep learning model presented in Section 3.4 is shown to outperform existing objective metrics in reproducing human assessment of material similarity, presumably thanks to (i) the subjective measures used during the training, as well as (ii) the ability of the model to learn deep features. It has been shown that deep learning features can lead to a representation that correlates with perceptual judgements [350]. Consequently, we consider here to what extent each of the two characteristics above contribute to its success, and whether subjective measures are actually needed.

We train two additional networks: a *classification network* (trained to classify each image given the represented material), a *human similarity network* (original network described in Section 3.4) and a *BRDF similarity network* (analogous to the latter but trained to mimic the *Cube root cosine weighted BRDF* metric [91]).

We compute Representation Dissimilarity Matrices (RDMs) between a 10% random subset of the training images. The dissimilarity is measured as the l_2 norm between normalized feature vectors. By organizing these matrices according to various properties over the images, we can understand how the space is organized and which are the properties that are best captured by the network. We used image properties (shape and illumination) and material properties (color of the material and reflectance properties).

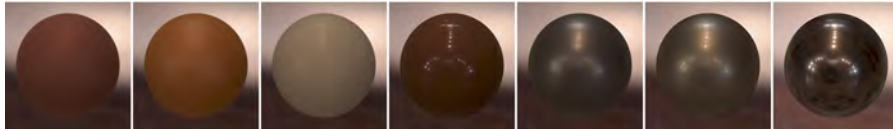


Figure 3.14: Example of database summarization for the Extended MERL dataset. These seven samples represent the variety of material appearances in the dataset.

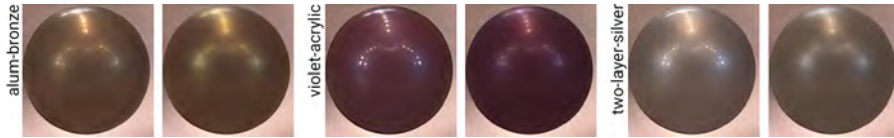


Figure 3.15: Our similarity metric can be used for gamut mapping applications, by minimizing the perceptual distance of our feature vectors. Each pair shows the groundtruth (left), and our in-gamut result (right).

3.7.1 Analysis

ANALYSIS OF RDMS In Figure 3.16, we show the RDMS organized by the reflectance properties of the materials. The classification network creates strong correlations only between images showing the same material but does not create relationships between ones with similar reflectance. Although classification networks have been used as a reference for perception metrics in recent works [350], we show here that the resulting feature space does not align with material perception.

On the other side, the human similarity network strongly correlates materials that have a similar reflectance. We see a clear distinction between metallic and non metallic materials and big blocks for the completely diffuse materials (top-left block) and highly specular plastics (bottom-right blocks in the non-metallic materials). This shows that the network learned to distinguish between different reflectance properties in a way that correlates with human perception.

The RDM of the BRDF similarity network exhibits a less clear structure and reveals an unevenly distributed space. Diffuse materials appear to be in a small cluster in the space but all other materials are spread out with a majority of large distances. Notably, the network does not make any clear separation between metallic and non metallic materials, but seems to cluster materials with a similar glossiness (very glossy plastics and very glossy metals).

AGREEMENT WITH HUMAN PERCEPTION We measure the agreement of the distances in feature space against human judgements in the same way as explained in Section 3.5, both for color and for gray-scale images. Although color seems to play a minor role for the two networks trained on material similarity, that is not the case for the classification network where accuracy increases significantly with gray-scale images (59% vs. 68% agreement). This network thus seems to use colors in a way that human did not use when making their judgement. Notably, the BRDF similarity network (80% agreement) yields very similar results to the human similarity network (82% agreement). In Section 3.5, it was shown that using the BRDF metric, by itself, aligns only 67% of the time with human judgements. However, the network trained with this metric creates a space that reflects human judgement similar

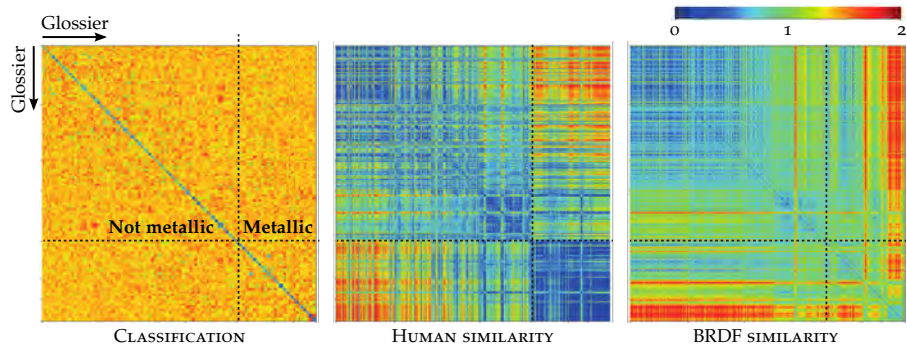


Figure 3.16: The RDMs are organized by reflectance for the three networks. The x - and y -axis are organized by surface reflectance properties of increasing glossiness: starting with diffuse materials to very glossy plastics, and following with metals. The separation between metallic and non metallic materials is depicted by the dotted black line.

to the one trained to directly mimic human perception. This suggests that the structure of the deep network leads to a representation that aligns well with human perception, as soon as the network is fed with a metric that sufficiently represents human judgement.

3.8 DISCUSSION

We have presented and validated a model of material appearance similarity that correlates with the human perception of similarity. Our results suggest that a shared perception of material appearance does exist, and we have shown a number of applications using our metric. Nevertheless, material perception poses many challenges; as such there are many exciting topics not fully investigated in this chapter. Several factors come into play that influence material appearance, i.e., the visual impression of a material, in a highly complex manner; fully identifying them and understanding their complex interactions is an open, fundamental problem. As a consequence of these interactions, the same material (e.g., plastic) may have very diverse visual appearances, whereas two samples of the same material may look very different under different illumination conditions [320, 86]. In aiming for material appearance similarity, we aim for a material similarity metric that can predict human judgements. There is a distinction, common in fields like psychology or vision science, between the distal stimulus—the physical properties of the material—, and the proximal stimulus—the image that is the input to perception—. The key observation here is that human perceptual judgements usually lie between these two, and our training framework and loss function are designed to take both into account. We combine the information about the physical properties of the material contained in the images, by having the same material under different geometries and illuminations, with the human answers on appearance similarity. In other words, a pure image similarity metric would not be able to generalize across shape, lighting or color, while a BRDF-based metric would be unable to predict human similarity judgements.

We do not attempt to identify nor classify materials (Figure 3.17). Our loss function could, however, incorporate additional terms (such as the cross-entropy and batch-mining triplet loss term discussed in the appendix) to help with classification tasks. We have carried out some tests and found

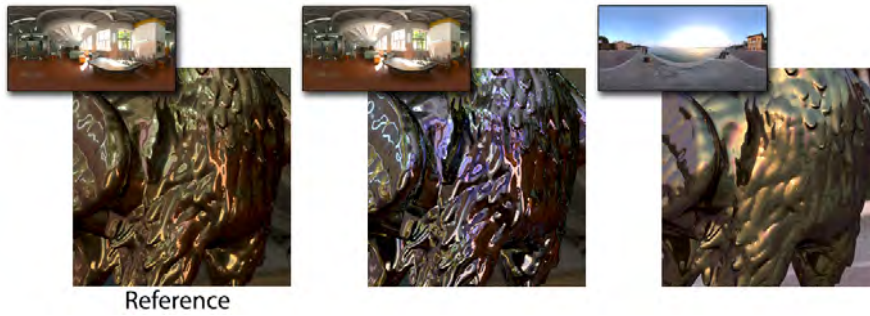


Figure 3.17: In the feature space defined by our model, the middle image (*chrome*) is closer in appearance to the reference (*brass*) than the image on the right (*brass*). The insets show the environment maps used. Our model is driven by appearance similarity, and does not attempt to classify materials.



Figure 3.18: Results using highly heterogeneous materials from the FMD dataset. We show the three closest results returned by our model, from the reference materials highlighted in violet. Note that the search was performed across all three categories shown, not within each category.

anecdotal evidence of this, but a thorough analysis requires a separate study not covered in this chapter.

Despite having trained our model on isotropic materials, we have found that it may also yield reasonable results with higher-dimensional inputs. Figure 3.18 shows three examples from the Flickr Material Database (FMD) [280], which contains captured images of highly heterogeneous materials. We have gathered all the materials from the *fabrics*, *metals*, and *plastics* categories in the database; taking one reference from each, we show the three closest results returned by our model, using an L2 norm distance in feature space. Images were resized to match the model’s input size, with no further preprocessing. Note that the search was not performed within each category but across all three, yet our model successfully finds similar materials for each reference. This is a remarkable, promising result; however, a more comprehensive analysis of in-the-wild, heterogeneous materials is out of the scope of this chapter.

We have also tested the performance of our model on grayscale images. In this case, we have repeated the evaluation conducted in Table 3.1 for our model, using grayscale counterparts of the images. Despite the removal of color information, we obtain results similar to those of our model on color images: A raw accuracy of 72.55 (vs 73.97 on color images), a majority accuracy of 78.64 (vs 80.69), a raw perplexity of 1.82 (vs 1.74), and a majority perplexity of 1.67 (vs 1.55). This further enforces the idea that we learn a measure of appearance similarity, and not image similarity.

To collect similarity data for material appearance, we have followed an adaptive sampling scheme [303]; following a different sampling strategy may translate into additional discriminative power and further improve our results. Our model could potentially be used as a feature extractor, or as a baseline for transfer-learning [282, 340] in other material perception tasks. A larger database could translate into an improvement of our model's predictions; upcoming databases of complex measured materials (e.g., Dupuy et al. [70]) could be used to expand our training data and lead to a richer and more accurate analysis of appearance. Our methodology for data collection and model training could be useful in these cases. Similarly, upcoming network architectures that may outperform our ResNet choice could be adopted within our framework. Finding hand-engineered features could also be an option and may increase interpretability, but it could also introduce bias in the estimation.

In addition to the applications we have shown, we hope that our work can inspire additional research and different applications. For instance, our model could be of use for designing computational fabrication techniques that take into account perceived appearance. It could also be used as a distance metric for fitting measured BRDFs to analytical models, or even to derive new parametric models that better convey the appearance of real world materials. We have made our data available for further experimentation, in order to facilitate the exploration of all these possibilities.

Selecting an optimal set of icons is a crucial step in the pipeline of visual design. It helps to structure and navigate through content, however, designing icons sets is usually a difficult task for which expert knowledge is required. To ease the process of icon set selection to the users, this chapter introduces a similarity metric which captures the properties of style and visual identity. We train a Siamese Neural Network with an on-line dataset of icons organized in visually coherent collections that are used to adaptively sample training data and optimize the training process. As the dataset contains noise, we further collect human-rated information on the perception of icons' similarity which will be used for evaluating the proposed model. We showcase several applications based on searches, kernel visualizations, and optimized set proposals that could be helpful while exploring large collections of icons.

This work was published in *Multimedia Tools and Applications* [170]. While I was the leading author in this project, Elena Garces and Diego Gutierrez helped by supervising my work, giving invaluable feedback, and on the manuscript text and figures.

M. Lagunas, E. Garces, & D. Gutierrez
Learning Icons Appearance Similarity
 Multimedia Tools and Applications, Vol 78 (8), 2019

4.1 INTRODUCTION

Visual communication is one of the most important ways to share and transmit information [203, 202]. In the same way as words are used for verbal communication, symbols or icons are the elements used to convey information in a universal and ubiquitous language [6, 131]. Icons are key elements to structure visual content and make it more appealing and comprehensible. Thus, finding the optimal set of icons is a very delicate task usually done by expert designers which involves semantic, aesthetic, and usability criteria. Recent works aim at automatizing this task and make it more accessible to the general public [18, 277, 278, 217], either by providing a unified icon representation and rules, such as Google Materials⁷, or with online datasets such as The Noun Project⁸ with more than one million elements. While these datasets are undoubtedly useful, they can be hard to explore due to their magnitude.

The following properties are desirable for an icon set to be effective: first, being appropriate for the meaning -usually, the icon's designer provide semantic labels. Second, being visually appealing by means of a coherent *style* and a carefully defined *visual identity* [278]. As seen in the literature [6] [15], we define style as the set of pictorial features in the icons such as stroke, fill, or curvature; and visual identity as the property that makes a set of icons visually identifiable and unique, it is a higher-level property usually linked to the shape of the object. Previous works have studied style in fonts [232], clip art [94], or infographics [263]. Although the definition of style for these domains shares certain properties with icons style, e.g. strokes, fills, or corner smoothness; icons have additional characteristics that make them unique and visually identifiable, and these are not taken into account in the existing metrics. For example, in Figure 4.1, the collections *notebooks* and *bags* have a different visual identity while their pictorial style can be considered similar.

⁷Explanation of Google's design language:
<https://material.google.com/>

⁸Icons database The Noun Project:
<https://thenounproject.com/>

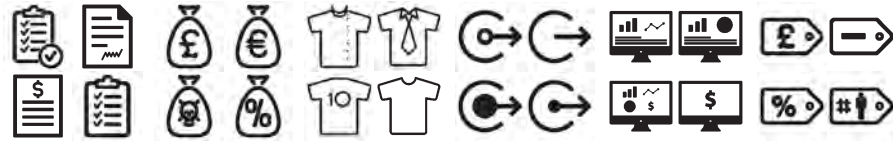


Figure 4.1: Example of six different collections of the dataset. Style and visual identity are preserved for each collection. From left to right, we see the collections labeled as: *notebook*, *bags*, *t-shirt*, *circle-arrow*, *monitor*, and *label*.

Note that each icon also has a unique semantic meaning independent of the collection's name.

On the other hand, the problem of choosing optimal icon sets is a recent topic of research. Previous works [59] [74] have proposed perceptual kernels for predefined icon sets based on crowdsourced data. These techniques learn directly a similarity matrix (or kernel) strictly for the icon selection. As they do not find a new low-level feature space for each icon, these techniques are not able to generalize outside the initial sample space of ten or twenty icons.

In this chapter, we present a learning-based similarity metric that captures the properties of style and visual identity for iconography. Our main contributions are:

- We present an icon dataset labeled by designers where each collection shares a coherent style and visual identity.
- We learn icons' appearance similarity using a Siamese Neural Network with a triplet loss function and adaptive sampling trained from our weakly-labeled dataset and evaluated with human ratings.
- We propose several applications including search by similarity and a method to create icon sets optimized for style and visual identity in order to help users on user-interface design tasks.
- We collect annotated ratings on the perception of appearance similarity for iconography.

We greedily gather an icon dataset from the Noun Project online database. Since the semantics of each icon is highly attached to the application, we assume that each icon is labeled with a keyword that represents its concept properly. The icons in this dataset are organized in collections, which share a style and have a particular visual identity (see Figure 4.1). As previous methods do not fully consider the pictorial properties of icons, we use the collected dataset to train a new Siamese Neuronal Network by adaptively sampling meaningful triplets of relative comparisons. However, as the labeling of the collections is very noisy, -there is no unified and homogeneous label set that we can completely trust- we need to gather new reliable data for testing the model. We numerically evaluate the performance of our distance metric on this test data, and compare its performance to existing similarity metrics. Finally, we propose an application to optimize icon sets for the properties of style and visual identity that can be used as a tool to help users while designing graphical interfaces. To validate the method we launch a crowdsourced survey to a group of 25 human-raters with experience in Computer Graphics or Graphic Design. Users reported that our method returns a set of icons sharing a representative appearance 75.25% of the times, while random icon sets share a representative appearance 29% of the times.

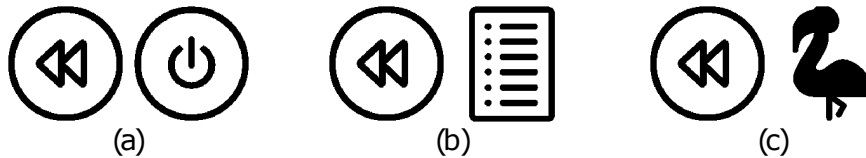


Figure 4.2: Examples of similarity between icons. (a) Icons with similar style and visual identity. Note that both icons have rounded shapes and medium-thick lines. (b) Icons with similar style yet different identity, one has rounded shape while the other is a rectangle. (c) Icons whose style is different and they also have different identities.

4.2 PROBLEM DEFINITION

Our main goal is to obtain a metric to measure style similarity and visual identity between icons. As mentioned in Section 4.1, an icon can be defined by its pictorial properties like outline stroke, fill or curvature [26], features that conform the pictorial style of the icon. In addition, a set of icons is also characterized by a particular visual identity [15] [6], i.e. one or more properties that make it unique and visually identifiable. Commonly, these properties relate to a particular shape or a motif, which repeats between icons of the same collection e.g. a silhouette circle, a notebook-like shape, an arrow, etc. (see Figure 4.1).

Finding clusters of perceptually different icon sets is really impractical given the subtle differences between them. Instead, as seen in previous work [94, 232, 192, 263], it is more intuitive to find a continuous metric space where the distances between the icons correspond to distances in the perceived similarity. Given that previous definitions of style use hand-crafted features for other domains that do not apply for icons, we aim to find a new similarity metric \mathcal{D} that measures differences in style and differences in visual identity:

$$\mathcal{D}(i, j) = \mathcal{D}_s(i, j) + \mathcal{D}_v(i, j) \quad (4.1)$$

where (i, j) is a pair of icons, the function $\mathcal{D}_s(i, j) \in \mathbb{R}^+$ measures style similarity, and the function $\mathcal{D}_v(i, j) \in \mathbb{R}^+$ measures visual identity. For icons with similar style and visual identity, \mathcal{D} should return small values, i.e. $\mathcal{D}_s \simeq 0$ and $\mathcal{D}_v \simeq 0$ (Figure 4.2, a). For icons with similar style but with different identity, $\mathcal{D} = \mathcal{D}_v$ (Figure 4.2, b). Finally, for icons where both properties are very different, the similarity function will also have a high value; $\mathcal{D} \gg 0$ with $\mathcal{D}_v \gg 0$ and $\mathcal{D}_s \gg 0$ (Figure 4.2, c).

4.2.1 Overview

An overview of the method can be seen in Figure 4.3. Our main goal is to obtain a similarity metric $\mathcal{D}(i, j)$ where i, j are a pair of icons. To train the similarity metric, we use a dataset which is annotated by icon designers. Since there is no unified way of labeling, we cannot completely trust the annotations and we might find noise in some of its classes. This kind of datasets are called *weakly labeled* and additional efforts are required to work with them. In our case, part of the dataset is used to launch crowd-sourcing surveys and gather human-ratings that will allow us to test and compare the proposed models (Section 4.3). The other part of the data will serve to train a Siamese Neural Network (SNN) to work as the similarity metric (Section 4.4).

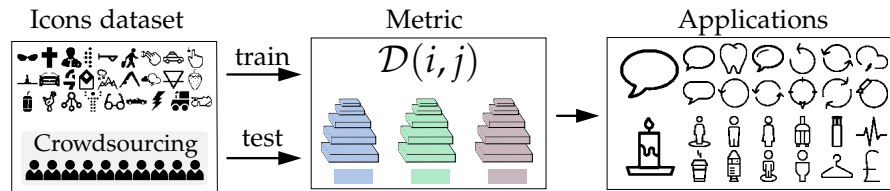


Figure 4.3: Overview of our framework: The leftmost part shows the data gathering process. First, we collect a dataset of icons and use it to train the similarity metric. Since the dataset contains icons labeled by the designers, we cannot completely trust their annotations and might find spurious data or noise. Due to that, we use part of the data gathered to launch crowdsourcing experiments in Amazon Mechanical Turk and obtain curated test data that we use to compare the trained models. Once the data is collected, we train a Siamese Neural Network (SNN) that works as our distance metric, returning small values for icons that share style and visual identity while returning large values for icons that do not share those properties. With the trained model we are also able to compare icons distances and perform similarity searches by returning the icons with the minimum distance to a reference in the learned Euclidean space.

The SNN maps the input icons into a new Euclidean feature space where they can be compared. The new mapping of the icons can be further used to propose different applications like searches by similarity, or propose icon sets optimized for the properties of style and visual identity.

The concept of weakly-labeled data might resemble weakly-supervised learning [49, 311, 324]. However, in weakly-supervised learning we have a constrained amount of annotated data, on the other hand, weakly-labeled data has no annotations but we know some meta-information about each sample. Moreover, in weakly-labeled data, we do not have any constraints on the amount of data used during training.

4.3 COLLECTING DATA

We obtain our icon dataset from the *Noun Project* website, which contains thousands of black and white icons uploaded by graphic designers. Using the provided API we greedily downloaded a total of 26027 different icons, grouped in 1212 collections or classes each one sharing a label decided by the author (see Figure 4.1 for a few examples). Each icon belongs to just one class and most of the icons per class share similar style and visual identity properties. As a first step, by means of stratified sampling, we split the dataset into three subsets: training (70%), validation (10%), and test (20%). We consider each class as the strata, then, we randomly select elements from each class proportionally (according to the given percentages) to sample the train, validation and test subsets. All the elements in each class are sampled and the subsets are mutually exclusive, meaning that each element is sampled only once and for one of the subsets. However, the labels provided by the designers are not disjoint and we might find different labels with the same style and identity and one label with different styles or identity. This kind of *weakly-labeled* [286] data may yield problems like not detecting if the model has overfitting or not allowing a fair comparison with other architectures at testing time. Thus, further data collection and adjustments are needed to take full advantage of the dataset.

COLLECTING CURATED DATA We collect valid data on the perception of icon’s similarity that will be used to test the proposed models and select the best one. We use *Amazon Mechanical Turk (MTurk)* to launch the experiments. Similar to previous works [94] [21] [192], we gathered data in the form of *relative comparisons*, since they are more robust and easier for human raters than Likert ratings [59] [259]. The structure of each test, or HIT, consisted of: first, a clear description of the task that human raters had to perform, then, a training phase where we show a small set of four manually picked relative comparisons displaying guidance messages if the user fails answering correctly. The last part corresponds to the test phase, where the rater has to answer a total of 60 relative comparisons where seven questions belong to a manually selected control set with an obvious answer. The duration of each HIT was approximately seven minutes, and all successful participants were paid.

We rejected all human raters that had more than one error (out of seven) in the control questions. In the end, we launched 6000 relative comparisons tests each of them answered by ten users, 962 HITs were approved and 38 rejected. To create the relative comparisons for each question, we randomly selected one icon per class from three different random classes. We allowed participants to do as many HITs as they wanted without repetition. A total of 213 users took part in the survey, 43% female. Among raters, 5.95% claimed some professional experience in user interface and interaction design, while 6.43% have had some professional experience with graphic design.

4.4 MODELING VISUAL APPEARANCE OF ICONS

Existing style similarity metrics [94, 263] use a handcrafted feature space only suitable for their respective domains, where only local style features are taken into account. On the contrary, besides style, our metric should measure also visual identity, which is usually a higher-level property related to the shape of the icon. Image similarity has been measured with existing deep models, such as VGG19 [288], pre-trained on natural images; and fine-tuning these networks has worked well for tasks such as interior design similarity [21]. However, we would need a huge amount of training data to improve the performance of any existing network, and given that our domain is much simpler than pictures of natural images, we choose to train a new network with our data. To make sense of the difference, the widely used network VGG19 has 144M of parameters, while our network has 47M parameters.

We use a Siamese Neural Network [36, 243, 270] consisting in three identical Convolutional Neural Networks (CNN) that share their parameters. This kind of architecture is really powerful for learning a new Euclidean space [270, 243, 229] where objects can be compared [36, 339]. Since the icons inside a collection in the dataset share the properties of style and visual identity, the SNN can be trained to map together the icons that share these properties while it separates icons with different style and visual identity. Each CNN has four convolutional layers that are followed by a batch normalization [137] layer and a max-pooling layer. The last pooling layer is connected to the linear classifier. The linear classifier contains three fully-connected layers where the first two have 4096 and 1024 features respectively. The last layer represents the final embedding $f(x)$ of the image x into the new feature space \mathbb{R}^d , where the value of d has been empirically set to 256. We also included two dropout [294] layers between the fully-connected ones with a dropout regularization rate of 30%. An example of the architecture we

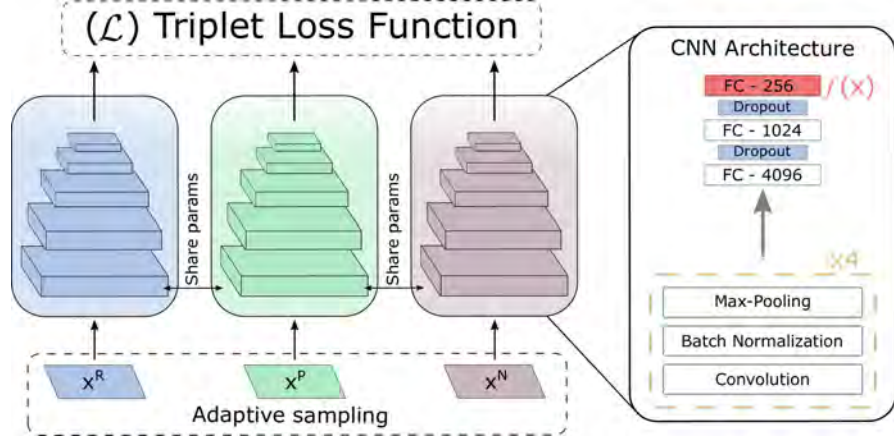


Figure 4.4: Architecture proposed to measure icons similarity. The Siamese Network has three inputs: Reference (x^R), Positive (x^P) and Negative (x^N); and three Convolutional Neural Networks (CNN) to obtain its embeddings ($f(x)$). With these three embeddings, we can compute the error of the network (\mathcal{L}) using the triplet loss function described in Equation 4.2. The CNNs share the same structure and parameters. Each of them has four convolutional layers, that are followed by a batch-normalization layer and a max-pooling layer. The last pooling layer is connected to a linear classifier with three fully connected layers (FC). First FC has 4096 features, second one has 1024, while the last FC has only 256, furthermore, last FC of each CNN corresponds to the embedding $f(x)$ of the input triplet $[x^R, x^P, x^N]$. Between the FC layers there are dropouts with regularization rate of 30%.

described is shown in the Figure 4.4, right. This architecture is trained using triplets of images: a reference x^R , a positive x^P (icon with similar properties to the reference), and a negative x^N (icon with different properties to the reference). To train the network we design a specific loss function which is explained below.

4.4.1 The Loss Function

We consider the output of the last fully-connected layer of the Convolutional Neural Network as an embedding $f(x) \in \mathbb{R}^d$ with input x . The embedding represents x in a new d -dimensional Euclidean space. Since we have a Siamese Neural Network formed by three CNNs that are identical with three inputs $[x^R, x^P, x^N]$, we get three embeddings as the output $[f(x^R), f(x^P), f(x^N)]$ where $f(x^R)$ corresponds to the embedding of a reference input while $f(x^P)$ is the embedding of an input of the same class as the reference and $f(x^N)$ is the input of an image that does not belong to the same class as the reference. We want to ensure that a reference icon x^R is closer to every icon of the same perceptual similarity (style and visual identity) x^P , than to the rest of icons with different image properties x^N . Thus the triplet loss function \mathcal{L} (Equation 4.2) has to ensure that the distance in the d -dimensional Euclidean space between the reference and the positive icon is minimum while it is large between the reference and the negative icon [270, 243].

$$\mathcal{L} = \sum_{i=1}^M \left[\|f(x_i^R) - f(x_i^P)\|_2^2 - \|f(x_i^R) - f(x_i^N)\|_2^2 + \alpha \right]_+ \quad (4.2)$$



Figure 4.5: Examples of the triplets sampled during training. The variables x^R , x^P and x^N refers to the reference, positive and negative icon respectively. The positive icon and the reference are selected from the same class and they have the larger Euclidean distance among the icons inside that class. The negative icon has the shorter Euclidean distance to the reference among the icons within a different randomly selected class.

Here M is the training set of triplets and α is a margin enforced between negative and positive pairs which was empirically set to 0.2. The value α prevents the function from evaluating to zero in cases where the distance between the reference and the negative sample is larger than the reference and the positive sample, thus letting it find larger margins while training.

ADAPTIVE SAMPLING If we would like to create all the possible triplets from the, approximately, 18200 icons in the training set we would have $\binom{18200}{3} \simeq 6.027 \cdot 10^{12}$ possible combinations, an unmanageable number using a standard desktop configuration. Furthermore, most of the generated triplets would easily satisfy the constraints of the loss function and not contribute to the training process at all, thus slowing it. For this reason, following the approach of Schroff et al. [270], we generate the triplets on the fly during the training process, selecting the ones that are active and help in the convergence. We generate triplets that violate the most the constraints imposed by the loss function. To do so, we randomly select one icon from the training set as the *reference*, then, we select the *positive* sample as the icon from the same class with the maximum distance in the Euclidean space to the reference: $\operatorname{argmax}_{x_i^P} \|f(x_i^R) - f(x_i^P)\|_2^2$. To obtain the negative icon, we randomly pick a different class and select the icon that has the minimum distance to the reference: $\operatorname{argmin}_{x_i^N} \|f(x_i^R) - f(x_i^N)\|_2^2$. We repeat this approach until a considerable number of triplets without repetition has been obtained. This process is applied before every epoch and it requires to compute the embedding for every icon at each iteration. In the first iteration, embeddings are directly obtained from the network whose parameters have been set using Xavier’s initialization [107]. Although it increases the training time, it also ensures that all input triplets are meaningful for the training. Figure 4.5 shows an example of the triplets sampled during training in the first iteration.

4.4.2 Training the Model

We use ADAM optimization [162] and the triplet sampling explained in Section 4.4.1. The mini-batch had a size of sixteen images and to update

the parameters of the network we use standard back-propagation [178, 108]. At training time, we perform two sequential operations with each image before feeding it to the network: first, *data augmentation* (randomly rotating or flipping the image) and second, *random crops*. For the crops, we randomly perform a crop of size 180x180 aligned to the corners in the original image, with size 200x200. We started the training with a learning rate of 10^{-4} that was reduced every 60 epochs by a factor of ten to let the model converge. To create the validation set we also use the adaptive sampling, moreover, each image is scaled to 180×180 instead of cropped and no data augmentation is applied. We need around two days and 140 epochs to train the model.

4.5 EVALUATION

We evaluate the performance of the models by comparing their *precision* and *perplexity* on the gathered data from the MTurk HITs. At testing time, no data augmentation is applied and the inputs are directly scaled to 180×180 without cropping. First, we obtain the embedding for the three inputs of the triplet $[f(x^R), f(x^P), f(x^N)]$, since they are in a 256-dimensional Euclidean space, we can calculate the Euclidean distance of each icon with respect to the reference $\mathcal{D}(x^R, x^P)$ and $\mathcal{D}(x^R, x^N)$. Actually, if we want to obtain the probability of choosing the icon x^P over x^N , what we are aiming to obtain is a function of similarity instead of a distance, thus we define the similarity between two icons $s(x^R, x^P)$ as:

$$s(x^R, x^P) = \frac{1}{1 + \mathcal{D}(x^R, x^P)}, \quad (4.3)$$

when the positive x^P and reference x^R icon are completely similar $\mathcal{D}(x^R, x^P) = 0$, their similarity is $s(x^R, x^P) = 1$. In the opposite case, if the pair of icons is completely dissimilar: $s(x^R, x^P) = 0$. Knowing that $\mathcal{D}(x^R, x^P)$ cannot be infinity, we can define the probability of choosing the icon x^P against x^N as:

$$\mathbb{P}(x^P) = \frac{s(x^R, x^P)}{s(x^R, x^P) + s(x^R, x^N)} \quad (4.4)$$

We can obtain $\mathbb{P}(x^N)$ similarly. Then, we compute precision and perplexity in two ways: assuming the correct answer relies on each turker opinion separately (*raw*) or assuming the majority opinion is the correct one (*majority*). We also compare our results with two baselines previously calculated: the Humans and the Oracle precision. To compute Humans baseline, we count the rater’s opinion and compare it to the majority. For the Oracle baseline, we count the opinion of the majority on each relative comparison, being the precision always one.

The precision \mathcal{P} tells us the percentage of icons that the model has predicted correctly according to our two criteria (*raw* and *majority*). The precision value is computed as:

$$\mathcal{P} = \frac{\text{Icons correctly predicted}}{\text{Number of total relative comparisons}} \quad (4.5)$$

The perplexity \mathcal{Q} is often used for measuring the usefulness of a model when predicting a sample. Its value is 1 when the model makes perfect predictions on every sample, while its value is 2 when the output is 0.5 for every sample, meaning total uncertainty. We define the perplexity of our model as

$$\mathcal{Q} = 2^{\left(-\frac{1}{M} \sum_{i=1}^M \log_2 \mathbb{P}(x_i^P)\right)} \quad (4.6)$$

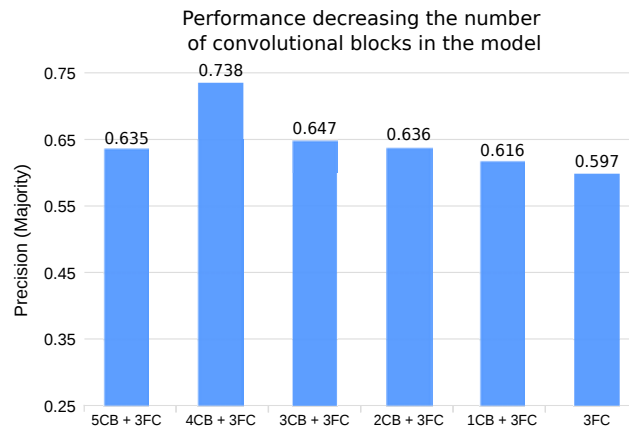


Figure 4.6: Model performance while varying the number of layers. The vertical axis shows the majority precision obtained while the horizontal axis shows the model description. In the models description, *CB* refers to the convolutional Blocks and *FC* to the Fully Connected layers. We can observe how the best model has four convolutional Blocks achieving nearly 74% majority precision. The models with less number of layers and parameters are not able to reach that performance. Also, the model with five convolutional blocks seems to overfit getting similar performance to the model with just two convolutional blocks.

To know which one is the positive sample x^P in the relative comparison we rely on raw and majority criteria as for the precision. The value $\mathbb{P}(x^P)$ will be the probability given by the model using Equation 4.4, M corresponds to the number of triplets we use for testing.

4.5.1 Other Architectures

We followed an incremental approach while designing the Siamese Neural Network. We tested out how the number of convolutional blocks (CB) affects model performance while keeping the same training parameters and same layers in each block (Convolution + Batch norm. + Pooling). Figure 4.6 shows how model performance varies, achieving best results with 4 convolutional blocks.

Once we know that the best accuracy is obtained with four convolutional blocks, we explore the performance varying the layers inside each block and the number of Fully Connected layers. Table 4.1 shows the precision and perplexity of the architectures described below. All the included architectures have four convolutional blocks. *Model-A* has max-pooling between the convolutions and two fully-connected (FC) layers. It has one of the worst results since it does not include layers to avoid overfitting or improve performance with non-linearities. *Model-B* includes max-pooling between convolutions and dropout between the two FC layers. The architecture is similar to *Model-A* and its result is the worst in terms of both, precision and perplexity. *Model-C* includes only max-pooling between convolutions and has three fully-connected layers with dropout between them. The new FC layer does not improve the performance of this model and its results remain lower in comparison to *Model-C*. Finally, *Model-D* includes max-pooling, batch-normalization and ReLUs between convolutions and it also has dropout between the three FC layers yet it does not improve the performance of *Model-D*.

Model	Precision (P)		Perplexity (Q)	
	Raw	Majority	Raw	Majority
Humans	0.771	0.842	-	-
Oracle	0.859	1	-	-
Garces [94]	0.609	0.627	1.578	1.591
VGG19 [288]	0.639	0.654	1.558	1.571
Model-A	0.519	0.521	1.603	1.617
Model-B	0.508	0.507	1.608	1.622
Model-C	0.671	0.702	1.543	1.556
Model-D	0.667	0.699	1.515	1.527
Best model	0.706	0.738	1.555	1.568

Table 4.1: Comparison of the precision and perplexity of different models and methods. We can observe how the chosen method outperforms the rest comparing the precision and it is the closest one to the human ratings. On the other hand, perplexity values are highly dependent on the formula used to obtain probabilities from distances, while precision only depends on turker’s answers. Due to that, our decision on choosing the best model has been more influenced by the results on the precision.

4.5.2 Comparing Against Previous Work

In Table 4.1 we also compare our best model with a well-known pretrained architecture VGG19 [288] and a hand-crafted feature vector for clip art style [94]. VGG19 model is able to achieve 63% of precision yet it was not designed to find a space where icons can be compared by similarity and its results are worse than most of the trained architectures. Also, the time needed to get the feature vector of an image is nearly two orders of magnitude higher than with our model, that just needs $9 * 10^{-4}$ seconds. The method of Garces et al. achieves worse accuracy than VGG19 and our model since the hand-crafted feature space was designed to measure style similarity in their specific dataset and it is not capable to model visual identity. Moreover, it is significantly slower than our method, using several seconds to compute the descriptors of an image.

In the end, *Model-C* outperforms other Convolutional Block configurations we tried and the previous works in terms of precision. Also, it is the closest one to the Human and Oracle baselines. Although our model has one of the best perplexity value, other architectures like *Model-D* and *Model-C* outperform it. The perplexity is computed using the probability of choosing x^P over x^N as the similar icon to x^R , that’s why its value is highly dependent on the formula used to compute the probability \mathbb{P} from a distance \mathcal{D} . Due to that, we trust more the values of the precision when choosing our model while we still consider the perplexity.

4.6 RESULTS AND APPLICATIONS

The trained Siamese Neural Network is capable to produce high-quality embeddings in a new Euclidean feature space which considers the properties of style and visual identity. We can visualize this space in 2D by using



Figure 4.7: Visualization created using the t-SNE algorithm. It reduces the dimensionality of the feature vectors that our model learns to a two-dimensional Cartesian space. Note how icons with similar appearance are grouped in the same regions.

non-linear dimensionality reduction techniques, such as t-SNE [315]. Results can be seen in Figure 4.7.

COMPARISON WITH PERCEPTUAL KERNELS As we show in Equation 4.1, for the same style, our metric measures the difference in visual identity, and, usually, this difference is linked to the shape of the object. Thus, we compare our metric with the perceptual kernel of Demiralp et al. [59] which is optimized for shape similarity (Figure 4.8 (a)). We take the same set of ten gray-scale icons, use our metric to compute the distances and normalize them between 0-1 range to obtain the matrix in Figure 4.8 (b). We also show in (c), and (d) the icons with maximum distances with Demiralp’s kernel and our distance \mathcal{D} , respectively. We observe that, although the results differ a little, both metrics perform very well in maximizing perceptual similarity. However, as opposed to Demiralp et al. work, our metric can be used with any input icon, while their kernel is strictly computed for that set of given icons. We additionally show in Figure 4.8 (e) the icons with maximum distances in our whole dataset. Note that differences in style and visual identity are maximal.

SEARCH BY SIMILARITY Our distance metric allows search by similarity. Given a query icon, we can search the k-nearest neighbors over the entire icon dataset. Results are shown in Figure 4.9. We compare our results with the output given by the method presented by Garces et al. [94] and the pretrained network VGG19 [288]. We can notice that while Garces et al. performs reasonably well to capture low-level style features like strokes and fills, it fails at higher-level elements, and the visual identity is not captured. This is due to the fact that their hand-crafted feature space does not include any feature to capture shape. The network VGG19 after being trained with millions of images can be used as a powerful image descriptor thanks to the knowledge it acquired regarding image features like contours, textures or shapes. The results of VGG19 seem to have coherent visual identity yet some fail in terms of style (see Figure 4.9 *candle* and *calendar* rows). This imprecision is also observable in the numerical evaluation of Section 4.5.2.

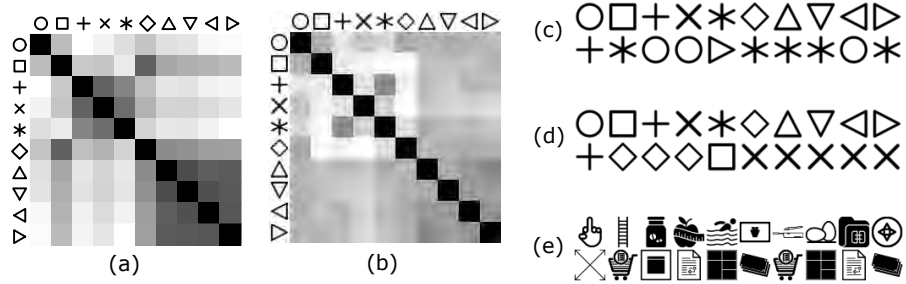


Figure 4.8: Comparison with the shape kernel of Demiralp et al. [59] (darker means more similar). (a) Shape kernel of Demiralp et al. using ten gray-scale icons. (b) Kernel obtained using our metric. Note that, as opposed to Demiralp’s kernel, the triangles using our kernel are not invariant to rotation. In (c) and (d) we show pairs of icons with maximum perceptual distances for Demiralp’s kernel (c) and our metric (d). Our model is capable to return coherent icons with maximum perceptual distance although we did not collect the data with this specific purpose. On the other hand, the method of Demiralp et al. can only be computed for their set of ten icons. (e) Pairs of icons with maximum distances using our whole dataset.

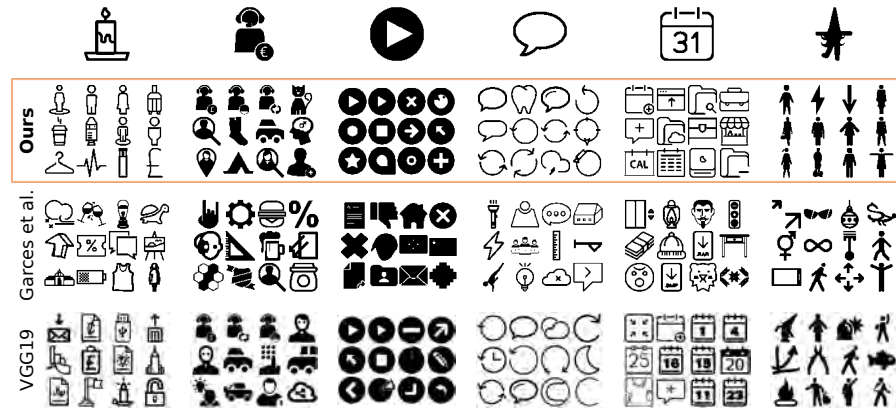


Figure 4.9: The following figure shows the most similar images given a reference and compares them against previous work. We can observe how our method returns visually appealing results considering both style and visual identity, that better represent the input image.

OPTIMIZED ICON SETS Our method can be useful helping designers in creating applications or graphical user interfaces. Given a set of semantic keywords, we can propose icon sets optimized for the properties of style and visual identity. In the example of Figure 4.10, we choose the keywords *animals* (A), *arrows* (B) and *buildings* (C) and we obtain three sets of icons $\{x_A\}$, $\{x_B\}$, $\{x_C\}$ with 36, 112, and 55 elements, respectively. We define a candidate icon set as a triplet $(x_A, x_B, x_C) \in T$, where T is the set containing all the possible combinations of icons for the selected keywords (note that we decided to have triplets as icon sets, but this could arbitrarily grow to icon sets of n elements with $n \in [1, \infty]$). For this case, T contains more than $2 \cdot 10^6$ possible triplets. The goal is to find: $argmin_{i,j,k} \mathcal{D}_{set}(x_{A_i}, x_{B_j}, x_{C_k})$, where $\mathcal{D}_{set}(x_A, x_B, x_C) = \mathcal{D}(x_A, x_B) + \mathcal{D}(x_B, x_C) + \mathcal{D}(x_A, x_C)$. The candidate sets are those whose distances are minimal. As we can see in the figure, the proposed icon sets are highly coherent.

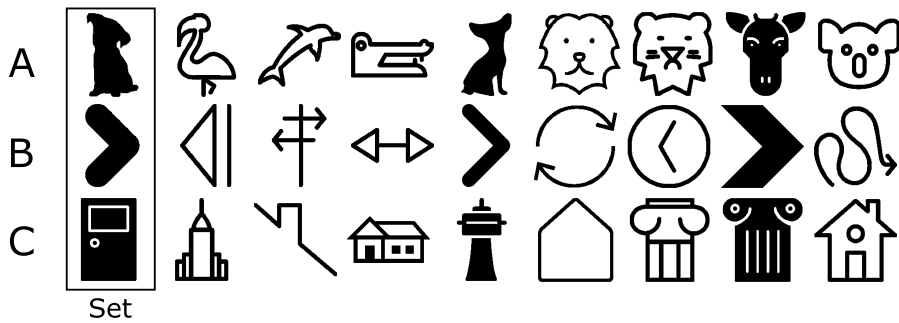


Figure 4.10: General icon set proposal for the keywords: animals (A), arrows (B) and buildings (C). Sets are optimized for the properties of visual identity and style using our method.

To evaluate how useful the proposed *optimized icon sets* are, we gather subjective judgements from annotation experts. We show several optimized icon sets to the rater and ask her two questions: "Do the icons in the set have a representative appearance?". The human-rater can only answer either yes or no. We created 100 sets using the method previously explained and 20 randomly sampling icons. Each survey contains 20 icon sets to be evaluated, 16 randomly sampled from the set of 100 created with our method and 4 randomly sampled from the set of random icon sets. Each icon set is made by four icons belonging to four different keywords. The keywords are also randomly sampled from a group of 9 candidates (animals, arrows, buildings, clothes, food, faces, music, humans and documents). Each keyword contains around 80 different icons from the test set with a wide variety of styles and visual identities. The Figure 4.11 shows a screenshot of the test carried out to validate the proposed icon sets. At the end we collected 25 subjective evaluations from raters with previous experience in Computer Graphics or Graphic Design, 8 raters are females and ages range between 20 to 32 years old with an average of 25 years old. Raters thought the visual appearance of the icons is representative within the sets returned by our method 75.25% of the times. On the other hand, raters found the appearance of the set representative only 28% of the times for sets with randomly sampled icons.

4.7 CONCLUSION AND FUTURE WORK

In this chapter, we have presented a model for measuring the properties of style and visual identity in iconography. As opposed to previous works, which only focus on low-level style features, our method is able to model high-level properties of the icons, capturing its visual identity. Our learned model maps each icon into a 256-dimensional feature space which allows direct comparisons by computing Euclidean distances. We have shown that our metric can be used to ease the process of icon set selection for users. Moreover, our approach is generalizable and can be used with any image outside the initial dataset.

There are many avenues for research following our work. The most immediate extension is to take into account color compatibility measures [231] to automatically colorize the icons to a particular color style. Similarity metrics can also be used as a guide to evaluate content generation methods, in our case, our metric could be used in combination with the work of Liu et al. [193] to automatically iconify pictures according to a desired style. In this regard, the success of deep generative methods for style transfer in

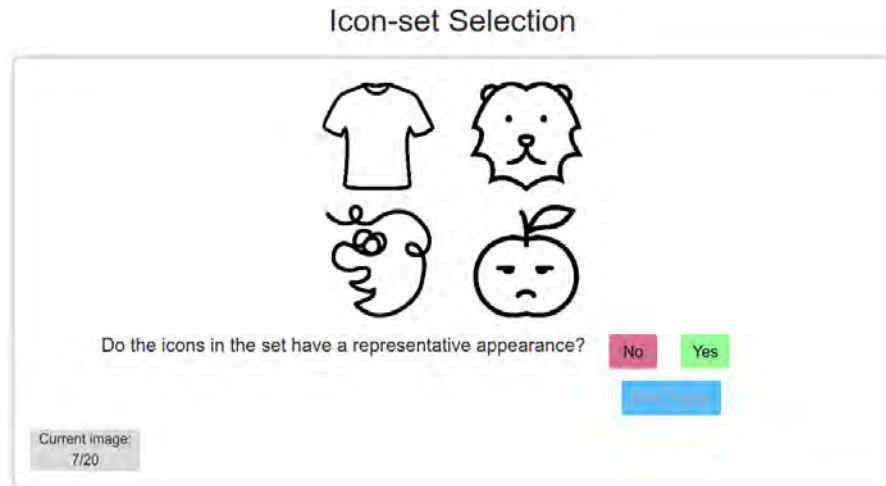


Figure 4.11: Screenshot of the test developed to validate the usefulness of the proposed icon sets. The icon set is made of four icons belonging to the keywords: clothes (top-left), animal (top-right), faces (bottom left) and food (bottom right). Below the images the question appears allowing for a binary answer (yes or no). The blue button goes to the next icon set and on the bottom left corner, white gray background, we can see the progress of the test.

fonts [314] suggests that such kind of techniques could be applied in this domain too. Moreover, Our network could be used in combination with semantic object labeling or object sketches to train better models that take into account object semantics besides depiction.

On the other hand, while CNNs have received a lot of attention for natural images, they are still highly unexplored for graphic designs. Since it is a domain with a simpler underlying representation, in theory, it should require less training data. We also believe that our work can inspire future works in the problem of extracting shape descriptors for 2D images. It is well known that Convolutional Neural Networks capture coarse shapes in the deeper layers of the hierarchy [344], but it is ongoing work to really understand how to disentangle this information to be used as a standalone shape descriptor.

Part IV

CONFOUNDING FACTORS IN MATERIAL PERCEPTION

This part focuses on the effect of confounding factors such as geometry, illumination, or motion in our perception of material appearance. In the first half, we focus on the joint role that geometry and illumination have in humans performance recognizing materials. Our main contribution is a comprehensive study analyzing the effect of such confounding factors, with special emphasis in the frequency domain. Moreover, we analyze the correlation of simple statistics and deep neural networks with human performance. In the second half, we focus on the effect that motion has in a set of material attributes. We launch two experiments, the first one focuses on understanding the role of motion in our perception of a set of attributes while the second one focuses on the brightness attribute for a set of stimuli with different degrees of motion and luminance. Our main contribution is therefore the proposed systematic experiments together with our findings on the effect of motion.

Observing and recognizing materials is a fundamental part of our daily life. Under typical viewing conditions, we are capable of effortlessly identifying the objects that surround us and recognizing the materials they are made of. Nevertheless, understanding the underlying perceptual processes that take place to accurately discern the visual properties of an object is a long-standing problem. In this chapter, we perform a comprehensive and systematic analysis of how the interplay of geometry, illumination, and their spatial frequencies affects human performance on material recognition tasks. We carry out large-scale behavioral experiments where participants are asked to recognize different reference materials among a pool of candidate samples. In the different experiments, we carefully sample the information in the frequency domain of the stimuli. From our analysis, we find significant first-order interactions between the geometry and the illumination, of both the reference and the candidates. In addition, we observe that simple image statistics and higher-order image histograms do not correlate with human performance. Therefore, we perform a high-level comparison of highly non-linear statistics by training a deep neural network on material recognition tasks. Our results show that such models can accurately classify materials, which suggests that they are capable of defining a meaningful representation of material appearance from labeled proximal image data. Last, we find preliminary evidence that these highly non-linear models and humans may use similar high-level factors for material recognition tasks.

While I led this line of work (under the supervision and help of Belén Masiá and Diego Gutiérrez), Ana Serrano collaborated by running the statistical analysis, and helping with the manuscript text and figures. This work was published in the *Journal of Vision* (JoV) [172].

M. Lagunas, A. Serrano D. Gutierrez, & B. Masia
The Joint Role of Geometry and Illumination on Material Recognition
Journal of Vision, Vol 21 (2), 2021

5.1 INTRODUCTION

Under typical viewing conditions, humans are capable of effortlessly recognizing materials and inferring many of their key physical properties, just by briefly looking at them. While this is almost an effortless process, it is not a trivial task. The image that is input to our visual system results from a complex combination of the surface geometry, the reflectance of the material, the distribution of lights in the environment, and the observer's point of view. To recognize the material of a surface while being invariant to other factors of the scene, our visual system carries out an underlying perceptual process that is not yet fully understood [67, 85, 2].

Then, *how does our brain recognize materials?* We could think that, similar to solving an inverse optics problem, our brain is estimating the physical properties of each material [250]. This would imply knowledge of many other physical quantities about the object and its surrounding scene, from which our brain could disentangle the reflectance of the surface. However, we rarely have access to such precise information, so variations based on Bayesian inference have been proposed [154].

Other approaches are based on image statistics, and explain material recognition as a process where our brain extracts image features that are relevant

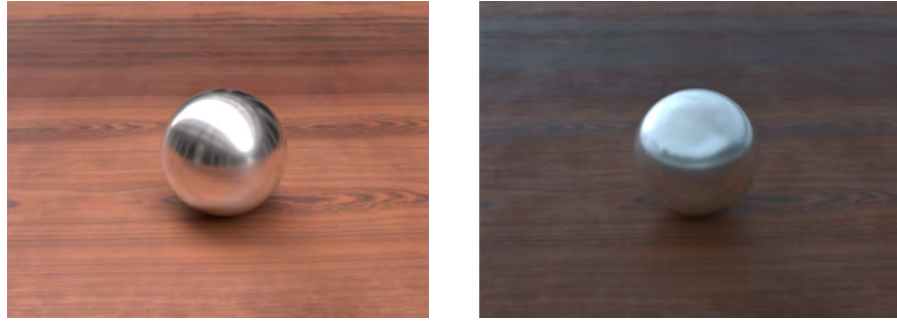


Figure 5.1: Two spheres made of silver, under two different illuminations, leading to completely different pixel-level statistics.

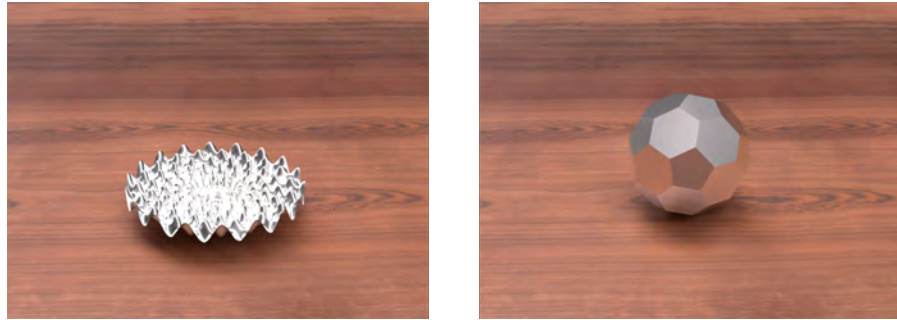


Figure 5.2: Two objects of different geometries but made of the same material, under the same illumination. The object on the left appears to be made of a shinier material.

to describe materials. Then, it would try to match them with previously acquired knowledge, in order to discern the material we are observing. In this approach our visual system would disregard the illumination, motion, or other factors in the scene and try to recognize materials by representing their typical appearance in terms of features instead of explicitly acquiring an accurate physical description of each factor. This type of image analysis can be carried out in the primary domain [100, 4, 82, 218, 227], or in the frequency domain [234, 35, 103]. However, it is argued if our visual system actually derives any aspects of material perception from such simple statistics [10]. For instance, Fleming and Storrs [89] have recently proposed the idea that highly non-linear encodings of the visual input may better explain the underlying processes of material perception.

In this chapter, we thoroughly analyze how the confounding effects of illumination and geometry influence human performance in material recognition tasks. The same material can yield different appearances due to changes in illumination and/or geometry (see Figures 5.1 and 5.2), while it is possible to have two different materials look the same by tweaking the two parameters [320]. We aim to further our understanding of the complex interplay between geometry and illumination in material recognition. We have carried out large-scale, rigorous online behavioral experiments where participants were asked to recognize different materials, given images of one reference material and a pool of candidates. By using photorealistic computer graphics, we obtain carefully controlled stimuli, with varying degrees of information in the frequency domain. In addition, we observe that simple image statistics, image histograms, and histograms of V1-like subband filters do not correlate with human performance in material recognition tasks. Inspired by Fleming

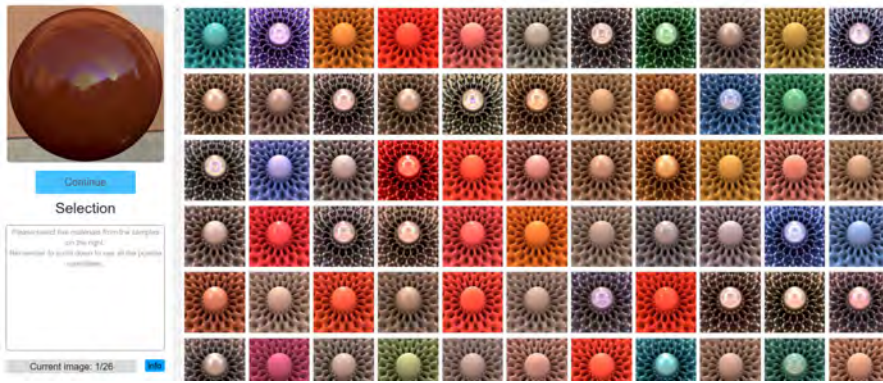


Figure 5.3: Graphical user interface of the online behavioral experiments. In particular, this screenshot belongs to the TEST SH. On the left, the user can see the reference material together with her current selection. On the right, she can observe all the candidate materials. To select one candidate material, the user clicks on the corresponding image and it is automatically added to the selection box on the left.

and Storrs’ recent work [89], we analyze highly non-linear statistics by training a deep neural network. We observe that such statistics define a robust and accurate representation of material appearance and find preliminary evidence that these models and humans may share similar high-level factors when recognizing materials.

5.2 METHODS

We carry out a set of online behavioral experiments where we analyze the influence of geometry, illumination, and their frequencies in human performance for material recognition tasks. Participants are presented with a reference material and their main task is to pick five materials from a pool of candidates that they think are closer to the reference. A screenshot of the experiment can be seen in Figure 5.3.

5.2.1 Stimuli

We obtain our stimuli from the dataset proposed by Lagunas et al. [171]. This dataset contains images created using photorealistic computer graphics, with 15 different geometries; six different illuminations ranging from real-world indoor scenarios to urban or natural landscapes; and 100 different materials measured from their real-world counterparts which were pooled from MERL database [211]. We sample the following factors for our experiments:

GEOMETRIES Among the geometries that the dataset contains, we choose the sphere and *Havran-2* geometry [118]. These are a low and high spatial frequency geometries, respectively, suitable to test how the spatial frequencies of the geometry affect the final appearance of the material and our performance at recognizing it.

- *Sphere*: Representing a smooth, and low spatial frequency geometry, widely adopted in previous behavioral experiments [77, 142, 153, 300].
- *Havran-2*:¹¹ It is a geometry with high spatial frequencies, and with high spatial variations that has been obtained through optimization

¹¹To simplify the notation we will refer to *Havran-2* geometry as *Havran*.

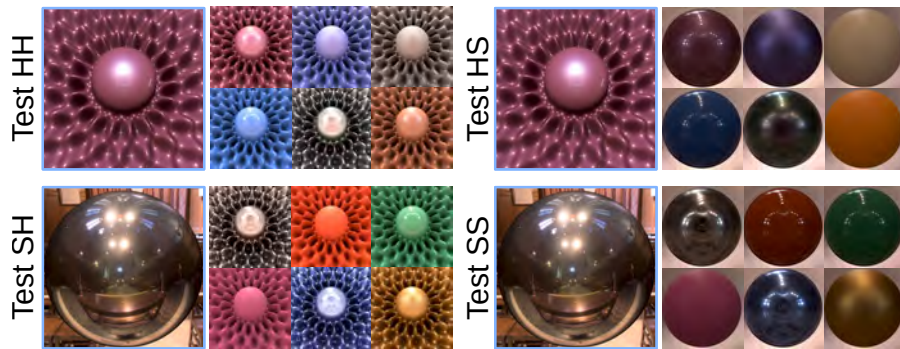


Figure 5.4: Examples of the stimuli in each different online behavioral experiment. On the left we show an example of the reference stimuli with one of the six illuminations. On the right, we show a small subset (six out of the 100 materials) of the candidate stimuli with *St. Peters* illumination.

techniques. *Havran-2* surface has had significant success in recent perceptual studies and applications [171, 113, 322, 111].

The geometry in the reference and candidate samples changes, the stimuli for each different experiment can be observed in Figure 5.4. The details are as follows:

- TEST HH: Both the reference and the candidates depict *Havran* geometry.
- TEST HS: The reference depicts *Havran* and the candidates depict the sphere.
- TEST SH: The reference depicts the sphere while the candidates depict *Havran*.
- TEST SS: Both the reference and the candidates depict the sphere geometry.

ILLUMINATIONS To prevent a pure matching task, we choose different illuminations between the reference and candidate materials for all behavioral experiments.

- The reference samples depict six different illuminations captured from the real-world. All illuminations can be observed in Figure 5.5. To have an intuition of the content in the captured illumination, the insets show the RGB intensity for the horizontal purple line. We use all illuminations in the dataset since they contain a mix of spatial frequencies suitable to empirically test how the spatial frequencies of the illumination may affect human performance on material recognition tasks. The illuminations *Grace*, *Ennis*, and *Uffizi* have a broad spatial frequency spectrum, *Pisa* and *Doge* mainly contain medium and low spatial frequency content, while *Glacier* mainly has low spatial frequency content. To simplify the notation, we will refer to them throughout the chapter as high-frequency, medium-frequency and low-frequency illuminations, respectively.
- The candidate samples depict the *St. Peters* illumination (except in an additional experiment discussed in Section 5.4 where they depict *Doge*

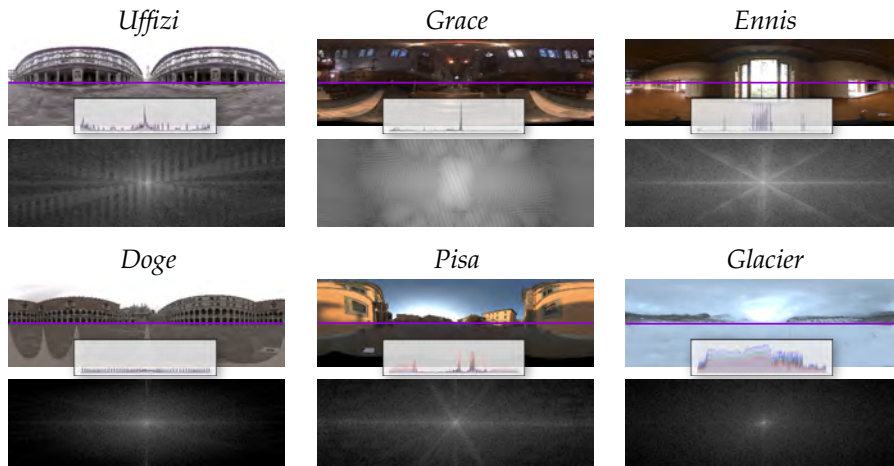


Figure 5.5: All illuminations depicted in the online behavioral experiments together with their frequency spectrum. The inset corresponds to the pixel intensity for the horizontal purple line.

illumination). *St. Peters* is an illumination that has been used in the past for several perceptual studies [86, 276], it can be seen in Figure 5.5. The inset shows the RGB pixel intensity for the horizontal purple line.

To quantify the spatial frequencies of the illuminations, we have employed the high-frequency content measure (HFC) [37]. This measure characterizes the frequencies in a signal by summing linearly-weighted values of the spectral magnitude, thus avoiding to arbitrarily choose a separation between high and low frequencies or visually assessing the slope of the $1/f$ amplitude spectrum. A high HFC value means higher frequencies in the signal. Figure 5.7 shows the HFC for each illumination.

MATERIALS We use all the materials from the Lagunas et al. dataset [171]. The reference trials are sampled uniformly to cover all 100 material samples in the dataset. Examples of the stimuli used in each behavioral experiment are shown in Figure 5.4 where the image on the left shows the reference material and the right area shows a subset of the candidate materials.

5.2.2 Participants

The online behavioral experiments were designed to work across platforms on standard web browsers and they were conducted through the Amazon Mechanical Turk (MTurk) platform. In total, 847 unique users took part in them (368 users belonging to the experiments explained in Section 5.3, and 479 belonging to the additional experiments explained in Section 5.4), 44.61% of them female. Among the participants, 62.47% claimed to be familiar with computer graphics, 25.57% had no previous experience and 9.96% declared themselves professionals. We also sampled data regarding the devices used during the experiments: 94.10% used a monitor, 4.30% used a tablet, and 1.60% used a mobile phone. In addition, the most common screen size was 1366 x 728 px. (42.01% of participants), minimum screen size was 640 x 360 px. (two people), and a maximum of 2560 x 1414 px. (one person). Users were not aware of the purpose of the behavioral experiment.

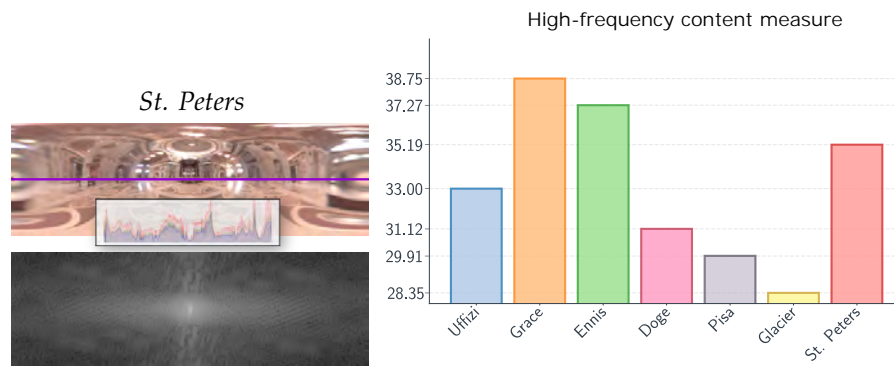


Figure 5.6: Candidate illumination employed in the online behavioral experiments together with its frequency spectrum. The inset corresponds to the pixel intensity for the horizontal purple line.

Figure 5.7: High-frequency content (HFC) measure computed for all the candidate and reference illuminations. We can observe how high-frequency illuminations (*Uffizi*, *Grace*, *Ennis*, *St. Peters*) also have a high HFC value, medium-frequency illuminations (*Pisa*, *Doge*) have a lower HFC value, and, last, low-frequency illuminations (*Glacier*) have the lowest HFC value.

5.2.3 Procedure

Subjects are shown a reference sample and a group of candidate material samples. Each experiment, *HIT* in MTurk terminology, consists of 23 unique reference material samples or *trials*, three of which are sentinels used to detect malicious or lazy users. Users are asked to "select five material samples which you believe are closer to the one shown in the reference image". Additionally, we instruct them to make their selection in decreasing order of confidence. We let the users pick five candidate materials because just one answer would provide sparse results. We launched 25 *HITs* for each experiment and each *HIT* was answered by six different users. This resulted in a total of 27.000 non-sentinel trials, 12.000 belonging to the four experiments analyzed in Section 5.3, and 15.000 of them belonging to the five additional experiments discussed in Section 5.4 (a total of nine different experiments with 25 *HITs* each, each *HIT* answered by six users and 20 non-sentinel trials per *HIT*). Users were not allowed to repeat the same *HIT*.

The set of materials in the candidate samples does not vary across *HITs*, however, the position of each sample is randomized for each trial. This has a two-fold purpose: it prevents the user from memorizing the position of the samples, and it prevents them from selecting only the candidate samples that appear at the top of their screen. The reference samples do not repeat materials during a *HIT* and the reference material is always present among the candidate samples. During the experiment, stimuli keep a constant display size of 300 x 300 px. for the reference, and of 120 x 120 px. for the candidate stimuli (except for some of the additional experiments explained in Section 5.4 where both reference and candidate stimuli are displayed at either 300 x 300 px. or 120 x 120 px.). Figure 5.3 shows a screenshot with the graphical user interface during the behavioral experiments. On the left-hand side, we can observe the selection panel with the current trial and the selection of the current materials. The right-hand side displays the set of candidate materials whereof users can pick their selection. Users were not

able to go back and re-do an already answered trial but they could edit their current selection of five materials until they were satisfied with their choice. Additionally, once the 23 trials of the *HIT* are answered, to have an intuition about the main features that humans use for material recognition, we asked the user: “Which visual cues did you consider to perform the test?”.

To minimize worker unreliability, the user performs a brief training before the real test [330]. To avoid giving the user further information about the test, we use a different geometry (*Havran-3* [118]) during the training phase. In this phase, the items of the interface are explained and the user is given guidance on how to perform the test using just a few images [94, 258, 169].

SENTINELS Each sentinel shows a randomly selected image from the pool of candidates as the reference sample. We consider user answers to the sentinel as valid if they pick the right material within their five selections, regardless of the order. We rejected users that did not correctly answer at least one out of the three sentinel questions. In order to ensure that users answers were well thought and that they were paying attention to the experiment, we also rejected users that took less than five seconds per trial (on average). In the end, we adopt a conservative approach and rejected 19.8% of the participants, gathering 21.660 answers (9.560 belonging to the behavioral experiments explained in Section 5.3 and 12.100 belonging to the additional experiments explained in Section 5.4).

5.3 RESULTS

We investigate which factors have a significant influence on user performance and on the time they took to complete each trial in the four experiments: *TEST HH*, *TEST HS*, *TEST SH* and *TEST SS*. The factors we include are: the reference geometry *Gref*, the candidate geometry *Gcand*, and the illumination of the reference sample *Iref*, as well as their first-order interactions (recall that the illumination of the candidate samples remains constant in these behavioral experiments). We also include the *Order* of appearance of each trial. We use a general linear mixed model with a binomial distribution for the performance since it is well-suited for binary dependent variables like ours, and a negative binomial distribution for the time, which provides more accurate models than the Poisson distribution by allowing the mean and variance to be different. Since we cannot assume that our observations are independent, we model the potential effect of each particular subject viewing the stimuli as a random effect. Since we have categorical variables among our predictors, we re-code them to dummy variables for the regression. In all our tests, we fix a significance value (*p*-value) of 0.05. Finally, for factors that present a significant influence, we further perform pairwise comparisons for all their levels (least significant difference pairwise multiple comparison test).

5.3.1 Analysis of User Performance and Time

In our online behavioral experiments, we rely on the top-5 accuracy to measure user performance. This metric considers an answer as correct if the reference is among the five candidate materials that the user picked in the trial. Since participants picked five materials ranked in descending order of confidence, the top-1 accuracy could also be considered for our analysis. However, the task they have to solve is not easy and users have an overall

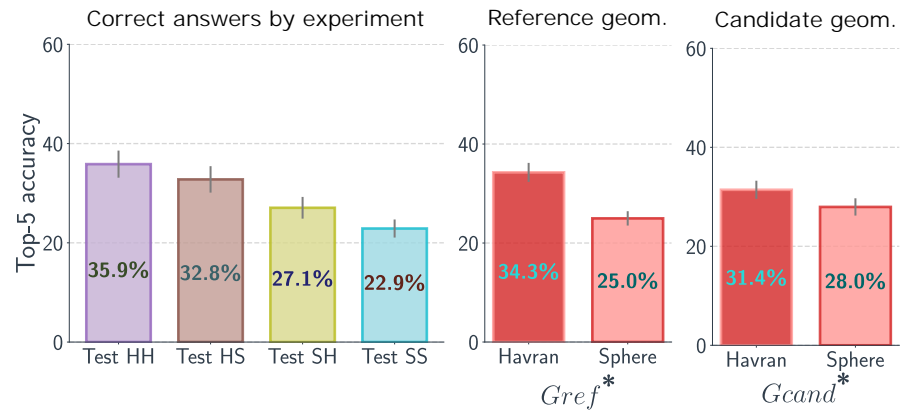


Figure 5.8: *Left*: Top-5 accuracy for each of the four behavioral experiment. *Center*: Top-5 accuracy for each reference geometry G_{ref} . *Right*: Top-5 accuracy for the candidate geometry G_{cand} . We can see how users seem to perform better when the candidate and reference are a high-frequency geometry. All plots have a 95% confidence interval. The names marked with * are found to have statistically significant differences.

top-1 accuracy of 9.21% which yields sparse results. A random selection would yield a top-1 accuracy of 1% and a top-5 accuracy of 5%.

INFLUENCE OF THE GEOMETRY There is a clear effect in user performance when the the geometry changes, regardless if that change happens in the candidate (G_{cand} , $p = 0.005$) or the reference geometry (G_{ref} , $p < 0.001$). This is expected, since the geometry plays a key role in how a surface reflects the incoming light and, therefore, will have an impact on the final appearance of the material. Figure 5.8 shows user performance in terms of top-5 accuracy with a 95% confidence interval when the reference and candidate geometry change jointly (left) or individually (center and right). Users seem to perform better when they have to recognize the material in a high-frequency geometry compared to a low-frequency one. Those results also suggest that changes in the frequencies of the reference geometry may have a bigger impact on user performance than changes in the frequencies of the candidate geometry (i.e. users perform better with a high-frequency reference geometry and low-frequency candidate geometry, compared to a low-frequency reference geometry and a high-frequency candidate geometry).

INFLUENCE OF THE REFERENCE ILLUMINATION We observe that the illumination of the reference image has a significant effect on user performance (I_{ref} , $p < 0.001$). This is expected since all the materials in a scene are reflecting the light that reaches them, therefore changes in illumination can significantly influence the final appearance of a material, and how we perceive it [32]. Figure 5.9, left, shows the top-5 accuracy for each reference illumination and groups of illuminations with statistically indistinguishable performance. We can observe how users seem to have better performance when the surface they are evaluating has been lit with a high-frequency illumination (*Ennis*, *Grace*, and *Uffizi*); while users appear to perform worse in scenes with an low-frequency illumination (*Glacier*); users show an intermediate performance with a medium-frequency illumination (*Doge* and *Pisa*). Moreover, we performed a least significant difference pairwise multiple comparison test to obtain groups of illuminations with statistically indistin-

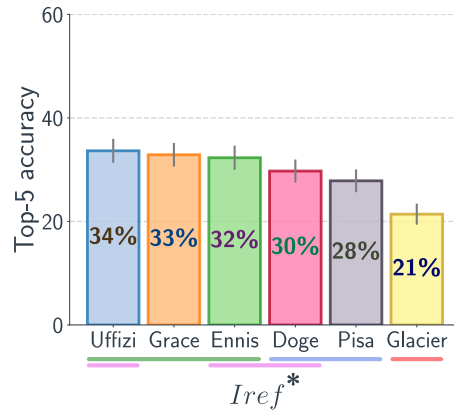


Figure 5.9: Top-5 accuracy for each reference illumination (I_{ref}). The horizontal lines under the x-axis represent groups of statistically indistinguishable performance. The reference illuminations marked with * denote significant differences. The error bars correspond to a 95% confidence interval.

guishable performance. These groups can be observed in Figure 5.9, under the x-axis. If we focus on I_{ref} we can see how high- (green), medium- (blue), and low-frequency (red) illuminations yield groups of similar performance. There is an additional group of statistically indistinguishable performance represented in pink color.

INFLUENCE OF TRIAL ORDER The order of appearance of the trials during the experiment does not have a significant influence in users performance ($Order, p = 0.391$).

FIRST ORDER INTERACTIONS We find that the interaction between the candidate geometry and the reference illumination has a significant effect on user performance ($G_{cand} * I_{ref}, p < 0.001$). Users seem to perform better with a high-frequency geometry (compared to a low-frequency one) when the reference stimuli features a high-frequency illumination ($I_{ref}=Uffizi, p = 0.019$; $I_{ref}=[Grace, Ennis], p < 0.001$). On the other hand, there appears to be no significant changes in performance between a high- and low-frequency candidate geometry when the reference stimuli has a medium- or low-frequency illumination ($I_{ref}=Doge, p = 0.453$; $I_{ref}=Pisa, p = 0.381$; $I_{ref}=Glacier, p = 0.770$). We argue that user performance is driven by the reference sample. When the reference material is lit with a low-frequency illumination, users seem to not be able to properly recognize it. Therefore, changes in the candidate geometry are not relevant to user performance. These results can be seen in Figure 5.9, center. Furthermore, under the x-axis, we can observe the groups with statistically indistinguishable performance where high-, medium-, and low-frequency illuminations yield groups of similar performance.

We also found out that the interaction between the reference geometry and the reference illumination has a significant impact in user performance ($G_{ref} * I_{ref}, p = 0.012$). Users seem to show better performance for all illuminations with a high-frequency reference geometry ($G_{ref}=Havran, I_{ref}=Uffizi, p = 0.002$; $I_{ref}=[Ennis, Pisa, Doge, Glacier], p < 0.001$), except for Grace illumination ($p = 0.176$), where the differences in humans performance are statistically indistinguishable. These results, together with the groups of statistically indistinguishable performance, can be seen in Figure 5.9, right.

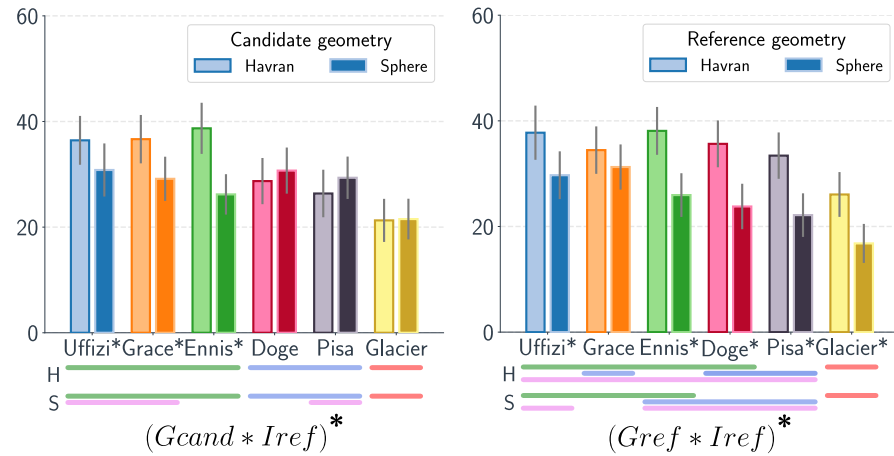


Figure 5.10: *Left*: Top-5 accuracy for each reference illumination when the candidate geometry (G_{cand}) changes. We can observe how users appear to perform significantly better with a high-frequency geometry (*Havran*) and illumination. On the other hand, for low-frequency illuminations, changes in the candidate geometry yield statistically indistinguishable performance. *Right*: Top-5 accuracy for each reference illumination when the reference geometry (G_{ref}) changes. We can observe how users seem to perform significantly better for all high-frequency illuminations, except for *Grace*. The horizontal lines under the x-axis represent groups of statistically indistinguishable performance. We can observe how the groups usually cluster high-, medium- and low-frequency illuminations. The reference illuminations marked with * denote significant differences in user performance between geometries for that illumination. The error bars correspond to a 95% confidence interval.

In general, we can not conclude that there are significant changes in performance due to the interaction between the candidate, and reference geometry ($G_{ref} * G_{cand}$, $p = 0.407$). Nevertheless, with a low-frequency reference geometry ($G_{ref} = sphere$), users seem to perform significantly better with a high-frequency candidate geometry ($G_{cand} = Havran$, $p = 0.009$).

5.3.1.1 Analysis of the Time Spent on Each Trial

To account for time, we measure the number of milliseconds that passed since the trial loaded in their screen and until they picked all five materials and pressed the "Continue" button. In Figure 5.11 we can see how the time spent to answer each trial becomes stable as the behavioral experiment advances.

INFLUENCE OF TRIAL ORDER We find that the *order* of the trials has a significant influence on the average time users spend to answer them ($p < 0.001$). Users spend more time in the first questions and that after few trials the average time they spend becomes stable at around 20 seconds per trial (recall that the *order* does not influence performance). This is expected as users have to familiarize with the experiment during the first iterations. As the test advances, they learn how to interact with it and the time they spend becomes stable.

INFLUENCE OF REFERENCE ILLUMINATION The reference illumination I_{ref} influences the time users spend to answer each trial ($p = 0.001$). Users

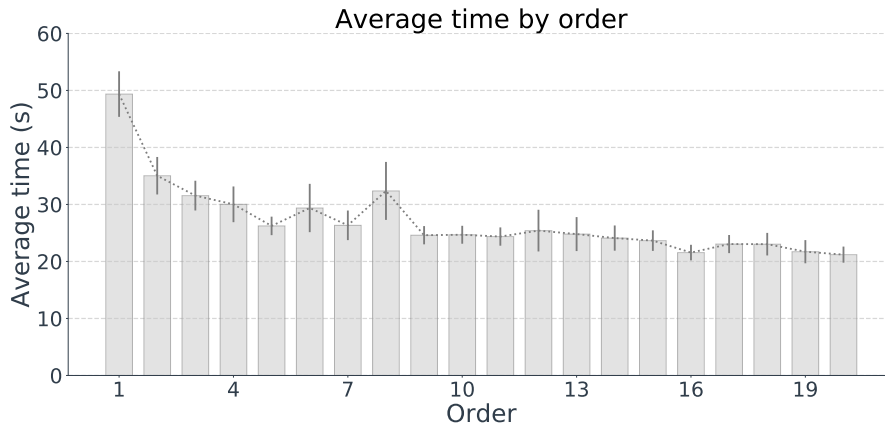


Figure 5.11: Average time the users spend for each trial according to the order of appearance during the online behavioral experiment. We can observe how, as the user progresses through the experiment, the time spent on each trial becomes stable. The error bars correspond to a 95% confidence interval.

spend more time when the stimuli are lit with *Ennis* illumination while they are the fastest when the illumination is *Doge*.

We did not find a significant influence of the reference geometry G_{ref} or candidate geometry G_{cand} in the average time each user spent to answer each trial.

FIRST ORDER INTERACTIONS We observe that users take significantly longer to answer the trials when the reference geometry and the candidate geometry change ($G_{ref} * G_{cand}$, $p = 0.001$). This happens in the case where the reference geometry has mostly low spatial frequency content and the candidate geometry changes ($G_{cand}=sphere$, $p=0.002$); and when the reference has mostly low spatial frequency ($G_{ref}=sphere$, $p=0.001$) and the candidate geometry changes.

5.3.2 High-level Factors Driving Material Recognition

In addition to the analysis, we also try to gain intuition on which high-level factors drive material recognition, investigate how simple image statistics and image histograms correlate with human answers, and analyze highly non-linear statistics in material classification tasks by training a deep neural network.

VISUALIZING USER ANSWERS To gain intuition on which high-level factors humans might use while recognizing materials, we employ a stochastic triplet embedding method called t-STE (t-Student stochastic triplet embedding) [316] directly on user answers. This method maps user answers from their original non-numerical domain into a two-dimensional space that can be easily visualized. Figure 5.12 shows the two-dimensional embeddings after applying the t-STE algorithm to the answers of each online behavioral experiment. Each point in the embedding represents one of the 100 materials from the Lagunas et al. dataset. The insets show the color of each material based on the color classification proposed by Lagunas et al. [171]. We can observe how materials are clustered by color and, if we focus in a single

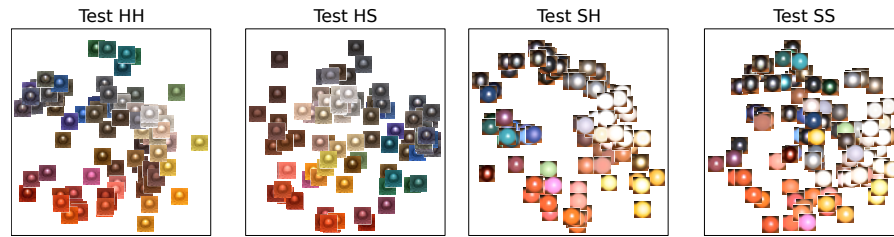


Figure 5.12: Visualizations of user answers to each of the four online behavioral experiments (namely TEST HH, TEST HS, TEST SH, and TEST SS) using the t-STE algorithm [316]. We can see how, for all experiments, materials with similar color properties are grouped together. Furthermore, if we explore the color clusters individually, we can see how there is a second-level arrangement by reflectance properties. These observations suggest that users may be performing a two-step process while recognizing materials where first, they sort them out by color, and second, by reflectance properties.

color, they seem to be clustered by reflectance properties (for instance, in TEST HH, red color cluster, we can observe how on the left there are specular materials while on the right there are diffuse materials). This suggests that users have followed a two-step strategy to recognize the materials, and that the high-level factors driving material recognition might be color first, and the reflectance properties second. At the end of the HIT, users were asked to write the main visual features they used to recognize materials. Out of 368 unique users from the experiments analyzed in Section 5.3, 273 answered that they have used the *colors*, and 221 answered that they relied on the *reflections*. Among them, 157 answered both *color* and *reflections* as some of the visual cues they have used to perform the task. This observation, together with the t-STE visualization, strengthens the hypothesis of a two-step strategy.

IMAGE STATISTICS Previous studies focused on simple image statistics as an attempt to further understand our visual system [218, 4]. Nevertheless, it is argued if our visual system actually derives any aspects of material perception using such simple statistics [10, 159, 235]. We tested out the correlation between the first four statistical moments of the luminance (considered as the ratio: $L = 0.3086 * R + 0.6094 * G + 0.0820 * B$), the pixel intensity for each color channel independently, and the joint RGB pixel intensity, directly against users top-5 accuracy. To measure correlation we employ a Pearson \mathcal{P} and Spearman \mathcal{S} correlation test. We found out that there is little to no correlation except for the standard deviation of the joint RGB pixel intensity where $\mathcal{P}^2 = 0.43$ ($p < 0.001$) and $\mathcal{S}^2 = 0.50$ ($p < 0.001$). Additional information can be found in Appendix B.1.

IMAGE HISTOGRAMS We also compute the histograms of the RGB pixel intensity (rgb), of the luminance (lum), of a Gaussian pyramid [179] (gaussian), of a Laplacian pyramid [38] (laplacian), and of log-Gabor filters designed to simulate the receptive field of the simple cells of the Primary Visual (V1) Cortex [80] (gabor). To see how such histograms would perform classifying materials, we train a support vector machine (SVM) that takes the image histogram as the input and classifies the material in that image. We use a radial basis function kernel (or Gaussian kernel) in the SVM. We use all image histograms that do not feature *Havran* geometry as the training set and leave the ones with *Havran* as test set. In the end, the best performing

SVM uses the RGB image histogram as the input and achieves a 24.17% top-5 accuracy in the test set. In addition, we compare the predictions of each SVM directly against human answers. For each reference stimuli we compare the five selections of the user against the five most-likely SVM material predictions for that stimuli. The best SVM uses the histograms of V_1 -like subband filters and agrees with humans 6.36% of the time. Moreover, we compare histogram similarities against human answers using a chi-square histogram distance [245]. For a reference image stimuli we measure its similarity against all possible candidate image stimuli and compare the closest five against participants' answers. The Gaussian pyramid histogram obtained the best result, agreeing with humans 6.29% of the time. These results show how simple statistics, and higher-order image histograms seem not to be capable of fully capturing human behavior.

Each of the trained SVM achieved a top-5 accuracy in the test set of: 24.17% (rgb), 15.16% (lum), 22.50% (gaussian), 6.33%, 7.52% (laplacian), and 16.33% (gabor), respectively. In addition, we have compared how the SVM predictions agreed with humans' answers from the online behavioral experiments. For each SVM the agreement is: 4.24% (rgb), 4.33% (lum), 4.34% (gaussian), 5.04% (laplacian), and 6.36% (gabor) respectively. Last, we have also computed the histogram similarity using a chi-squared distance. Then, we have taken the five closest samples and compared that with human answers. We do that for each of the five different histograms and each achieves the agreement: 5.95% (rgb), 5.45% (lum), 6.29% (gaussian), 4.97% (laplacian), and 5.07% (gabor) respectively.

IMAGE FREQUENCIES To understand if humans performance could be explained by taking into account the spatial frequency of the reference stimuli, at their viewed size, we have added the high-frequency content measure (HFC), and the first four statistical moments of the reference stimuli magnitude spectrum to the factors analyzed in Section 5.3. We found that the Skewness ($p < 0.001$) and Kurtosis ($p < 0.001$) of the magnitude spectrum seem to have a significant influence on humans performance; however, they present a very small effect size.

HIGHLY NON-LINEAR MODELS Recent studies suggest that, to understand what surrounds us, our visual system is doing an efficient non-linear encoding of the proximal stimulus (the image input to our visual system) and that highly non-linear models might be able to better capture human perception [89, 57]. Inspired by this hypothesis, we have trained a deep neural network called ResNet [123] employing a loss function suitable to classify the materials in the Lagunas et al. dataset. The images feature the same illuminations as the reference stimuli. We left out the images rendered with *Havran* geometries for validation and testing purposes, and use the rest during training. To know which material the network classifies we add a softmax layer at the end of the network. The softmax layer outputs the probability of the input image to belong to each material in the dataset. In comparison, the model used by Lagunas et al. [171] does not have the last fully-connected and softmax layer, and it is trained using a triplet loss function aiming for similarity instead of classification.

To train the 35 layers ResNet (34 of the original model plus an additional fully connected) [123] we have employed the dataset introduced by Lagunas et al. [171], which contains renderings of materials with different illuminations and geometries. We use a soft cross-entropy loss where samples that

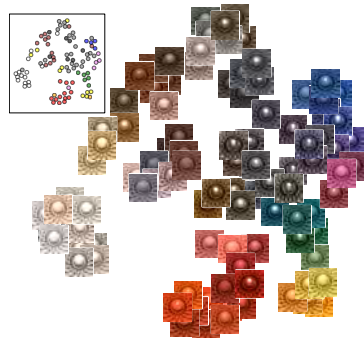


Figure 5.13: Two-dimensional visualization obtained using the UMAP algorithm [214]. The inset shows the color of each material. We can observe how materials are arranged by color clusters. Moreover, we can observe similarities between this visualization and the t-STE visualization on user answers.

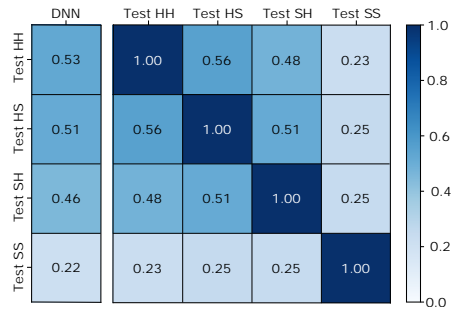


Figure 5.14: Normalized pairwise similarity for each online behavioral experiment and the deep neural network trained for material classification. We can observe how the pairwise similarity decreases as the stimuli in the experiments cover fewer frequencies in the spectrum, where *Test SS* has the lowest similarity.

do not belong to the same class are penalized [302]. The loss function takes the probabilities output of the softmax layer and penalizes when they give a high probability to the materials that do not belong to the input image. The images input to the model are resized to 224x224 px.

The parameters of the model are initialized using a pretrained version on ImageNet dataset [61]. We use the ADAM algorithm [161] as the optimizer. The model has been trained during 50 iterations starting at a learning rate of 10^{-3} and decayed by a factor of 10 at the iteration 20, 35, and 45; the batch-size was set to 64 images. We use the PyTorch framework and use an Nvidia 2080Ti GPU.

At the end of the training, the model achieves a top-5 accuracy of 89.63% on the test set suggesting that such models are actually capable of extracting meaningful features from labeled proximal image data. To gain intuition on how the network has learned we have used the UMAP (Uniform Manifold Approximation and Projection) algorithm [214]. This algorithm aims to reduce the dimensionality of a set of feature vectors while maintaining the global and local structure of their original manifold. Figure 5.13 shows a two-dimensional visualization of the test set obtained using the 128 features of the fully-connected layer before softmax. We can observe how materials seem to be grouped first by color and then by its reflectance properties suggesting that the model may have used similar high-level factors to humans when classifying materials.

We additionally assess the degree of similarity between the high-level visualization of each online behavioral experiment and the high-level visualization of the deep neural network. We calculate the similarity in a pairwise fashion where we choose a material sample and retrieve its five nearest neighbors in two different low-dimensional representations. Then, we compute the percentage of materials that are the same in both groups of nearest neighbors. We repeat this process for all the materials and calculate the similarity as the average.

The low-dimensional representations are obtained with stochastic methods, where the same input can have different results if we vary the parameters. To evaluate the degree of self-similarity, we run the t-STE algorithm [316] on each behavioral experiment using five different sets of fully randomly sampled parameters. We obtain a self-similarity value of 0.66, on average across experiments. For comparison, a set of random low-dimensional representations have a similarity of 0.06, on average. Figure 5.14 shows the average pairwise similarity normalized by the value of self-similarity and random similarity for all experiments and the deep neural network visualization. If we compare between behavioral experiments, we can observe a decreasing degree of similarity as their stimuli feature fewer frequencies in the spectrum, where TEST SS yields the lowest similarity in each of the pairwise comparisons. We argue that TEST SS has the lowest similarity because it is the experiment where users have the worst performance thus yielding a *blurry high level visualization*. On the other hand, the network is very accurate classifying materials and yields a high-level visualization with well-defined material clusters. Moreover, if we focus on the deep neural network visualization, we can observe how its similarity values are, in general, on par with those obtained by users in TEST HH, TEST HS, and TEST SH. This result further supports the hypothesis that both humans and deep neural networks may rely on similar high-level visual features for material recognition tasks. However, this is just a preliminary result that may highlight a future avenue of research, and a thorough analysis of the perceptual relationship between deep learning architectures and humans is out of the scope of this chapter.

5.4 DISCUSSION

From our online behavioral experiments, we have observed that humans appear to perform better at recognizing materials in stimuli with high-frequency illumination and geometry. Moreover, our performance when recognizing materials is poor on low-frequency illuminations and it remains statistically indistinguishable irrespective of the spatial frequency content in the candidate geometry.

ASYMMETRIC EFFECT OF THE REFERENCE AND CANDIDATE GEOMETRY

It is also interesting to observe that humans seem to have better performance with a high-frequency reference geometry, compared to a high-frequency candidate geometry ($p = 0.001$, see Figure 5.15, left). The number of candidates with respect to the reference could be used as an explanation for this observation, since users may devote more time to inspecting the single reference than the higher number of candidates. At the same time, a lower performance with a high-frequency reference geometry may speak against an inverse optics approach since having multiple candidate materials with the same geometry and illumination could provide a strong cue to inferring the material.

One potential factor that may explain this difference in performance is the different display sizes of the reference (300x300 px.) and the candidate (120x120 px.) stimuli. To test this hypothesis, we have launched two additional experiments where we collect answers on TEST HS and TEST SH displaying the candidate and the reference stimuli at size 300x300, and other two additional experiments where they are displayed at 120x120 px. We sample the stimuli to cover all the possible combinations of illuminations and materials and keep other technical details as explained in Section 5.2.

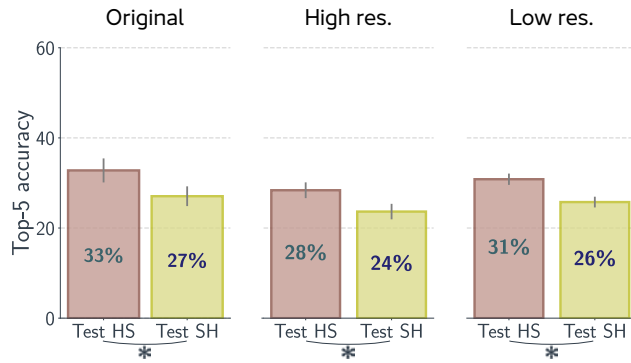


Figure 5.15: Top-5 accuracy obtained by participants in the original experiment (left), when the stimuli are displayed at 300x300 px. (middle), and at 120x120 px. (right). We can observe how the asymmetric effect of participants performing better when *Havran* is the reference geometry (TESTHS) compared to when it is the candidate (TESTSH) remains present when the participants observe the reference and candidate stimuli at identical sizes (middle and right). The * denotes significant differences. The error bars correspond to a 95% confidence interval.

We perform an analysis of the gathered data similar to the one explained in Section 5.3, but using the different experiment type as a factor. From our results we observe that such asymmetric effect remains present when the stimuli are displayed at 300x300 px. ($p < 0.001$) and when they are displayed at 120x120 px. ($p < 0.001$). Those results can be seen in Figure 5.15, middle and right. It is also interesting to observe how users have slightly worse performance when the stimuli are displayed at 300x300 px. At such display size only three candidate stimuli per row could be displayed taking into account the most used display size. Thus, seems reasonable to think that the need for additional scrolling could be hampering participants' performance.

INFLUENCE OF THE CANDIDATE ILLUMINATION We have seen that humans seem to be better at recognizing materials under high-frequency reference illuminations. However, in Figure 5.5 and 5.7 we can see that the *St. Peters* candidate illumination features a similar frequency content to the reference illuminations where users have better performance. To assess if *St. Peters* illumination contains a set of frequencies that aids recognizing materials under reference illuminations with a similar set of frequencies, we have launched an additional behavioral experiment. In this experiment we use *Doge*, a medium-frequency illumination, as the candidate illumination. We sample the stimuli to cover all materials and reference illuminations in TEST HH. Other technical details are kept as explained in Section 5.2. From the data collected (see Figure 5.16), we can observe how, using *Doge* as the candidate illumination, humans performance follows a similar distribution to the original experiment (with *St. Peters* as the candidate illumination). Participants seem to perform better with high-frequency reference illuminations (*Uffizi*, *Grace*, *Ennis*), they perform worse with medium-frequency ones (*Pisa*), and have their worst performance with low-frequency reference illuminations (*Glacier*). In addition, participants seem to have slightly better performance with a high-frequency candidate illumination (*St. Peters*) compared to a medium-frequency one (*Doge*).

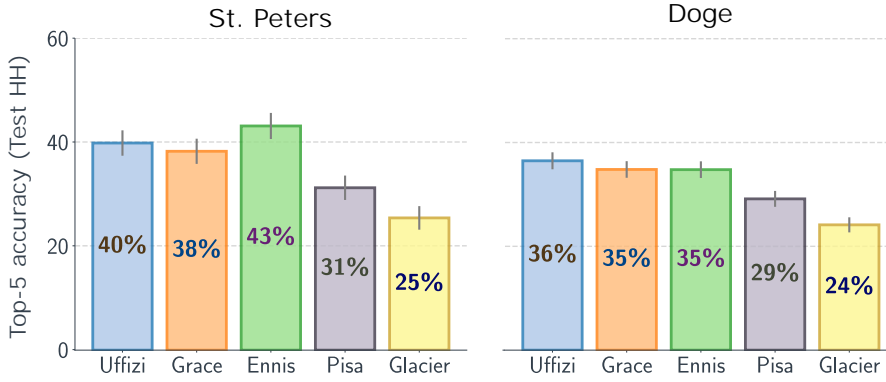


Figure 5.16: Left: Top-5 accuracy for each reference illumination when *St. Peters*, a high-frequency illumination, is the candidate illumination. Right: Top-5 accuracy for each reference illumination when *Doge*, a medium-frequency illumination, is the candidate illumination. Both results have been obtained for TEST HH. We can observe how, for both candidate illuminations, participants seem to perform better with high-frequency reference illuminations (*Uffizi*, *Grace*, *Ennis*), they perform worse with medium-frequency ones (*Pisa*), and have their worst performance with low-frequency reference illuminations (*Glacier*). In addition, we also observe that participants have slightly better performance when *St. Peters* (high-frequency illumination) is the candidate illumination. The error bars correspond to a 95% confidence interval.

INTERPLAY BETWEEN MATERIAL, GEOMETRY, AND ILLUMINATION We have looked into how geometry, illumination, and their frequencies affect our performance in material recognition tasks. Our stimuli were rendered images in which we varied the frequency of the illumination, and of the underlying geometry of the object present. To better understand how our factors (illumination and geometry) affect the generated stimuli, and thus the proximal stimulus, we offer here a brief description of the rendering equation, providing an explanation of the probable effect of how the frequencies of the geometry and illumination in the 3D scene affect the final, rendered image that is used as a stimulus in our experiments. From the point of view of the rendering equation, the radiance L_o at point x in direction ω_o , assuming a distant illumination and non-emissive surfaces can be approximated as

$$L_o(x, \omega_o) \approx \int_{\Omega} L_i(\omega_i) F(\omega_i, \omega_o) T(x, \omega_i, \omega_o) d\omega_i, \quad (5.1)$$

where L_i accounts for the incoming light, the variable F accounts for the reflectance of the surface, and T depends on the point of the surface we are evaluating, therefore, on the geometry.

The simulation of the radiance L_o can be seen as a convolution (spherical rotation) [254] between each signal: incoming radiance L_i , material F and geometry T . Moreover, if we analyze L_o in the frequency domain (where \mathcal{F} is the Fourier transform), and apply the convolution theorem ($f * g = \mathcal{F}(f) \cdot \mathcal{F}(g)$) the value of $\mathcal{F}(L_o)$ becomes

$$\mathcal{F}(L_o) \approx \mathcal{F}(L_i) \mathcal{F}(F) \mathcal{F}(T). \quad (5.2)$$

Equation 5.2 shows how the frequency of the radiance L_o in the final image is a multiplication of all the other signals, L_i , F , and T in the frequency domain.

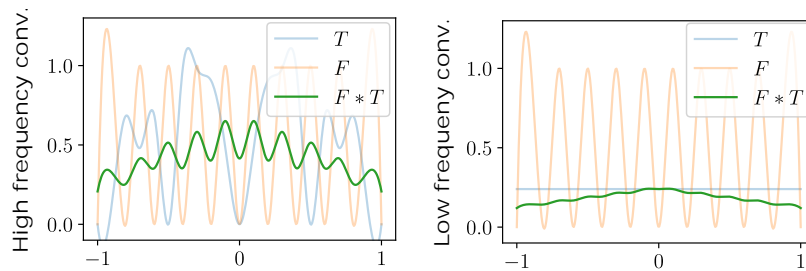


Figure 5.17: Example of a convolution ($F * T$, green line) between a material (F , orange line) and a geometry (T , blue line) with different frequency content. *Left*: We can see how when we convolve a geometry and a material with high spatial frequencies, the resulting convolution also retains high-frequency content. *Right*: We observe how when geometry has low spatial frequencies and the material has high spatial frequencies, the resulting convolution does not retain high-frequency content. Note that T and F are not necessarily related to a real BRDF or shape from the ones reported in this chapter.

Thus, the final image will only have the frequencies that are contained within the three other signals. Figure 5.17 shows how when we convolve two high-frequency signals, the resulting one keeps the high-frequency content; on the other hand, when we convolve a high- and a low-frequency signal, the resulting one has most of its frequencies masked.

We can relate the observations made from Equation 5.2 to the results on user performance that we obtained from the online behavioral experiments. We have seen that users seem to consistently perform better when they recognize materials from high-frequency geometries and illuminations. This finding is supported by Equation 5.2 since, to avoid filtering the frequencies of the material in the stimuli, it should have a high-frequency geometry and illumination. Moreover, a low-frequency geometry (or illumination) could filter out the frequencies of the illumination (or geometry) and the material, thus yielding fewer visual features on the final image and, as a result, worse users performance. This is consistent with our findings from the analysis of first-order interactions for users performance in Section 5.3.1.

MATERIAL CATEGORIES We have seen that the reflectance properties seem to be one of the main high-level factors driving material recognition. In this regard, we have also investigated users' performance using the classification by reflectance type proposed by Lagunas et al. [171], where the MERL database is divided into eight different categories with similar reflectance properties. On average, users perform best on *acrylics*, with a top-5 accuracy of 45.45%, while they have their worst performance with *organics*, with an accuracy of 10.22%. Figure 5.18 shows the top-5 accuracy for each category in each reference illumination. Firstly, we observe that users seem to perform better with high-frequency illuminations (*Uffizi*, *Grace*, *Ennis*). However, we can see how *fabrics* and *organics* do not follow this trend. We argue that *fabrics* and *organics* contain mostly materials with a diffuse surface reflectance (low-frequency) that clamp the frequencies of the illumination and, therefore, yield fewer cues in the final stimulus that is input to our visual system.

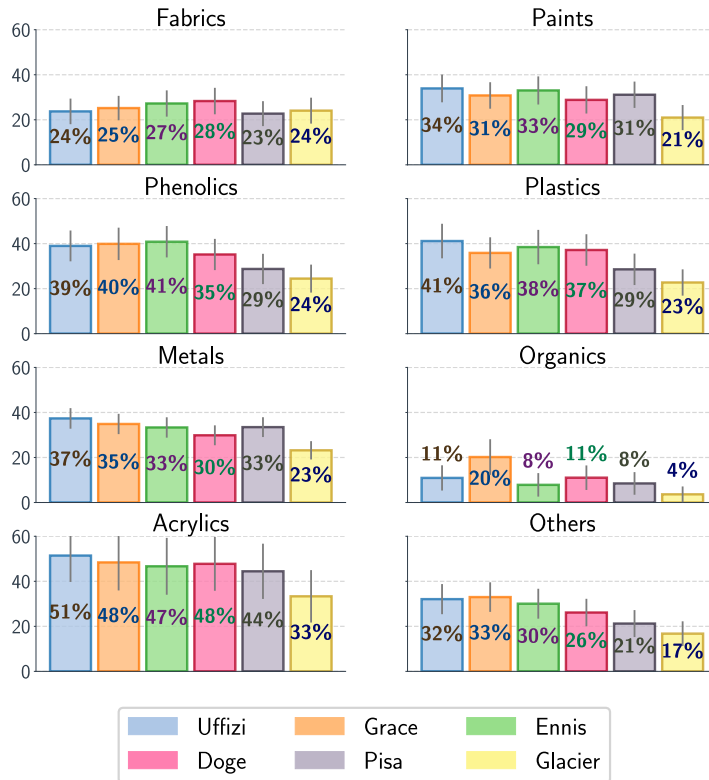


Figure 5.18: Users’ performance, in terms of top-5 accuracy, for material recognition tasks taking into account the reflectance of the materials. We can observe how, on average, users perform better for high-frequency illuminations (*Uffizi*, *Grace*, and *Ennis*). Also, we can see how for classes, like *fabrics* or *organics*, containing materials with diffuse surface reflectance (low-frequency), users do not have better performance with broad frequency content illuminations. We argue that, since they have a low-frequency surface reflectance, even though there is a high-frequency illumination, those frequencies cannot be represented on the final stimulus that is input to our visual system.

5.5 CONCLUSIONS

In this chapter, we have presented a thorough and systematic analysis of the interplay between geometry, illumination, and their spatial frequencies in human performance recognizing materials. We launched rigorous crowd-sourced online behavioral experiments where participants had to solve a material recognition task between a reference and a set of candidate samples. From our experiments, we have observed that, in general, humans appear to be better at recognizing materials in a high-frequency illumination and geometry. We found that simple image statistics, image histograms, and histograms of V_1 -like subband filters are not capable of capturing human behavior, and, additionally, explored highly non-linear statistics by training a deep neural network on material classification tasks. We showed that deep neural networks can accurately perform material classification, which suggests that they are capable of encoding and extracting meaningful information from labeled proximal image data. In addition, we gained intuition on which are the main high-level factors that humans and those highly non-linear statistics use for material recognition and find preliminary evi-

dence that such statistics and humans may share similar high-level factors for material recognition tasks.

LIMITATIONS AND FUTURE WORK To collect data for the online behavioral experiment we have relied on the Lagunas et al. [171] dataset which contains images of a diverse set of materials, geometries, and illuminations that faithfully resemble their real-world counterparts. This database focuses on isotropic materials, which are capable of modeling only a subset of real-world materials. A systematic and comprehensive analysis of other heterogeneous materials, or an extension of this study to other non-photorealistic domains, remains to be done. Our stimuli were rendered using the sphere and *Havran* geometry, although those surfaces have been widely used in the literature [171, 276, 142, 118], introducing new geometries could help to further analyze the contribution of the spatial frequencies of the geometry in our perception of material appearance [227]. Moreover, to select our stimuli, we characterized the frequency content of real-world illuminations using the high-frequency content measure [37]. We focus on real-world illuminations, which are by definition broadband, therefore, we do not impose nor limit their frequency distribution in our analyses; carefully controlling the spatial frequency of the stimuli via filtering in order to isolate frequency bands and study their individual contribution to the process of material recognition is an interesting avenue of research.

In our additional experiments, we have investigated the asymmetric effect in performance with a high-frequency reference geometry, compared to a high-frequency candidate geometry when all stimuli are displayed at the same size. A rigorous study of the interplay between display size, the spatial frequencies of the stimuli, and how this affects humans' performance on material recognition remains an interesting line of future work. Furthermore, despite the fact that our neural network was trained to classify materials, without any sort of perceptual information, it achieved an agreement with participants answers of 22.43%. This does not prove that the neural network follows the same mechanisms as humans do when performing these tasks. However, this result together with the increase in popularity of deep neural networks, makes the analysis of the perceptual relationship between learned features and the features that our visual system uses to recognize materials a promising avenue to explore. Last, we hope that our analyses will provide relevant insights that will help shed light on the underlying perceptual processes that occur when we recognize materials and, in particular, on how the confounding factors of geometry and illumination affect our perception of material appearance.

In this chapter, we analyze the effect of motion in the perception of material appearance. First, we create a set of stimuli containing 72 realistic materials, rendered with varying degrees of linear motion blur. Then we launch a large-scale study on Mechanical Turk to rate a given set of perceptual attributes, such as brightness, roughness, or the perceived strength of reflections. Our statistical analysis shows that certain attributes undergo a significant change, varying appearance perception under motion. In addition, we further investigate the perception of brightness, for the particular cases of rubber and plastic materials. We create new stimuli, with ten different luminance levels and seven motion degrees. We launch a new user study to retrieve their perceived brightness. From the users' judgements, we build two-dimensional maps showing how perceived brightness varies as a function of the luminance and motion of the material.

This work was presented at the *Symposium on Applied Perception (SAP)* [207]. While I was not the leading author in this work, my contributions lie on the design and launch of the second experiment (Section 6.4.1) together with writing the manuscript.

R. Mao, M. Lagunas, B. Masia, & D. Gutierrez
The Effect of Motion on the Perception of Material Appearance
Symposium of Applied Perception, No. 16, 2019

6.1 INTRODUCTION

The perception of material appearance is affected by confounding effects such as the shape of the object, illumination, viewing conditions, etc. However, being able to distinguish between materials and to infer their key properties by sight is an inherent process in humans, which is invaluable for multiple tasks. How this is done exactly remains unclear, since human perception is a complex process influenced by a large set of variables. In fact, a unified theory that fully explains such a process does not exist [82, 9].

To reduce the dimensionality of the perception of material appearance, many works have focused on developing applications for material synthesis [355], editing [276], or filtering [142]. Other works focus on investigating individual variables that may affect our perception, such as translucency [105], or gloss [246, 331].

With a few exceptions, most of the research on material perception has focused on static stimuli. Doerschner et al [64] identified three motion cues the brain could rely on to distinguish between matte and shiny surfaces. Aliaga et al. [8] explore the interplay of cloth and dynamics, while Sakano et al. [262] studied how self-motion influences the perception of gloss. Nevertheless, a generalized approach that tries to explain the effect of motion on a wide variety of attributes remains an open avenue of research.

In this chapter, we study the influence of linear motion in the perception of material appearance relying on a set of intuitive, high-level perceptual attributes [276]. Inspired by previous large-scale studies [258, 276], we rely on crowdsourced data, from a set of 72 varied, realistic materials. To create a dense space of stimuli, we add additional variations by modifying their luminance, and render them with different levels of motion. This creates a total of 284 stimuli, used in two different experiments.

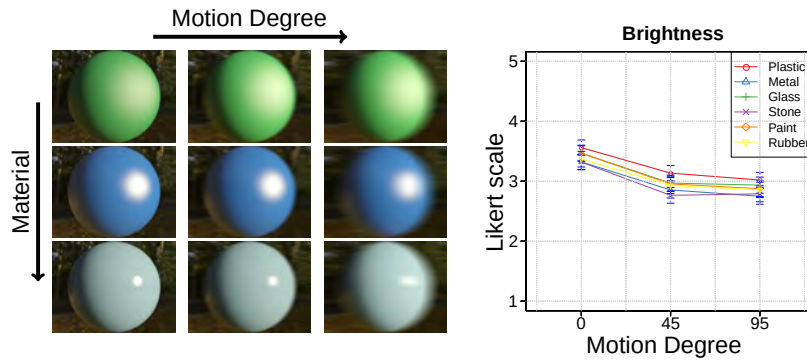


Figure 6.1: *Left*: Representative samples of the stimuli used in our first experiment, in which we analyze the perception of a series of high-level attributes for varying degrees of motion and different materials. *Right*: The variation of the high-level attribute *brightness* for different motion degrees and material categories tested in our first experiment.

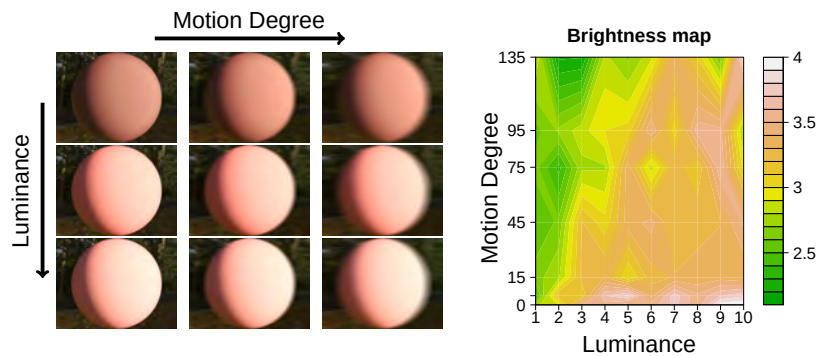


Figure 6.2: *Left*: Representative samples of the stimuli used in our second experiment, in which we look at the influence of varying luminance levels and motion degree on perceived brightness. *Right*: Results of our second experiment, showing a non-linear influence of motion degree and luminance level on perceived brightness.

In the first experiment (see Figure 6.1), we statistically analyze the significant change in how several of these high-level attributes are perceived according to the degree of motion. While brightness decreases significantly, more diffuse appearances remain constant as motion increases. This may be explained by the fact that motion blur has a higher impact on the high frequencies of an image which are usually associated with more specular materials.

In a second experiment (see Figure 6.2), we focus on the perception of brightness, where the first experiment showed a larger influence of motion. We select rubber and plastic materials, with seven motion degrees and ten luminance levels. Using human judgements through Mechanical Turk, we build a brightness map as a two-dimensional function that shows the average perceived brightness for each level of luminance and degree of motion.

Our contributions represent just a step towards a better understanding of how motion affects material perception, a relatively unexplored topic compared to studies on static stimuli. To encourage further research on this topic, all our data will be made publicly available, including the material database, code, and the users' responses.



Figure 6.3: Summary of the process followed to generate stimuli of homogeneous materials with different degrees of motion. From left to right, the first column shows the frameworks used for rendering. The second column is the rendering result using *vMaterials* and ray tracing engine OptiX [242], including texture and small geometric details. The third column represents the homogeneous material removing additional information provided in *vMaterials*. The last column shows the stimuli with different degrees of motion.

6.2 STIMULI CREATION

The following section explains the process followed in order to create the stimuli for our experiments.

MATERIALS We use the material library *vMaterials*, a large collection of realistic materials and lights described in Material Definition Language (MDL) [155]. We did not use a measured material database since they will not give us the efficiency needed to, in the future, use high-level material features to control and edit material appearance in real-time. Among all the available options, we selected a subset of 72 materials that span six categories, including glass, metal, paint, plastic, rubber, and stone. We decided to choose materials categories that represent daily items to avoid introducing bias due to unawareness in the participants' answers.

The materials in *vMaterials* include additional layers that add texture or small geometric details at rendering time. These features can alter the information that arrives at our visual system and distract the participant from its real purpose which is to observe the material itself. Therefore, in order to focus the attention of the user, we decided to remove all additional layers included in the material description, obtaining at the end, a homogeneous representation of the material. This procedure is depicted in Figure 6.3.

ADDING MOTION To approximate the effect of motion in our stimuli, we use an image algorithm consisting in a two-dimensional filter [34], described in Equation 6.1.

$$\mathcal{I}_o(x, y) = \sum k(x', y') \times \mathcal{I}_i(x + x', y + y') \quad (6.1)$$

where $\mathcal{I}_i(x, y)$ and $\mathcal{I}_o(x, y)$ are the input and output images respectively; and (x, y) are the coordinate of the pixel in the image. The expression $k(x', y')$ means a convolution kernel, whose index in the kernel is (x', y') . The range of $k(x', y')$ is $1 \leq x' \leq k_c$ and $1 \leq y' \leq k_r$, where k_c and k_r are the number of columns and rows in k . When the aperture is partially outside the image, we interpolate outlier pixel values by mirroring \mathcal{I}_i .

Our experiments only intend to investigate material perception for linear movement, therefore $k(x', y')$ is a horizontal vector, whose value is $1/k_c$. The size of the kernel (k_c), can be considered as the motion degree. In a static situation, the motion degree is 0.

SCENE We use a sphere as the 3D model, a well-known surface, widely used in previous user-studies [142, 86]. The 3D model is placed in the

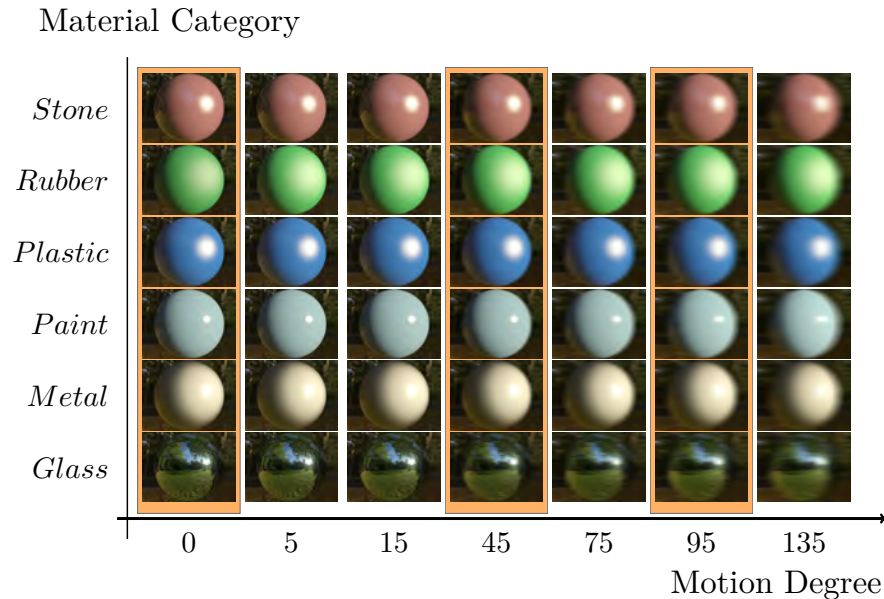


Figure 6.4: Subset of the stimuli used in the user experiment. Each row correspond to a material from a material category. The x -axis shows the degree of motion, going from 0 (a static stimuli) to 135 (fastest stimuli). For the first experiment, we only use the motion degrees 0, 45, and 95 since they provide significant visual changes in the final renderings.

centre of the scene. In order to render our stimuli, first, we use OptiX [242], a general purpose ray tracing engine, with the homogeneous material definition to generate an image of a static material, then we apply the two-dimensional filter explained in Equation 1 to add motion to the scene. Figure 2 summarizes the process of stimuli creation.

6.3 FIRST EXPERIMENT: RATING MATERIAL ATTRIBUTES

This section describes the approach followed in order to collect human ratings of perceptual attributes on materials under the influence of a certain degree motion.

STIMULI We choose a set of 72 different realistic materials, that span six different material categories, and three degrees of motion: 0, 45, and 95 with a significant difference between them in order to provide a notable change in the stimuli appearance. Further explanation about the stimuli creation is given in Section 6.2. A subset of the stimuli is shown in Figure 6.4, all the stimuli are shown in the supplementary material.

PARTICIPANTS Similar to previous work dealing with large-scale experiments in computer graphics [258, 276, 171], we rely on Amazon Mechanical Turk (MTurk) to launch our experiments. In the end, a total of 540 users performed the task. Users were unaware of the purpose of the experiment.

PROCEDURE Our user-study deals with the perception of material appearance under the influence of motion. We ask the users to rate a set of perceptual attributes regarding the material in the stimuli. They rate the attributes in a Likert-like scale, which has proven to work reliably in

multi-modal problems [342]. Following previous work [69, 345, 276], our scale ranges from 0 (very little) to 5 (a lot). We consider this range a good compromise between the complexity to fill in the survey and the number of available options for the participants.

Following the work of Serrano et al. [276], participants rate 14 perceptual attributes, namely: *plastic-like*, *rubber-like*, *metallic-like*, *fabric-like*, *ceramic-like*, *soft*, *hard*, *matte*, *glossiness*, *brightness*, *roughness*, *tint of reflections*, *strength of reflections*, and *sharpness of reflections*. Covering both, high and mid-level features of the material, allowing for a comprehensive evaluation of the participants' perception.

Each test consisted of 12 different trials, 4 groups of three spheres with the same material and different motion degrees, randomly displayed. The participants had to rate all the 14 perceptual attributes in all the trials of the user study (HIT, Human Intelligence Tasks, in MTurk terminology). The experiment was developed to work in standard web browsers. Before the real experiment, there is a thorough description of the task, and a brief training session in order to minimize worker unreliability [330]. During the real test, participants were shown one stimulus at a time. In the end, each stimulus was rated by 30 different users and we collected a total of 6480 answers for 14 attributes, yielding 90.720 ratings.

6.3.1 Analysis

Among the 14 perceptual attributes rated by the participants, we select a subset of six to perform the statistical analysis, those include *brightness*, *glossiness*, *matte*, *roughness*, *sharpness of reflections* and *strength of reflections*. We conducted a Friedman rank sums test, a non-parametric version of ANOVA [251], in order to analyze how the factors: *material category* and *motion degree* have affected participants' answers. The *motion degree* contains three levels: 0, 45, and 95 while the *material category* has six: glass, metal, paint, plastic, rubber, and stone. We decided to use a Friedman test since it is suitable for Likert-like ratings and samples do not need to be normally distributed. We chose a significance level of $p = 0.05$ in all our tests.

THE INFLUENCE OF MOTION A summary with all the p -values for each rated attribute and material category can be found in Table 6.1, a complete table with all the p -values can be found in the supplementary material. We observe that *glossiness*, *brightness*, *strength of reflections*, and *sharpness of reflections* have a significant change in all the material categories ($p < 0.05$). On the other hand, the attributes *matte*, and *roughness* are not influenced by motion and do not have a significant change for all the material categories ($p > 0.05$). These results may suggest that motion has a bigger effect on our perception of the attributes that describe the specularity of the material instead of on the ones that are used to characterize diffuse appearances.

THE INFLUENCE OF THE MATERIAL The attributes *matte*, and *roughness* are not influenced by motion in all the material categories. Only metal and paint materials for the *matte* attribute will have a significant change ($p < 0.05$). We argue that this significant change is produced because these materials have a characteristic narrow specular lobe that under the effect of motion blur will broaden, giving them a more diffuse and dimmer final appearance.

Attribute	Material	χ^2	p -value	Attribute	Material	χ^2	p -value
Matte	Plastic	0.301	0.860	Glossiness	Plastic	24.958	0.004
	Metal	13.942	0.001		Metal	19.080	0.072
	Glass	0.833	0.659		Glass	52.166	0
	Stone	1.240	0.538		Stone	64.045	0
	Paint	10.033	0.007		Paint	29.110	0
	Rubber	2.898	0.235		Rubber	19.648	0.054
Brightness	Plastic	70.490	0	Roughness	Plastic	2.924	0.232
	Metal	52.211	0		Metal	1.769	0.413
	Glass	62.027	0		Glass	2.233	0.327
	Stone	72.240	0		Stone	1.347	0.510
	Paint	71.274	0		Paint	1.327	0.515
	Rubber	59.049	0		Rubber	0.003	0.999
Strength of reflections	Plastic	41.848	0	Sharpness of reflections	Plastic	45.868	0
	Metal	42.993	0		Metal	51.283	0
	Glass	66.824	0		Glass	101.800	0
	Stone	29.444	0		Stone	48.995	0
	Paint	18.885	0.079		Paint	38.979	0
	Rubber	14.236	0.001		Rubber	34.243	0

Table 6.1: Results from Friedman rank test for all the materials and attributes over all motion degrees. The last two columns show the results from the Friedman test (χ^2 and p -values).

Given the previous findings, we further investigate the significance of each factor individually. We conduct a Nemenyi post hoc test with a single-step p -value adjustment, suitable to find statistically significant groups after the Friedman test. Table 6.2 shows the p -values for the attributes *glossiness*, *brightness*, *strength of reflections*, and *sharpness of reflections*; and both factors: *material category*, and *motion degree*.

We can observe a clear trend for all four attributes, there is a significant change in participants' perception when the stimuli changes from static to a middle motion degree ($p < 0.05$) and if the stimulus moves from static to the highest analyzed motion degree ($p < 0.05$). However, if the stimulus changes from a middle motion degree (45) to the highest motion degree (95), there are no significant changes in participants' answers ($p > 0.05$). This showcases the non-linear nature of human perception of material appearance [307] and tells us that linear changes in motion do not necessarily correlate with linear changes on the perception we have about the stimuli [343, 41]. However, the rubber material category does not show significant changes for attributes *glossiness*, and *strength of reflections* when the degree of motion changes from static to a middle motion degree ($p > 0.05$), but shows significant changes when we move from a static to the highest motion degree ($p < 0.05$). This could be due to the diffuse appearance of rubber, which lacks reflections. Therefore, small changes in motion do not produce perceptually visible changes in the specular appearance of the material, requiring higher motion degrees to perceive those changes.

Figure 6.5 shows the trends of participants' ratings for each of the six perceptual properties and motion degrees. As previously discussed *glossiness*, *brightness*, *sharpness of reflections* and *strength of reflections* attributes have significantly different answers under motion. For each attribute, we observe

Attribute	Material	Motion Pairs			Attribute	Material	Motion Pairs		
		0-45	0-95	45-95			0-45	0-95	45-95
Glossiness	Glass	0	0	0.209	Strength of reflections	Glass	0.006	0	0.357
	Metal	0.269	0.003	0.178		Metal	0.014	0.004	0.108
	Paint	0.017	0	0.373		Paint	0.029	0.012	0.946
	Plastic	0.004	0.800	0.892		Plastic	0.003	0.005	0.274
	Rubber	0.070	0.003	0.553		Rubber	0.056	0.040	0.991
	Stone	0.002	0	0.665		Stone	0.003	0	0.768
Brightness	Glass	0	0	0.988	Sharpness of reflections	Glass	0	0	0.592
	Metal	0.019	0	0.649		Metal	0.003	0.002	0.192
	Paint	0	0	0.645		Paint	0.002	0.071	0.703
	Plastic	0.002	0	0.380		Plastic	0	0.008	0.729
	Rubber	0.005	0	0.783		Rubber	0.006	0	0.728
	Stone	0	0	0.946		Stone	0.027	0.004	0.918

Table 6.2: Results from the Nemenyi post hoc test. In the left and right tables, the first column is the name of the attribute rated by the participants. The names of the material categories are placed in the second column. From the third to the fifth column we have the p-values for each change in motion degree.

a falloff as the motion degree is increased. This is expected as the blur introduced by the movement of the stimulus dims the image content, softens highlights. Also, if we observe the *sharpness of reflections* attribute, we see that material categories with clear reflections in their materials — like *glass* — have higher ratings in the static stimulus and a steeper decreasing slope. On the other hand, *matte* and *roughness* attributes, not considered statistically significant, have slight changes as we increase motion, showing an almost uniform behavior.

6.4 SECOND EXPERIMENT: BRIGHTNESS MAP CONSTRUCTION

Given the results of the previous experiments (see Section 6.3), where we found that the attributes that describe the specularity of material significantly change under motion, we further investigate the behavior of our perceived brightness. In the following section, we explain the steps carried on in order to understand the effect of motion on our perception of brightness.

6.4.1 Influence of brightness

To further investigate the influence of motion on perceived brightness, we run new experiments using rubber and plastic materials, including additional motion degrees, and adding more levels of luminance in the stimuli.

STIMULI We generate a series of stimuli for the materials rubber and plastic, with ten continuous-changed luminance levels (1 to 10) under seven different motion degrees (0, 5, 15, 45, 75, 95, 135). The stimuli used in this experiment for the material rubber can be seen in Figure 6.2.

PARTICIPANTS Similar to the previous user-study, we relied on Amazon Mechanical Turk to launch the experiment. A total of 210 unique workers took part in it. Users were unaware of the purpose of the experiment.

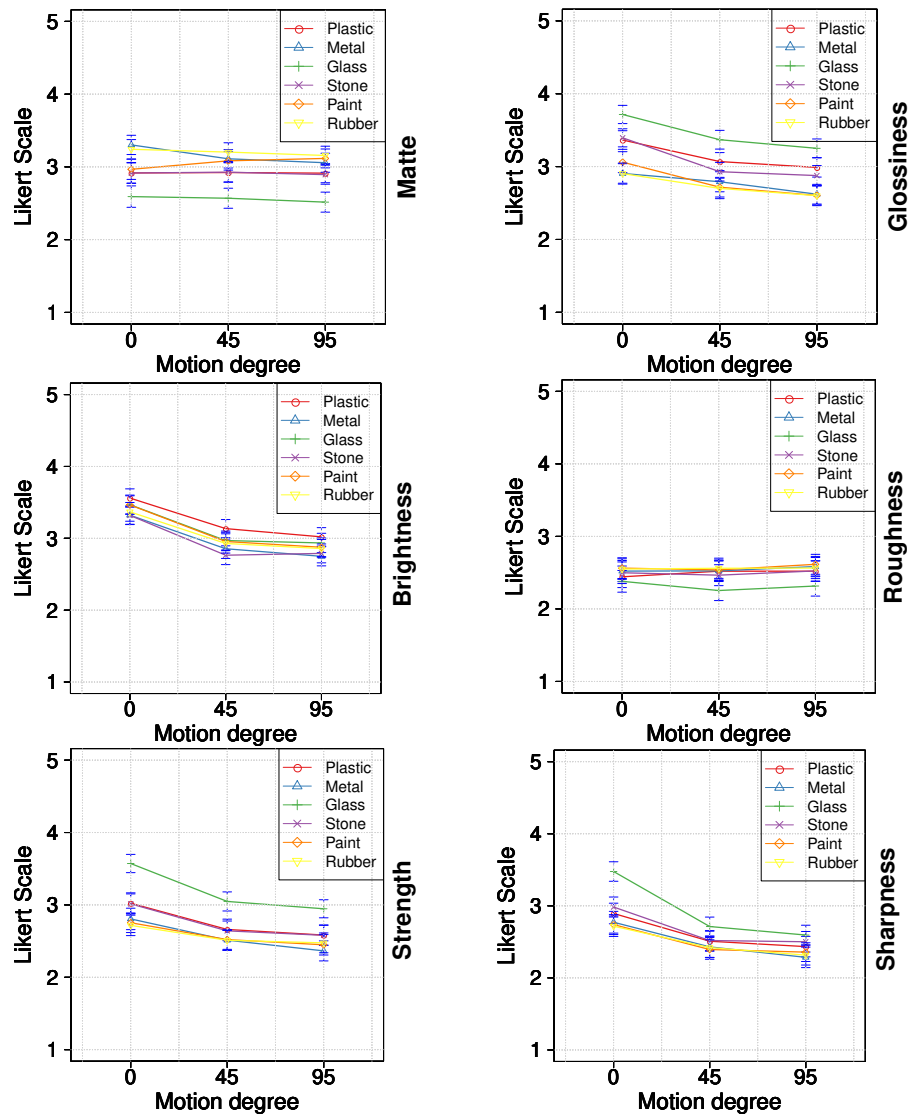


Figure 6.5: Average rating given in the user-study for each of the six perceptual attributes analyzed for each material type and for all the motion degrees. The x -axis shows the variation in motion of the stimuli, the y -axis displays the average rating given by the participants. Each color of the line plot represents one of the material categories. Error bars correspond to a 95% confidence interval. In attributes *glossiness*, *brightness*, *sharpness of reflections* and *strength of reflections*, the rating decreases while for *matte* and *roughness* it shows an almost uniform behavior. All the plots for each attribute can be found in the supplementary material.

PROCEDURE This experiment follows a similar scheme to the one explained in Section 4. Participants are asked to rate a subset of six different attributes (introduced in Section 6.3) on a continuous 5-point Likert-like scale. Also, although we are only interested in brightness, in order to avoid participants to know the real purpose of the experiment, they are asked to rate all the subset of 6 attributes. We reduce the number of attributes to rate from 14 to 6 in order to avoid distractions in the user [62]. Each experiment has 10 different trials containing all the brightness levels and random motion degrees. The trials are presented randomly to the participants. We collect

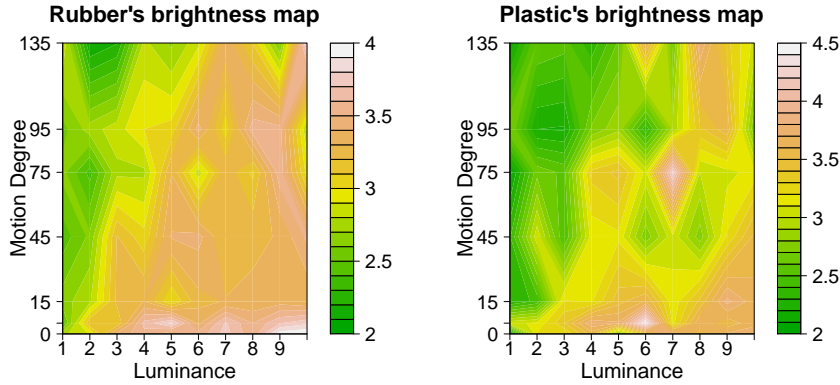


Figure 6.6: *Left*: Rubber brightness map created using the stimuli from the experiments. *Right*: Plastic brightness map. In both brightness maps, the x -axis is the luminance of the stimuli while the y -axis is its motion degree. Both figures show how as motion degree increases, brightness decreases non-linearly.

15 answers for each sphere, brightness, and motion degree. In the end, we gather a total of 2100 answers.

6.4.2 Analysis

After collecting all the data, we have a set of brightness' ratings for stimuli with variations in their motion degree and luminance. With it, we can build two-dimensional maps that tell us the average perceived brightness for each luminance and motion degree level.

BUILDING A BRIGHTNESS MAP In Figure 6.6, we can see the brightness maps for the materials rubber (left) and plastic (right) The brightness maps are constructed for all 10 luminance levels, in the x -axis, and 7 motion degrees, in the y -axis, of the stimuli. In order to build the brightness map, we get the average perceived brightness values of the participants using the ratings of the experiment. Since we cannot generate an infinite set of stimuli, the points that are not sampled are generated using linear interpolation.

LUMINANCE COMPENSATION Both brightness maps confirm the results from the previous experiment (Section 6.3), as motion increases our ratings of perceived brightness decrease. However, when the stimulus reaches a high motion degree our perception of its brightness keeps almost constant. Also, If we follow an isocontour in the brightness maps — the lines that have the same average perceived brightness —, we can observe how, in order to keep a constant perceived brightness, the slope of the isocontour bends. Moreover, if we compare both brightness maps, we can observe how the plastic has an overall less perceived luminance. We argue that this happens due to the blur in the highlights of the plastic, a specular material, produced by the effect of motion blur. While on the rubber, since it is a diffuse material, there are no significant highlights and the effect of motion has a reduced impact compared to the case of plastic.

LUMINANCE SPLIT Isocontours show how in order to keep the same level of perceived brightness, luminance levels have to increase non-linearly when motion degree is increased linearly. Moreover, if we follow the isocontour generated between the values 2 and 3 of luminance, we can see a clear separation; on the left, we would have the dim area with values of perceived brightness around 2.5 while on the right side of the isocontour we can see a constant area with perceived brightness of 3.5. These two regions are separated by the isocontour where our perceived brightness takes a value of 3. This exhibits the non-linear behavior of human perception, and also how, although the values of luminance and motion degree are changing, our perception of brightness remains constant.

6.5 DISCUSSION AND FUTURE WORK

6.5.1 Conclusion

We have presented an analysis of the effects of motion blur on our perception of a wide set of material attributes, for different material categories. To do it, we analyze 38,880 ratings (72 (*materials*) \times 3 (*motion degrees*) \times 30 (*answers/stimulus*) \times 14 (*attributes*)) given for six different material attributes: *brightness*, *glossiness*, *matte*, *roughness*, *sharpness of reflections* and *strength of reflections*.

A reasonable concern when using Mechanical Turk as a source for participants in user studies, is the possible effect of uncontrolled viewing conditions (such as display characteristics or viewing environment). Nevertheless, previous studies have shown that MTurk can actually be used in visual psychophysical experiments, and its results matched those under controlled lab conditions, since a large number of participants reduces variance (e.g. [126, 142]). Similarly, participants were not screened for proficiency in English, so it may be that some subtleties in the description of material appearance were lost.

Our results indicate that motion blur has a significant effect on our perception of attributes related to the specular nature of the material, namely *glossiness*, *brightness*, and *strength* and *sharpness of reflections*, whereas no significant influence exists on the attributes *matte* and *roughness*. The four aforementioned attributes—*glossiness*, *brightness*, and *strength* and *sharpness of reflections*—are given significantly lower ratings as the motion degree increases. This can be due to the removal of high frequencies of the stimuli as a consequence of motion blur, since it has been shown that these relate to specularity [71]. When looking at the nature of this influence through post hoc analyses, we observe that, among lower motion degrees the influence of motion blur on the perception of attributes is larger than among higher ones; actually, there is no significant difference between motion degrees 45 and 95, while there is between degrees 0 and 45. This seems to indicate a stabilization of perceived attribute magnitudes for higher motion degrees, i.e., from a certain motion degree, increased motion blur does not change the perception of these attributes.

Moreover, we select an attribute and material category to do a more in-depth analysis of the influence of motion blur on perceived brightness. In this second experiment, we not only vary the motion degree, but also the luminance of the material shown in the stimuli. We seek to observe how the isocontours (i.e., lines of constant brightness) behave as we vary both motion degree and luminance level. Despite the limited extent of our experiment, we

observe how the non-linear behavior with motion degree persists. Further, this behavior is different for different luminance levels, falling back to an almost constant (independent of motion degree) behavior for both very high and very low luminance levels. This intricate relationship suggests the need for more detailed experiments, covering a wider range of appearances and perceived attributes; this work takes the first steps in this direction, but a comprehensive model of the influence of motion blur in material perception remains as future work.

6.5.2 *Future work*

In this chapter, we have just investigated the effect of motion blur due to linear motion in our perception of material appearance; other kinds of motion, including rotation, multidirectional paths, accelerated movement, etc., would require further analysis, for which we hope our work can provide a solid basis. Similarly, we make the reasonable assumption that motion blurred images are a good proxy for actual moving stimuli for the purposes of our test.

Our stimuli have been rendered using homogeneous material files in MDL. Exploring how the missing information encoded in the heterogeneities of some materials (like stone, or wood) affects appearance perception is an interesting extension, not included in our work. Some materials, especially specular ones, may require a wider dynamic range than what a typical display provides, an aspect of appearance perception not covered in this chapter since our stimuli were tone mapped. Also, since users, in general, tend to avoid extrema in Likert-based tests, it would be interesting to find if using a 7-point scale would have an influence in their responses. Finally, extending this work to find a correlation between image statistics and the set of perceptual attributes analyzed remains unexplored. Human perception of appearance is a sophisticated process, not fully understood, which poses many challenges and opens interesting avenues of future research, and we hope our work will inspire future exploration of the influence of motion in our perception. For instance, our findings could be used to optimize the performance of rendering pipelines where motion can have a big influence like in the emerging fields of virtual reality or real-time ray-tracing.

Part V

INTUITIVE APPLICATIONS FOR APPEARANCE EDITING

This part is devoted to the development of intuitive applications that manipulate visual appearance. First half focuses on the well-known problem of relighting for the particular case of scenes with full-body humans. We leverage a large dataset of synthetically generated humans, precomputed radiance transfer, and deep neural networks to propose a framework that outperforms state of the art in single-image full-body human relighting. Second half explores a framework for intuitive editing of material appearance, from RGB images, by using material attributes. We rely on crowd-sourced data on such attributes together with a newly proposed generative neural network architecture to obtain a plausible and intuitive single-image material editing framework.

This chapter presents a single-image data-driven method to automatically relight images with full-body humans in them. Our framework is based on a realistic scene decomposition leveraging precomputed radiance transfer (PRT) and spherical harmonics (SH) lighting. In contrast to previous work, our framework lifts the assumptions on Lambertian materials and explicitly model diffuse and specular reflectance in our data. Moreover, we introduce an additional light-dependent residual term that accounts for errors in the PRT-based image reconstruction. We introduce a new deep learning architecture, tailored to the decomposition performed in PRT, that is trained using a combination of L_1 , logarithmic, and rendering losses. Our model outperforms the state of the art for full-body human relighting both with synthetic images and photographs.

This work was done as part of both internships in Adobe Research, which continued as a collaboration afterwards. This was presented at the *Eurographics Symposium on Rendering* (EGSR) [173]. While I led this line of research (with the guidance and help of my supervisors Belén Masiá and Diego Gutiérrez). Xin Sun also helped giving invaluable feedback, setting up the rendering engine to generate the synthetic dataset of humans, and collaborating with the manuscript text.

M. Lagunas, X. Sun, J. Yang, R. Villegas, J. Zhang, Z. Shu, B. Masia, & D. Gutierrez
Single-image Full-body Human Relighting
Eurographics Symposium on Rendering (EGSR), 2021

7.1 INTRODUCTION

The growth in mobile computing, together with the increasing demand for visual social media has led to a tremendous rise in the popularity of consumer digital photography. In full-body photographs lighting plays an important role in transmitting the desired appearance of the subject, and changes in the illumination can lead to drastically different renditions. However, these photographs usually lack controlled illumination conditions.

We present a single-image relighting method that acts as a post-processing step, allowing a casual user to plausibly change and manipulate the illumination on a subject in a photograph. Human relighting usually benefits from multiple images as input, and requires solving an inverse rendering problem; in the general case, illumination information needs to be disambiguated from geometry and material appearance, based on simple pixel values. This is a well-studied but ill-posed problem, for which no definite solution exists. This chapter takes a data-driven approach to the problem, requiring only one photograph and a user-specified target illumination as input (see Figure 7.1). Our method relies on precomputed radiance transfer [291] (PRT) and spherical harmonics lighting [254] (SH). Based on this, a convolutional neural network (CNN) decomposes the image into its albedo, illumination, and light transport components; from which the shading can be easily computed. Disentangling the illumination from all other factors in the scene allows for effective relighting, while the PRT-based scheme enables fast, efficient rendering. Our work lifts the assumption of Lambertian materials present in previous single-image human relighting methods [274, 148]. We model the PRT decomposition in our framework by approximating material reflectance using an Oren-Nayar [237] and GGX microfacet model [325] for the diffuse



Figure 7.1: Relighted results given a single image as input for different illumination maps. Please refer to Figure 7.12 for more details about the reconstructions.

and specular components, respectively. In addition, we extend the image reconstruction formulation by adding a *residual* term learned by our model, which accounts for errors in image reconstruction that would be obtained using only the terms proposed by PRT.

To train our model, we create a synthetic dataset containing almost 140,000 images with a rich variety of humans (105), poses (5), and illumination maps (266). We quantitatively and qualitatively evaluate relighting results on both synthetic images and real photographs, and perform extensive ablation experiments to validate our design choices in the model architecture, reflectance model for data generation, and loss functions. Compared with the current state of the art in full-body single-image human relighting [148], our model yields more accurate reconstructions of relighted images for both synthetic images and real photographs.

7.2 BACKGROUND

In this section, we briefly review the building blocks of our technique: Spherical harmonics (SH) lighting [254], and precomputed radiance transfer (PRT) [291].

PRT [291] and SH lighting [254] enable rendering dynamic low-frequency environments with realistic highlights and real-time shading. They estimate the amount of radiance reflected at a point in the scene by solving a simplified version of the rendering equation:

$$R(x) = \int_{S^2} L(x, \omega_i) T(x, \omega_i) d\omega_i, \quad (7.1)$$

where R is the reflected radiance or image intensity computed over the sphere S^2 of incoming directions ω_i , L is the incoming light at point x from direction ω_i , and T is a transport function computed for each vertex that includes the material reflectance f_r , visibility term V that is 1 if the point is not occluded and 0 otherwise, and the cosine term which uses the normal \mathbf{n} at point x . The function T can be expressed as:

$$T(x, \omega_i) = f_r(x, \omega_i) V(x, \omega_i) (\omega_i \cdot \mathbf{n}). \quad (7.2)$$

The formulation presented by PRT expands the illumination L and the transport T using (real) spherical harmonics basis functions $Y_{l,m}$, such that:

$$\begin{aligned} L(x, \omega_i) &= \sum_{l=0}^{\infty} \sum_{m=-l}^l L_{l,m}(x) Y_{l,m}(\omega_i), \\ T(x, \omega_i) &= \sum_{l=0}^{\infty} \sum_{m=-l}^l T_{l,m}(x) Y_{l,m}(\omega_i), \end{aligned} \quad (7.3)$$

where $L_{l,m}$ and $T_{l,m}$ are the corresponding coefficients for illumination and transport, respectively (see [253, Sections 3 and 4] for additional details on how to obtain $T_{l,m}$ and $L_{l,m}$). The integral in Equation 7.1 then becomes:

$$R(x) = \sum_{l=0}^{\infty} \sum_{m=-l}^l L_{l,m}(x) T_{l,m}(x). \quad (7.4)$$

This formulation has two advantages: It allows to approximate the rendering equation as a fast dot product, and it disentangles the illumination and the transport in the scene. In this way, relighting a scene only requires computing the coefficients of the new illumination $L'_{l,m}$, while keeping $T_{l,m}$ fixed.

Traditionally, relighting methods based on the estimation of illumination and transport coefficients from a single image soften the problem by assuming that the scene has a light source at a sufficient distance to neglect the angular variation between points, i.e., $L(x, \omega_i) \approx L(\omega_i)$. They also estimate a transport function T encoding only the cosine term [274], or the cosine term together with the visibility function [148]. These methods assume all materials to be Lambertian, removing the reflectance term from the transport $T_{l,m}(x)$, and modeling it as a constant for each point of the scene represented by the albedo $\rho(x)$. With this, expressing $L_{l,m}$ as a vector \mathbf{L} and $T_{l,m}(x)$ as a vector per point of the scene $\mathbf{T}(x)$, $R(x)$ can be approximated as (hereafter, we omit the dependency on x for clarity):

$$R \approx \underbrace{\rho}_{\text{albedo}} \cdot \underbrace{(\mathbf{T}^T \cdot \mathbf{L})}_{\text{shading } S}, \quad (7.5)$$

where the dot product between transport and illumination yields the shading $S(x)$ of a point in the scene, then scaled by the albedo ρ . The error of the approximation will be related to the number of coefficients used to estimate the illumination and transport in Equations 7.3 and 7.4; this number of coefficients depends on the number of terms used to approximate the infinite term summation of Equation 7.4, $l = [0..N]$.

To increase realism in the inferred and rendered images, we lift the Lambertian material assumption of previous work and include a better approximation of material reflectance in the transport function \mathbf{T} . We approximate the reflectance term in Equation 7.2 by keeping the albedo ρ as a constant and using a white material with an Oren-Nayar [237] for the diffuse component, and a GGX model with Smith shadowing factor and Fresnel [325] for the specular reflection. Then, we use Equation 7.3 to encode such reflectance in a new transport function \mathbf{T} , later used to render new images with Equation 7.5. As Figure 7.2 shows, this allows to better capture the directionality of specular reflections. Our reflectance model employs the following parameters: albedo, roughness, metallic, and transparency (refer to Section 7.4 for additional details). Both the Oren-Nayar and the GGX models share the same roughness parameter. The final reflectance model is



Figure 7.2: Comparison between the data generated with our framework and that of the recent work by Kanamori and Endo [148], used to train the respective models. Our transport function \mathbf{T} takes into account angular dependencies in the reflectance term, better capturing specular reflections and improving high-frequency details in the shading.

defined as a combination of up to seven BSDFs, which can be either a diffuse Oren-Nayar microfacets model or a specular GGX model.

7.3 OUR IMAGE RECONSTRUCTION FORMULATION

This section describes our image reconstruction formulation, including the motivation behind the addition of a new residual term.

Since using a large number of basis coefficients in Equation 7.5 to approximate R with a low error is computationally expensive, we introduce an additional residual vector \mathbf{E} , leading to:

$$R \approx \underbrace{\rho}_{\text{albedo}} \cdot \underbrace{(\mathbf{T}^T \cdot \mathbf{L})}_{\text{shading } S} + \underbrace{(\mathbf{E}^T \cdot \mathbf{L})}_{\text{residual } E}, \quad (7.6)$$

where the dot product between the residual vector \mathbf{E} and the illumination \mathbf{L} yields a residual value per point $E(x)$. Again, the dependency on x is omitted for clarity, but Equation 7.6 applies to each point in the scene, yielding the corresponding images; in the following, we will use S to denote the shading image, and E for the residual image. The residual vector \mathbf{E} does not have a physical meaning; instead, it is a set of learned coefficients that aim to model the errors in image reconstruction that we would obtain using only the terms (albedo, transport, and illumination) with a limited number of coefficients.

7.3.1 Problem Formulation

Our main goal is to relight an image ψ with a full-body human in it, given a user-specified target illumination \mathbf{L}' :

$$\hat{\psi} = \mathcal{R}(\psi, \mathbf{L}'), \quad (7.7)$$

where \mathcal{R} is a relighting function, and $\hat{\psi}$ is the resulting relighted image with target illumination \mathbf{L}' .

Using a model such as the one in Equation 7.6, one can change \mathbf{L} to \mathbf{L}' to obtain the relighted image. However, given a single image as input, the transport \mathbf{T} , illumination \mathbf{L} , residual \mathbf{E} , and albedo ρ are unknown. To obtain an approximation of \mathbf{T} , \mathbf{L} , \mathbf{E} , and ρ , we introduce the parametric function \mathcal{G} , which takes as input the image ψ and a set of parameters β :

$$\{\mathbf{T}, \mathbf{L}, \mathbf{E}, \rho\} \approx \mathcal{G}(\psi, \beta). \quad (7.8)$$

In particular, we model \mathcal{G} using a convolutional neural network whose parameters are represented by β . Note that \mathcal{G} tries to approximate each of the terms $\{\mathbf{T}, \mathbf{L}, \mathbf{E}, \rho\}$ irrespective of the underlying reflectance model previously used to generate them. With the output of \mathcal{G} and a given user-specified illumination \mathbf{L}' , we can use Equation 7.6 to directly approximate the relighting function \mathcal{R} .

7.4 DATASET

To learn the parametric function \mathcal{G} introduced in Section 7.3 we have created a synthetic human image dataset of almost 140,000 images including a rich variety of humans, poses, and illuminations, which we describe in this section.

HUMAN 3D MODELS Existing models captured using photogrammetry mostly consist only of diffuse and normal maps. To fully exploit the capabilities of our framework and go beyond Lambertian materials, we purchase rigged 3D human models and clothing from the DAZ website [52], which include realistic materials and texture maps for diffuse color, specular, opacity, roughness, metallic, translucency, and normals. In total, we collected 105 different clothed models; augmented with five poses each, this yields a total of 525 different renditions. For each pose we simulate cloth interaction after posing the model, and, to foster diversity, perform subtle random changes to the hue of the diffuse color.

ILLUMINATIONS We used freely-available spherical high-dynamic range images (HDRIs) from HDRIHaven [121], corresponding to both indoor and outdoor scenarios. To normalize the HDRIs, we compute a *reference radiance* for each image by obtaining the mean shading in Equation 7.5, where \mathbf{L} are the coefficients of the HDRI, and \mathbf{T} is obtained analytically by sampling all unit directions in the sphere. We scale all the illuminations \mathbf{L} to have a reference radiance in the range $[0.7, 0.9]$. In total we gathered 266 different HDRIs.

RENDERING We used Monte Carlo path tracing to render realistic images and to obtain the transport vector \mathbf{T} for each scene. To generate \mathbf{L} for each illumination, we integrate over the unit sphere of directions. We fix $N = 4$ ($l = [0..4]$), which leads to 25 spherical harmonics coefficients in \mathbf{T} and \mathbf{L} (in contrast, the work of Kanamori and Endo [148] estimates only Lambertian materials and uses $N = 2$). Among all the available maps defining reflectance for each purchased model, during rendering we employ the albedo (diffuse color), roughness, metallic, and transparency maps. In total, we render 139,650 different scenes. For each scene, we generate: Its path-traced (PT) image, the PRT image computed using Equation 7.5, an alpha mask of the human, the shading, the normals, the albedo, and a material map containing the roughness, transparency, and metallic, each of them encoded in separate channel of an RGB image. All images are rendered with a resolution of 768×768 pixels; using 256 samples per pixel for the PT image, and 1,024 for the transport \mathbf{T} and all other scene properties. Figure 7.3 shows two samples from our dataset, cropped down from the squared aspect ratio.

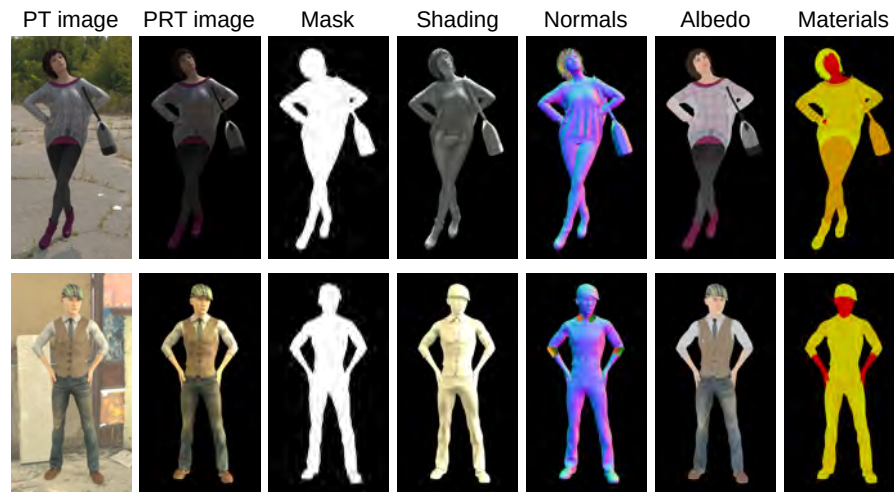


Figure 7.3: Two examples in our dataset. For each scene we obtain its path-traced (PT) rendered image, its PRT image rendered with our image reconstruction formulation, the alpha mask, the shading, the normals, the albedo, and a material map describing the roughness, transparency, and metallic (each encoded in a separate channel of an RGB image).

7.5 OUR MODEL

In this section we explain our model architecture and its components, together with an intuition behind our design choices; in addition, we provide details on our training, hyper-parameters, and loss function.

7.5.1 Model Architecture

To represent our parametric function \mathcal{G} we use a convolutional neural network based on a UNet-like model [257]. Figure 7.4 shows an overview. It consists of a shared encoder that receives the input image ψ , and several decoders responsible for estimating albedo ρ , transport \mathbf{T} , residual coefficients \mathbf{E} , and the illumination of the input image \mathbf{L} . We add skip-connections between the shared encoder and each decoder to encourage better reconstructions, except for the light decoder. Last, we have a rendering layer based on Equation 7.6 that generates the shading S , the residual E , and the final relighted image $\hat{\psi}$.

SHARED ENCODER Our encoder has a standard architecture consisting of several convolutional blocks with batch-normalization (BN) that sequentially decrease the resolution of the features by a factor of two. The features between convolutional blocks are used as skip-connections with the decoders.

DECODERS Each decoder has a *residual block* (similar to ResNet [123]), and a *generator block* except for the light decoder that only has a generator block. The generator block varies between decoders. The output of the albedo, transport, and residual coefficients decoder has the same spatial resolution as the input image. We only add batch-normalization to the albedo decoder. The architecture of each generator is as follows (see also Figure 7.5):

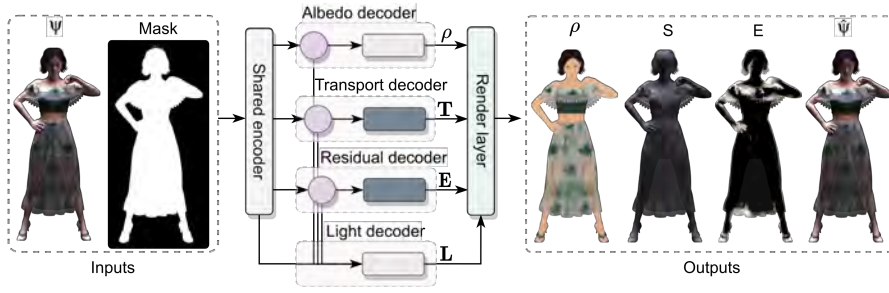


Figure 7.4: Our model architecture. The masked input image goes through a shared encoder that converts it into a feature map. Such feature map simultaneously serves as an input to the albedo, transport, and residual decoders. The three decoders output the albedo ρ , transport \mathbf{T} , and residual coefficients \mathbf{E} , respectively. The features from these three decoders and from the shared encoder are concatenated and fed to the light decoder, which outputs the illumination coefficients \mathbf{L} . Last, the rendering layer outputs the albedo (equal to the output from the albedo decoder), shading, residual image, and the final relighted image.

- The albedo decoder has several convolutional blocks with skip connections. In each convolutional block features are scaled by a factor of two. The output of the albedo decoder is clamped to lie in the range $[0, 1]$.
- To properly capture geometry and material reflectance in the scene, a good estimation of the transport matrix \mathbf{T} is needed. The transport and residual decoders feature a generator tailored for the PRT decomposition in Equation 7.6. Deep neural networks, by design, add non-linear functions that clamp negative values. However, the transport coefficients are defined with both positive and negative values. Thus, we would rely on the last convolution without non-linearities to generate all the negative content in the coefficients. To give additional degrees of freedom to the decoders, we decompose the coefficients as $\mathbf{T} = \mathbf{T}^+ - |\mathbf{T}^-|$ where \mathbf{T}^+ corresponds to the positive part and $|\mathbf{T}^-|$ is the absolute value of the negative part. Instead of directly predicting \mathbf{T} , we add two generators (similar to the albedo one) to predict \mathbf{T}^+ and $|\mathbf{T}^-|$, respectively, and later we reconstruct the coefficients \mathbf{T} . We apply a similar strategy to the residual coefficients \mathbf{E} .
- The light decoder differs from the previous as its input is the output of the shared encoder and the residual blocks of the albedo, transport, and residual decoders. Those features go straight to a generator that follows a similar decomposition as for the transport and residual decoder, however, the generator architectures differ. The generator has several convolutional blocks that reduce the spatial dimensions of the features by a factor of two. After the convolutions, we perform an average pooling making the features one-dimensional, and a fully-connected layer outputs the positive and negative illumination coefficients in each generator, with shape $3 * 25$ (25 being the total number of coefficients when $N = 4$). Then, we reconstruct \mathbf{L} using the positive and negative part.

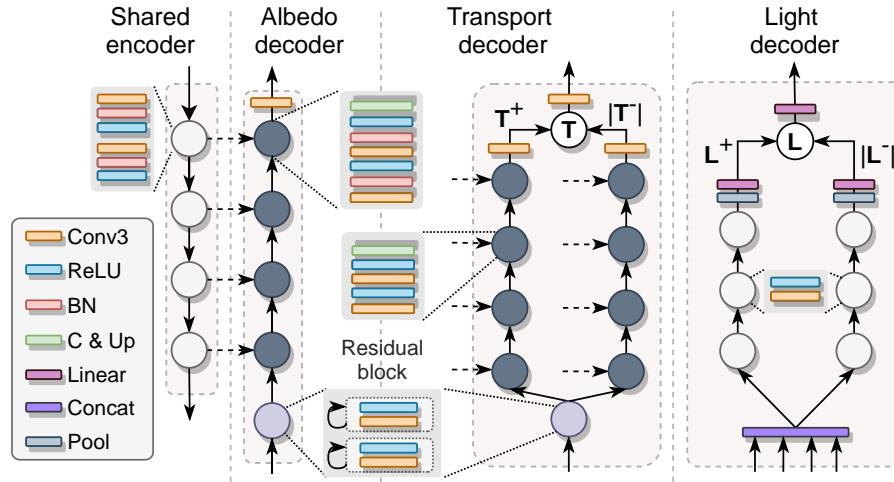


Figure 7.5: Workflow of each component of our model. The shared encoder contains several convolutional blocks that reduce the spatial dimensions by two and output a feature map of the input. Such feature map goes to the albedo, transport, and residual decoders. Each decoder (except for the light one) has a residual block and a generator. The generator concatenates skip-connections and upscales (C & Up) the spatial resolution of the features. The output of the decoders has the same spatial resolution as the input image. Last, the light decoder uses the features of the encoder, together with the features from the residual block of each decoder, to predict the illumination in the scene.

7.5.2 Training

The dataset in Section 7.4 is split into training and validation, where we select 7 clothed models (with all their poses) that are representative of challenging scenes as the validation set. The rest of the humans with their poses are used for training. The input to our model are images rendered with PRT, where we crop the human using the bounding-box defined by the mask with a padding of 20 pixels. Since our network is fully-convolutional it allows inputs of arbitrary resolution. We normalize the image pixels to lie in the range $[-1, +1]$ and multiply it by the alpha mask before forwarding it through the model. For training we use the Adam optimization algorithm [163] with the decoupled weight decay regularization [198]. The learning rate has a value of $5 \cdot 10^{-5}$. We set an effective batch size of 16. We use the PyTorch framework [244] with PyTorch-Lightning [73] to design our model and experiments. The model is trained for 25 epochs on eight Tesla V100-SXM2-16GB, lasting 55 hours approximately.

7.5.3 Loss Functions

Our loss function \mathcal{L} can be expressed as:

$$\mathcal{L} = \mathcal{L}_\rho + \mathcal{L}_T + \mathcal{L}_L + \mathcal{L}_S + \mathcal{L}_{\hat{\phi}}, \quad (7.9)$$

where each term supervises the prediction of albedo, transport, illumination, shading, and the final relighted image. Note that the residual coefficients are not directly supervised. Instead, we let the network freely learn a set coefficients \mathbf{E} that aim to improve the quality of the rendered images. Each

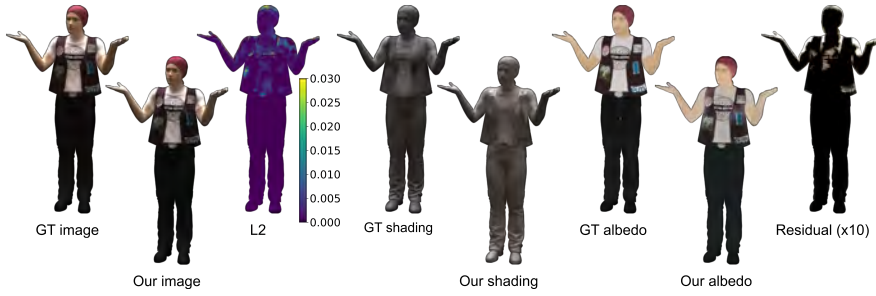


Figure 7.6: Example result of our model for a synthetic image (see also Table 7.1, synthetic images). Neither the human nor the illumination were used for training. We show direct comparisons with the groundtruth (GT), the L2 error in the final image, and our residual term scaled by a factor of 10 for visualization purposes.

of the terms in \mathcal{L} is additionally composed of different losses. We linearly combine the different terms using a weight of 1 for all of them.

- **Reconstruction loss (\mathcal{L}_{L1})** We apply an L1 loss function to each predicted map with respect to groundtruth data. To encourage a better reconstruction, we leverage the architecture tailored for PRT rendering, and additionally include an L1 loss between the positive and negative coefficients in \mathcal{L}_T and \mathcal{L}_L .
- **Render loss (\mathcal{L}_r)** The terms in Equation 7.6 are computed using the albedo, transport, illumination, and residual vectors. For each of those vectors (except the residual E), there is both a predicted (which is being learned) and a groundtruth vector. To increase robustness, we introduce in \mathcal{L}_S and $\mathcal{L}_{\hat{\psi}}$ an L1 error term for each possible way of generating the shading and relighted image in Equation 7.6 from the predicted and groundtruth vectors.
- **Log loss (\mathcal{L}_{\log})** The transport, and the illumination coefficients have an unbounded range. To compress it, we apply a logarithmic loss of the following form:

$$\mathcal{L}_{\log} = \|\log(|x| + 1) - \log(|\hat{x}| + 1)\|_2^2$$

in \mathcal{L}_T and \mathcal{L}_L . We apply $|x|$ in the logarithmic loss to avoid errors on the negative values of the coefficients. We leverage the PRT decomposition to apply the logarithmic loss also to the positive and negative decomposition of transport and illumination.

7.6 RESULTS

We show and evaluate results of our model on both synthetic images, where groundtruth data is available, and real photographs. Throughout the evaluation, we show the reconstructed albedo ρ , shading S (resulting from the combination of transport T and target illumination L' , see Equation 7.6), the final rendered result $\hat{\psi}$, and the residual image E . We also include ablation studies to clearly demonstrate the influence of each component in the final relighted images

SYNTHETIC IMAGES We use the validation subset of our dataset (see Section 7.5.2) rendered with six new illuminations not used for training: *ennis*,

Model	ALBEDO			SHADING			IMAGE		
	L1 (x100)	L2 (x100)	PSNR	L1 (x100)	L2 (x100)	PSNR	L1 (x100)	L2 (x100)	PSNR
Ours	2.88	0.44	24.18	3.77	0.71	24.05	1.64	0.19	28.94
Kanamori and Endo	4.95	1.19	20.68	6.75	1.90	18.29	2.94	0.47	26.06
Without E	3.67	0.66	23.05	6.71	2.74	17.89	1.97	0.24	27.13
Without PRT decomposition	4.54	1.00	21.08	10.57	5.69	14.43	2.13	0.24	26.80
Without \mathcal{L}_{\log}	4.02	0.83	21.84	10.34	5.35	14.55	2.14	0.24	27.82
With $N = 2$	3.31	0.58	23.33	8.60	4.21	16.18	1.83	0.22	28.33
With T^*	3.68	0.76	21.74	7.53	3.54	16.65	2.25	0.31	27.29
Lambertian materials	3.58	0.68	22.66	7.22	2.91	17.09	1.91	0.21	27.92

Table 7.1: Quantitative results of our model for synthetic images and real photographs, measured with three metrics: L1 and L2 distances, and PSNR. Note that the L1 and L2 metrics have been scaled by a factor of 100. We also include a comparison to the model of Kanamori and Endo [148], which our model consistently outperforms. Boldface highlights the best result in each case.

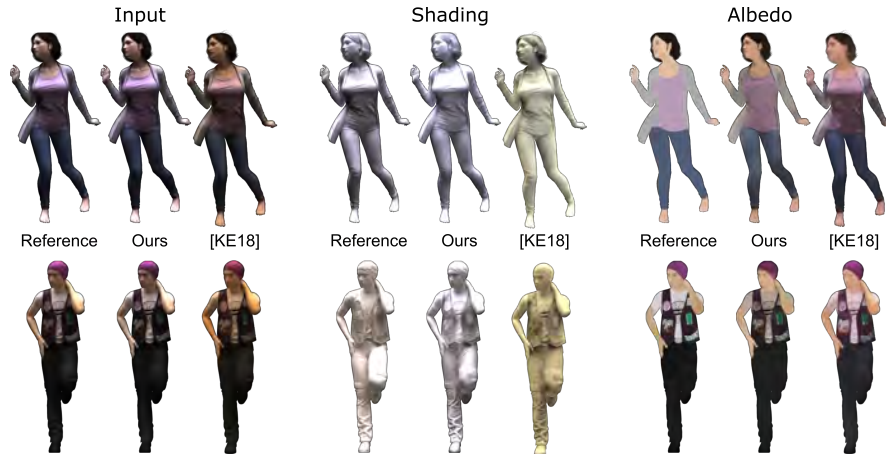


Figure 7.7: Comparison between our model and the model provided by Kanamori and Endo [148] in two examples of the validation dataset. We can see how our model outperforms them rendering the input image, albedo, and shading. Note that the shading encodes both the transport and the illumination of the scene.

grace, pisa, doge, glacier and *uffizi* [168]. We render the final relighted (target) image using the predicted illumination of the scene to reconstruct the shading and the residual. Since groundtruth data is available, we also compute quantitative error measures for the albedo, shading, and final rendered image. Specifically, we compute the L1 and L2 distances, as well as PSNR, averaged across the dataset. Table 7.1 (synthetic images) shows the results, including a comparison with the pretrained model of the recent work by Kanamori and Endo [148]. Our more complete material reflectance formulation, together with our residual term (see ablation studies in Subsection 7.6.1) lead to significantly lower L1 and L2 values, and higher PSNR for the albedo and shading, as well as the final relighted image. Figure 7.6 shows a direct comparison of our reconstructed image with the groundtruth; both images match with a very small L2 error. Figure 7.7 shows a comparison between our model and the pretrained model given in the work of Kanamori and Endo on synthetic images. We can see how our model better estimates the shading and albedo, leading to more accurate results where directional effects are better reproduced (see the highlights in the face of the first image, for instance).

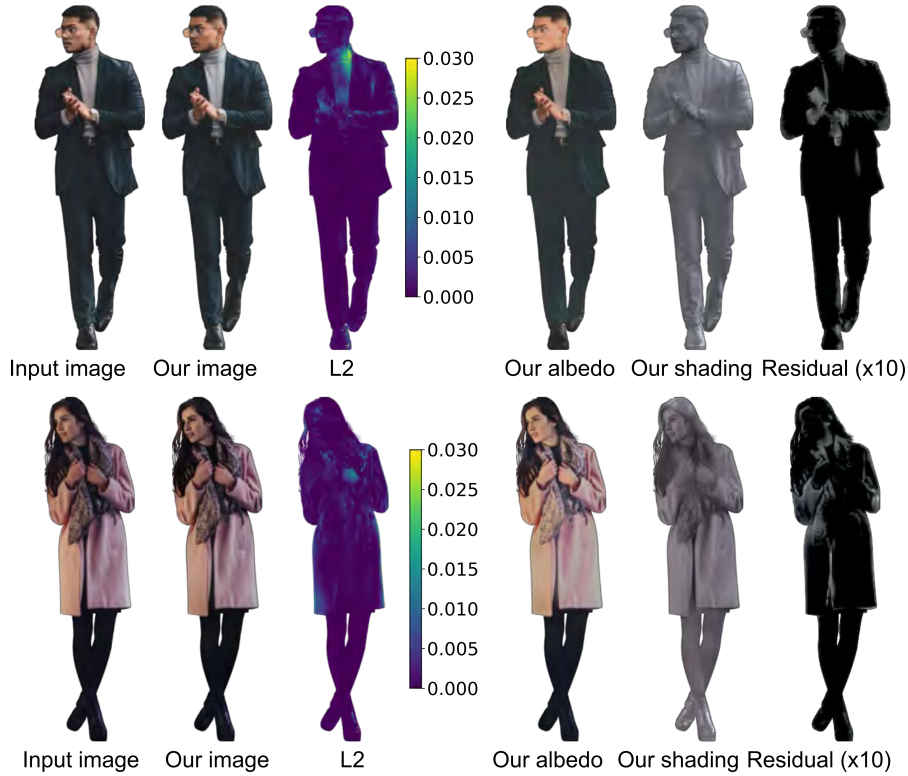


Figure 7.8: Example results of our model on real photographs (see also Table 7.2). For each image, from left to right: groundtruth input image, resulting image relighted with our model, L2 error, albedo, shading, and residual term scaled by a factor of 10 for visualization purposes.

REAL PHOTOGRAPHS To test our model on real photographs we use free-license images downloaded from Unsplash [313]. To obtain the alpha mask we rely on freely available APIs [147]. In total we collected 10 different images with a single human in them. Error metrics for the resulting rendered images, averaged over the 10 photos, can be found in Table 7.1 (real photographs). As with synthetic images, our results significantly outperform previous work [148]. Maybe surprisingly, the error metrics indicate better results with real photographs (both for our method and using the pretrained model of Kanamori and Endo) than using synthetic images. This is possibly due to the fact that the synthetic validation dataset contains some quite extreme illuminations (e.g., *glacier* or *grace*), while the photographic dataset has more natural illuminations that the two models are able to reproduce better. Figure 7.8 shows the reconstruction performed by our model for two different input photographs, including albedo and shading components, while a direct comparison with previous work is shown in Figure 7.9. Again, we see how our model is able to better capture directional effects (see, e.g., the faces or the highlights in the jackets) and overall produce more accurate reconstructions. Finally, in Figure 7.12 we show a variety of relighting results under different illuminations (refer to the Appendix D.1 for additional results). For each input photo and illumination map we show the final relighted image, and the reconstructed shading and residual terms.

Model	L1 (x100)	L2 (x100)	PSNR
Ours	1.12	0.08	31.46
Kanamori and Endo	2.14	0.20	28.38
Without E	2.64	0.29	26.24
Without PRT decomposition	2.55	0.30	26.66
Without \mathcal{L}_{\log}	2.31	0.25	26.85
With $N = 2$	2.08	0.18	27.76
With \mathbf{T}^*	1.75	0.14	29.43
Lambertian materials	1.82	0.18	29.15

Table 7.2: Quantitative results of our model and Kanamori and Endo [148] using real photographs, measured with three metrics: L1 and L2 distances, and PSNR. Note that the L1 and L2 metrics have been scaled by a factor of 100. Boldface highlights the best result in each case.

7.6.1 Ablation Studies

We evaluate the contribution of our design choices with a series of ablation experiments performed on both the synthetic images and the real photographs. In particular, we first compare the performance of our model (*Ours*) without the residual generator predicting \mathbf{E} (*Without E*) and without including the PRT decomposition in the architecture of the generators (*Without PRT decomposition*). Then, we evaluate the impact of the logarithmic loss \mathcal{L}_{\log} in the prediction of \mathbf{T} and \mathbf{L} (*Without \mathcal{L}_{\log}*), as well as the performance of our model when using only nine coefficients (*With $N = 2$*). To avoid using a constant albedo in Equation 7.6, we combine the different terms that define reflectance (*With \mathbf{T}^**) into a single vector $\mathbf{T}^* = (\rho * \mathbf{T} + \mathbf{E})$. Last, to showcase the benefit of our reflectance, we have trained a model using purely Lambertian materials in our data (*Lambertian materials*).

Tables 7.1 and 7.2 show the results (including albedo and shading for synthetic images) for the L1, L2, and PSNR metrics for all the ablation studies. All options yield significantly inferior results when compared with our full model. Figure 7.10 further illustrates this on an example using a real photograph. One could think that the model *With \mathbf{T}^** would obtain better performance since it does not need to assume a constant albedo ρ in the reflectance. However, \mathbf{T}^* requires estimating 25 different RGB maps (with $N = 4$), leading to additional complexity that hinders convergence and produces higher errors.

7.7 DISCUSSION

We have presented a model for human relighting that requires a single image as input. We lift the assumption on Lambertian materials and include a better approximation of material reflectance in our transport function. Moreover, we introduce an additional residual term which further mitigates errors in the PRT-based final reconstruction. This additional term becomes increasingly relevant for challenging illuminations, such as backlighting, where the overall dark appearance of the image does not allow for an accurate estimation of the PRT terms. The resulting errors are absorbed by our residual, helping







Input	Ours	[KE18]	Input	Ours	[KE18]
					
L1 (x100)	2.29	5.17	L1 (x100)	1.40	2.15
L2 (x100)	0.19	0.72	L2 (x100)	0.10	0.19
PSNR	27.27	21.41	PSNR	29.83	27.27

Figure 7.9: Image reconstructions obtained by our model, and the model provided by Kanamori and Endo [148]. We can see how our model outperforms them in the three metrics (see also Table 7.2). Note that the L1 and L2 metrics have been scaled by a factor of 100. In addition, our model better captures skin and cloth albedo, and the directionality of the illumination.



Figure 7.10: Reconstruction results obtained on the different ablation experiments. We can clearly observe how our full model better captures the appearance of the input photograph.

to produce good final reconstructions. Overall our results show compelling estimations of albedo and shading (transport and illumination), leading to accurate relighting reconstructions for both synthetic images and real photographs.

Nevertheless, our work is not free of limitations. Figure 7.11 shows a difficult case with a real photograph as input. While our reconstruction is still plausible, the strong presence of stray light (especially on top) leads to an excessively flat, milky estimation of the albedo in the head and shoulders area. Also, our shading reconstruction carries traces of texture details in the T-shirt, which remains an open problem in intrinsic images decomposition.

Human relighting poses many challenges not fully investigated in this chapter. Besides making the model more robust to poorly lit input images, being able to take into account other lighting effects such as subsurface scattering [143], anisotropy in cloth materials [7], or more complex reflectance models, remain interesting open problems. Moreover, one implicit problem of SH-based lighting is the need for a large number of coefficients to reconstruct high-frequency details. While we mitigate this problem by introducing the residual term, complex high-frequency effects are still an open challenge.

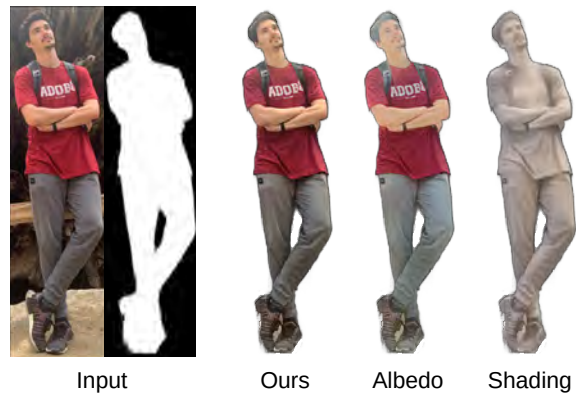


Figure 7.11: Example of the limitations of our model. The strong presence of stray light in the input image leads to an excessively flat albedo, seen especially in the head and shoulders area, while some texture details appear in the shading image.

Another exciting avenue of future work is to extend the potential of our approach, for instance by using contrastive loss functions, or proposing self-supervised schemes that would avoid having to generate additional synthetic data.



Figure 7.12: Relighting results for two different illuminations (*ennis*, and *pisa*) and five different input images. Last two columns feature the same illumination and two different rotations. In each case, we show the relighted image, and the reconstructed shading and residual terms. Our model is capable of producing a compelling relighting result for a varied set of input images and illuminations, including both indoors and outdoors cases. The residual term has been scaled by a factor of 10 for visualization purposes.

Single-image appearance editing is a challenging task traditionally requiring the estimation of additional scene properties such as geometry or illumination. Moreover, the exact interaction of light, shape, and material reflectance that elicits a given perceptual impression is still not well understood. This chapter presents an image-based editing framework that allows to modify the material appearance of an object by increasing or decreasing high-level perceptual attributes, using a single image as input. Our framework relies on a two-step generative network, where the first step drives the change in appearance and the second produces an image with high-frequency details. For training, we augment an existing material appearance dataset with perceptual judgements of high-level attributes, collected through crowdsourced experiments, and build upon training strategies that circumvent the cumbersome need for original-edited image pairs. We demonstrate the editing capabilities of our framework on a variety of inputs, both synthetic and real, using two common perceptual attributes (*Glossy* and *Metallic*), and validate the perception of appearance in our edited images through a user study.

After a first reviewing cycle, this work was referred to the *Computer Graphics Forum* (CGF) [56]. While I was not the lead author of this line of work; I collaborated by rendering the database used to train the models, and by taking part in the technical decisions and ideas regarding the novel generative neural network architecture.

J. Delanoy, M. Lagunas, J. Condor, B. Masia, & D. Gutierrez
A Generative Framework for Image-based Editing of Material Appearance using Perceptual Attributes
Computer Graphics Forum (CGF), major revisions, 2021

8.1 INTRODUCTION

Material appearance is one of the most important properties that determine how we perceive an object. The visual impression that it elicits, whether it appears metallic, glossy, or matte, strongly impacts how we manipulate such objects and expect them to behave. This appearance does not only depend on the intrinsic properties of the material itself, but also on external factors such as the geometry or the illumination of the scene. Editing material appearance based on a single image is therefore a very challenging task. A common approach is to estimate illumination, geometry, and reflectance properties (inverse rendering), and modify the latter. This approach faces two problems. First, inaccuracies in the estimation of any of those scene properties can strongly impact the final result. Second, even if they are correctly estimated, modifying the reflectance parameters to obtain a certain visual impression of the material is not a trivial task, since the exact interaction of light, shape, and material reflectance that elicits a given perception of appearance is still not well understood.

We present an image-based method for appearance editing that does not rely on any physically-based rendering of the image, but instead modifies directly the image cues that drive the perception of the material. It takes a single image of an object as input and modifies its appearance based on varying the intensity of high-level perceptual attributes (see Figure 8.1). However, since the image cues that drive the perception of such attributes can not be captured in a few image statistics [89, 295], we rely on generative

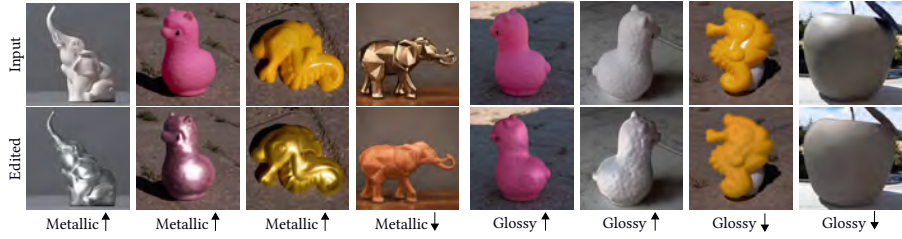


Figure 8.1: Given a single photograph as input (top row), our framework allows to edit the appearance of objects using high-level perceptual attributes. It produces realistic edits (bottom row) for a variety of input images depicting objects with different material appearance, illumination, and geometry. Note how illumination conditions are preserved in the edited results even though they were not explicitly modeled in the framework. Arrows indicate a high (pointing up) or low (pointing down) value of the target perceptual attribute.

neural networks to learn their relationship with appearance, and generate novel images with the edited material. Our networks additionally take as input a normal map that helps preserve the high-frequency details of the input geometry in the reconstructed images. Since normal maps are not available in photographs, we provide a normal map predictor that extends the applicability of our method to real input images.

A possible approach to training our framework would be to collect pairs of (original, edited) images, where the edited exemplars were manually produced given a target high-level attribute value. This is not only cumbersome, but could also lead to high variability that could hamper the learning process. Instead, and taking inspiration from existing works on face editing [174, 46, 191, 166], we train our system using perceptual judgements of the attributes of a large set of training images, that we collect through crowdsourced experiments. While these works benefit from a fixed camera location and exploit the fact that faces share similar geometry and features, we deal with a more unconstrained and varied set of potential input images. We thus devise a two-step framework, where the first step drives the change in appearance, while the second produces an image with high-frequency details.

To demonstrate the editing capabilities of our framework on a varied set of synthetic and real images, we focus on two attributes that are both common and easy to understand by participants: *Metallic* and *Glossy*. Without loss of generality, this allows us to collect robust human judgements of such attributes, while additionally assessing the perception of the appearance in our edited images through a user study. We validate our framework qualitatively, and by means of the aforementioned user study, as well as ablating each of its components. We will make our dataset of perceptual judgements publicly available to foster further research.

8.2 OUR FRAMEWORK

8.2.1 Goal and Overview

The goal of our method is to take as input an image \mathcal{I} of an object, whose material appearance we wish to edit, and a target value $b_A \in [-1, 1]$ for a high-level perceptual attribute A (e.g., *Glossy*, or *Metallic*), and from them

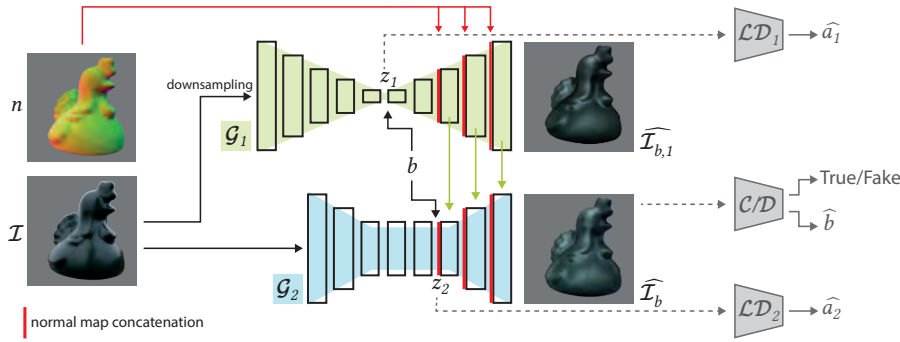


Figure 8.2: Overview of the different components of our framework. The two networks \mathcal{G}_1 and \mathcal{G}_2 both take as input the image \mathcal{I} , its normal map n , and the target attribute value b . The image first goes through \mathcal{G}_1 , whose decoder features are forwarded to the decoder of \mathcal{G}_2 (green arrows). \mathcal{G}_2 is in charge of producing the final output image $\hat{\mathcal{I}}_b$. The three auxiliary networks ($\mathcal{LD}_1, \mathcal{LD}_2$ and \mathcal{C}/\mathcal{D}), shown in gray, are used at training time to guide the networks towards correctly interpreting the target attribute value b .

produce a new image $\hat{\mathcal{I}}_b$ that exhibits the same content as \mathcal{I} , but features a change in appearance according to the desired value of the perceptual attribute, b_A (hereafter, we drop the subindex A for clarity). Our method thus needs to extract or disambiguate the information of such attribute from the input image, and allow its subsequent manipulation to generate the final one. We leverage the success of generative neural networks on image-based editing tasks, and propose a framework based on them.

Producing a representation of \mathcal{I} in which the information of the attribute has been disambiguated requires a *deep* model that can produce a compact latent code; however, such a model typically encompasses the loss of high-frequency details from the input image, hindering the reconstruction of the final image. We therefore propose a framework based on two generative networks, \mathcal{G}_1 and \mathcal{G}_2 . \mathcal{G}_1 is a deeper network that aims at producing a compact latent code of \mathcal{I} that is easy to control, and can be used to produce the final target appearance. Meanwhile, \mathcal{G}_2 is a shallower model that has the task of reconstructing the final image with high-frequency details, *guided by* the intermediate features of \mathcal{G}_1 that encode the relevant information on the final target appearance. An overview of our framework is shown in Figure 8.2, while the remainder of this section provides the details on the architecture, loss functions and training scheme used.

8.2.2 Model Architecture

Both networks, \mathcal{G}_1 and \mathcal{G}_2 , are based on an encoder-decoder architecture, in which the target attribute value b is concatenated at the bottleneck of each network (see Figure 8.2). Each encoder consists of a series of convolutional blocks that downscale the image by a factor of two, followed by a series of residual blocks. The output of these residual blocks is the latent code z_i ($i \in \{1, 2\}$), which we train to encode a representation of the input image \mathcal{I} that does not contain information about the perceptual attribute. In particular, we have six convolutional blocks for \mathcal{G}_1 , and three for \mathcal{G}_2 . Each decoder consists of a series of convolutional blocks followed by bilinear upsampling that restore the original resolution of the image. The complete description of the architecture of each network can be found in the supplementary material.

One of the main drawbacks of encoder-decoder architectures such as ours is the loss of high-frequency information when reconstructing the image from the latent code z_i . A popular strategy to recover the missing information is to use *skip connections*, that forward feature maps between the encoder and the decoder, explicitly allowing to generate high frequencies. In our case, however, this strategy cannot be applied: our latent space is trained to be invariant to the attribute, so that the decoder can reconstruct the image with the target attribute value; adding skip connections would hamper this by forwarding information from the encoder to the decoder. We alleviate this problem by providing high-frequency information to the decoder through a normal map n of the object. This normal map is concatenated to the feature maps of the decoder at different scales (illustrated in red in Figure 8.2), allowing it to incorporate high-frequency information into the reconstruction of the target image. In the case of real images, where the normal map is not directly available, it can be obtained through a normal map predictor network (see Section 8.4).

Even with the use of normal map information, a single network such as \mathcal{G}_1 can succeed in obtaining an attribute-invariant latent code z_1 , but struggles when generating a detailed reconstructed image: image $\widehat{\mathcal{I}}_{b,1}$ in Figure 8.2 has the desired appearance, but lacks fine detail. We therefore use \mathcal{G}_1 not to produce the final result, but as a means to generate a series of feature maps that encode a representation of the edited image with the target appearance. These feature maps will be used by the second network, \mathcal{G}_2 , a shallow network capable of reconstructing high-frequency details. More precisely, we use the three last feature maps from \mathcal{G}_1 , which include information at multiple scales, and concatenate them to the feature maps of \mathcal{G}_2 (as illustrated by the green vertical arrows in Figure 8.2). In this way, \mathcal{G}_2 is able to provide the output image $\widehat{\mathcal{I}}_{b,2} = \widehat{\mathcal{I}}_b$, which features the desired appearance specified by the target attribute value b while preserving the relevant high-frequency information of the input. As we will show in Section 8.4.1, the latent space of \mathcal{G}_2 alone has too much information from the input image \mathcal{I} to allow for manipulation of the desired attribute.

As explained, we need to train the latent spaces from \mathcal{G}_1 and \mathcal{G}_2 to be invariant to the attribute of interest, while learning to generate a realistic target image $\widehat{\mathcal{I}}_b$. To do this, *during training*, we use three auxiliary networks. Two latent discriminators (\mathcal{LD}_i in Figure 8.2) push the latent spaces z_i to not contain information on the attribute, while an attribute predictor and discriminator \mathcal{C}/\mathcal{D} , trained in an adversarial manner, guides the network towards generating a realistic image with the target attribute value b . The next subsection explains the training process and objectives.

8.2.3 Loss Functions and Training Scheme

IMAGE RECONSTRUCTION LOSS The first goal of each encoder-decoder network \mathcal{G}_i (for clarity, we will use \mathcal{G} instead of \mathcal{G}_i hereafter) is to reconstruct the input image \mathcal{I} when given the groundtruth perceptual attribute value a , and the normal map n . We use the L_1 loss between pixels as a measure of error, and define the reconstruction loss as:

$$\mathcal{L}_{rec}(\mathcal{G}) = \|\mathcal{I} - \mathcal{G}(\mathcal{I}, n, a)\|_1. \quad (8.1)$$

ATTRIBUTE-INVARIANT LATENT SPACE LOSS In order to force the decoder to exploit the target attribute b , we draw inspiration from Fader-

Net [174], and push the encoder to produce a latent space that does not contain information about the attribute. This is achieved with an adversarial training on the latent space, for which a latent discriminator \mathcal{LD} is introduced. The goal of \mathcal{LD} is to predict the groundtruth attribute value a from the latent code z ,

$$\mathcal{L}_{lat}(\mathcal{LD}) = \|a - \mathcal{LD}(z)\|_1, \quad (8.2)$$

while the goal of \mathcal{G} is to prevent \mathcal{LD} from being able to predict a from z :

$$\mathcal{L}_{lat}(\mathcal{G}) = -\|a - \mathcal{LD}(z)\|_1. \quad (8.3)$$

This adversarial training effectively pushes the encoder to generate an attribute-invariant latent space z , thus forcing the decoder to use the groundtruth attribute a to reach a good reconstruction.

ATTRIBUTE PREDICTOR AND DISCRIMINATOR LOSSES Until this point, the model has no feedback on its ability to edit images, since the target attribute is the groundtruth attribute value of the input image, $b = a$ (recall that the training data lacks original-edited image pairs). Therefore, in order to provide additional feedback to the model regarding the edited image, we introduce an attribute predictor \mathcal{C} . This predictor is trained to predict the attribute value of an image, using the following loss:

$$\mathcal{L}_{attr}(\mathcal{C}) = \|a - \mathcal{C}(\mathcal{I})\|_1. \quad (8.4)$$

Meanwhile, the network \mathcal{G} is trained so that the attribute value of the edited image is correctly predicted by \mathcal{C} , using:

$$\mathcal{L}_{attr}(\mathcal{G}) = \|b - \mathcal{C}(\mathcal{G}(\mathcal{I}, n, b))\|_1. \quad (8.5)$$

However, trying to satisfy the attribute predictor can lead \mathcal{G} to the generation of unrealistic artifacts in the reconstructed image. Thus, to additionally push the network to generate images that feature the same distribution as the original input data, we introduce a GAN loss together with an image discriminator \mathcal{D} . In particular, we use the losses from WGAN-GP [112] on both networks \mathcal{G} and \mathcal{D} , $\mathcal{L}_{adv}(\mathcal{G})$ and $\mathcal{L}_{adv}(\mathcal{D})$ (the complete formulation can be found in the supplementary material).

FINAL LOSS FUNCTIONS \mathcal{G}_1 is trained jointly with its latent discriminator \mathcal{LD}_1 , by using the losses $\mathcal{L}_{lat}(\mathcal{G}_1)$ and $\mathcal{L}_{rec}(\mathcal{G}_1)$. We do not include the attribute predictor and discriminator module because \mathcal{G}_1 is intended to create a compact and editable latent space, rather than a high-quality output image. The resulting loss functions are:

$$\mathcal{L}(\mathcal{G}_1) = \lambda_{rec}^{\mathcal{G}} \mathcal{L}_{rec}(\mathcal{G}_1) + \lambda_{lat}^{\mathcal{G}} \mathcal{L}_{lat}(\mathcal{G}_1), \quad (8.6)$$

$$\mathcal{L}(\mathcal{LD}_1) = \lambda_{lat}^{\mathcal{LD}} \mathcal{L}_{lat}(\mathcal{LD}_1). \quad (8.7)$$

\mathcal{G}_2 is trained jointly with its latent discriminator \mathcal{LD}_2 , as well as the attribute predictor and discriminator module \mathcal{C}/\mathcal{D} . The resulting loss functions are:

$$\mathcal{L}(\mathcal{G}_2) = \lambda_{rec}^{\mathcal{G}} \mathcal{L}_{rec}(\mathcal{G}_2) + \lambda_{lat}^{\mathcal{G}} \mathcal{L}_{lat}(\mathcal{G}_2) + \lambda_{adv}^{\mathcal{G}} \mathcal{L}_{adv}(\mathcal{G}_2) + \lambda_{attr}^{\mathcal{G}} \mathcal{L}_{attr}(\mathcal{G}_2) \quad (8.8)$$

$$\mathcal{L}(\mathcal{LD}_2) = \lambda_{lat}^{\mathcal{LD}} \mathcal{L}_{lat}(\mathcal{LD}_2) \quad (8.9)$$

$$\mathcal{L}(\mathcal{C}/\mathcal{D}) = \lambda_{adv}^{\mathcal{D}} \mathcal{L}_{adv}(\mathcal{D}) + \lambda_{attr}^{\mathcal{C}} \mathcal{L}_{attr}(\mathcal{C}). \quad (8.10)$$

In practice, \mathcal{C} and \mathcal{D} share the same convolutions and are trained as a unique network, thus the joint loss in Equation 8.10.

TRAINING DETAILS We optimize all losses using the Adam optimizer [163] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To train the generators, we use a learning rate of 10^{-4} . \mathcal{G}_1 is trained with the following loss weights: $\lambda_{rec}^{\mathcal{G}} = 1$, $\lambda_{lat}^{\mathcal{G}} = 5$, while \mathcal{G}_2 is trained with $\lambda_{rec}^{\mathcal{G}} = 1$, $\lambda_{lat}^{\mathcal{G}} = 2.5$, $\lambda_{adv}^{\mathcal{G}} = 0.02$ and $\lambda_{attr}^{\mathcal{G}} = 2$. Both latent discriminators are optimized with a learning rate of $2.5 \cdot 10^{-5}$ for 12 iterations for every iteration on the generator. \mathcal{C}/\mathcal{D} is optimized with a learning rate of 10^{-4} for seven iterations for every iteration on the generator with loss weights $\lambda_{adv}^{\mathcal{D}} = 1$ and $\lambda_{attr}^{\mathcal{C}} = 3$. Our model is trained individually for each attribute. We first train \mathcal{G}_1 for 300 epochs, then train \mathcal{G}_2 for 50 epochs, freezing parameters for \mathcal{G}_1 . We implemented our models using the Pytorch framework [244] and trained them using a Nvidia 2080Ti GPU. In total, training our framework took two days per attribute.

8.3 TRAINING DATASET

Training our model to edit a certain attribute of material appearance requires images with realistic depictions of materials, on objects with different shapes and a variety of illuminations. For each of these images, we require the corresponding value for the attribute of interest. Since we are targeting high-level perceptual attributes of material appearance, this value needs to be obtained from subjective data gathered through subject responses. These image-attribute (\mathcal{I}, a) pairs are used to train our network towards correctly interpreting such attributes.

IMAGE DATA We leverage the recent dataset by Lagunas et al. [171], designed specifically for learning tasks related to material appearance. It is composed of realistic renderings of 13 geometries of varied complexity (with two additional viewpoints, leading to 15 different scenes), illuminated with six captured environment maps [53]. The objects are rendered with 100 measured BRDFs from the MERL dataset [211], using the physically-based renderer Mitsuba [139]. The dataset comprises a total of 9,000 renderings, of which representative samples are shown in Figure 8.3.

SUBJECTIVE ATTRIBUTES The image dataset we use [171] includes associated subjective data, but in the form of similarity judgements between pairs of images, unsuitable for our goal. Other datasets include subjective measures of high-level perceptual attributes of material appearance for the materials in the MERL dataset, but for a single shape and illumination [276]. Since shape and illumination play an important role in the perception of material appearance [172, 320, 227], we set out to gather our own subjective data of high-level perceptual attributes for the Lagunas et al. image dataset.

To do so, we follow the same methodology as Serrano et al. [276]: we carry out a perceptual experiment in which, for each image in the Lagunas et al. dataset, participants had to rate a number of high-level attributes on a Likert-type, 1-to-5 scale. To further increase the robustness of the obtained ratings, we augment Lagunas et al.’s dataset by creating, for each combination of material \times shape \times illumination, five different images with slight variations in the viewpoint (randomly sampled within a 45 degrees cone around the original viewpoint). Examples of such images for the *bunny* shape are shown in Figure 8.4. Similar to previous large-scale studies, we relied on Amazon Mechanical Turk to collect the ratings

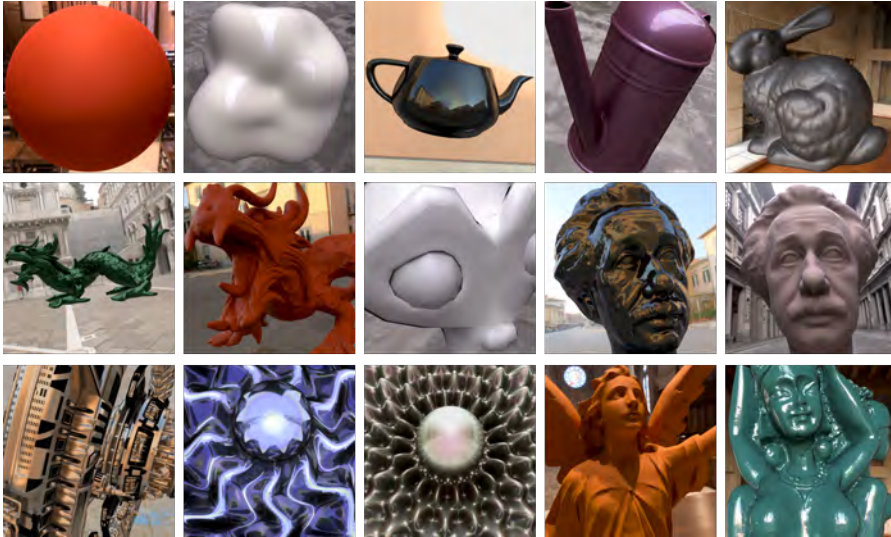


Figure 8.3: Representative samples of the image dataset used for training. The images show each of the 15 scenes in the dataset (13 distinct geometries, two of them with two different viewpoints, for a total of 15 scenes), featuring different materials and illuminations.

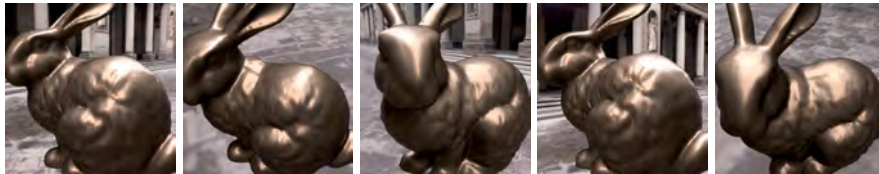


Figure 8.4: Example of the five viewpoints used in the perceptual study on the *bunny* shape rendered with the *Uffizi* illumination and *alum-bronze* material.

TRAINING OF PARTICIPANTS Participants of our perceptual study first had to go through a training session in which they were shown with a text description and a few example images depicting materials with low and high score values for each attribute. We then show them the same screen as in the study and ask them to answer the attributes for two easy examples (shown in Figure 8.6, left). If answers of the participants were not the expected ones, we instructed them to look again at the image and check the description of the attributes.

CONTROL QUESTIONS In addition to the 15 stimuli, we added four control images in order to detect lazy users. These images contains materials with clear expected answers (shown in Figure 8.6 right). We rejected participants answering wrongly to more than one of these questions and rejected 20% of the participants based on this criteria.

A total of 2,600 paid subjects participated in the study, each of them seeing 15 different random images. Figure 8.5 shows a screenshot of the perceptual study, as seen by the participants. The stimuli is shown on the left part of the screen while the list of attributes to score are shown on the right. Through our perceptual study we gather, for each attribute, 39,000 ratings ($13 \text{ shapes} \times 6 \text{ illuminations} \times 100 \text{ materials} \times 5 \text{ viewpoints}$), leading to that number of image-attribute pairs. It is important to note that, due to the vast size of our dataset, we only gather one response per condition (per combination of material \times shape \times illumination \times viewpoint), which can lead to variability

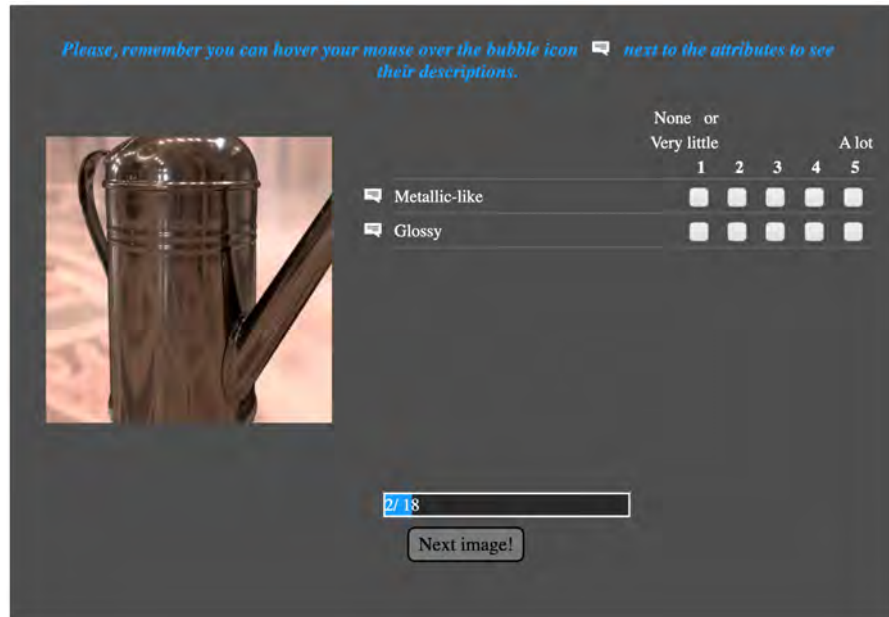


Figure 8.5: Screenshot of the perceptual study as seen by participants. Stimuli is shown on the left, the participant have to select a score for the two attributes shown on the right.

in the data that may hinder the convergence of the training. In order to reduce it, we pool the perceptual ratings over viewpoint and shape by means of the median, more robust to outliers than the mean.

8.4 RESULTS AND EVALUATION

In this section, we start by introducing our evaluation dataset, and showing results of our framework by applying it to two perceptual attributes: *Glossy* and *Metallic*. We then validate our design choices through a series of ablation studies (Section 8.4.1), and analyze the consistency of our editing across controlled geometry, illumination, and material variations (Section 8.4.2). Finally, we perform an additional user study to assess whether our edits of the attributes do correlate with human perception (Section 8.4.3).

EVALUATION DATA Our evaluation data is composed of both synthetic images and real photographs. The *synthetic images* evaluation dataset is composed of images never seen during training by our framework. They are rendered using eight shapes collected from free online sources, four illuminations obtained from *HDRIHaven* [121], and eight materials coming from Dupuy and Jakob’s database [70]. A representative subset is shown in Figure 8.7.

We collected *real images* for our evaluation dataset by browsing online catalogues of decorative items, as well as photographing objects ourselves in uncontrolled setups. Within each image, we masked the object of interest using an online API [147]. Since our framework requires a normal map, which is not directly available when using real photographs, we obtain the normal maps for these objects by using a *normal map predictor*. Inspired by image-to-image generative networks, we trained a new model to infer normal maps directly from the single-view RGB images. Our normal map



Figure 8.6: Left: the two images used in our training session. Right: the four images used as controls.



Figure 8.7: Representative images of our synthetic evaluation dataset, showing the eight shapes and materials used in it. Each column is rendered with one of the four illuminations used.

predictor consists of a modified Pix2Pix network[138]. We carefully designed our architecture and losses to minimize convolution artifacts, high variance noise in the resulting normals, and maintain as much geometrical detail from the original images as possible, while reducing the influence of varying reflectance and illumination conditions. The model was trained on synthetic data coupled with groundtruth normal maps. Additional details about the architecture and losses used to train the normal predictor can be found in the supplementary material. Representative examples of our real evaluation dataset, together with their predicted normal maps, can be seen in Figure 8.8.

RESULTS Figure 8.1 shows editing results for a variety of real-world objects photographed in uncontrolled setups under different conditions, for our two attributes *Glossy* and *Metallic*. They include indoor and outdoor scenarios, varied shape complexity, and different types of materials, yet our framework can handle them gracefully, producing compelling edits by just changing the high-level perceptual attribute. It is interesting to observe how, even though



Figure 8.8: Representative examples of our real images evaluation dataset, comprised of photos from online catalogues (top), and casually photographed objects (bottom). For each image, we also show its normal map, as obtained by our normal map predictor.

the illumination is not explicitly modeled during training, the edits seem to plausibly capture the lighting in the scene. Additionally, our framework is trained so that the attribute of interest can be sampled along its range, producing consistent results. This is shown in Figure 8.9 for two real images, where both attributes exhibit a coherent variation (see the supplementary material for additional results). Figures 8.1 and 8.9 also show that our normal map predictor is capable of yielding a normal map that allows for realistic editing of photographs.

8.4.1 Ablation Studies

We evaluate the utility of each of the components of our method through a series of ablation studies where the *Metallic* attribute is used. We generate five ablated versions of our framework, for which we show an illustrative result in Figure 8.10. First, the effect of the individual generative networks is shown in *Only \mathcal{G}_1* and *Only \mathcal{G}_2* . When using only \mathcal{G}_1 , the resulting image features the desired edit, but lacks high-frequency details. Meanwhile, \mathcal{G}_2 alone is able to reconstruct the fine detail of the input image, but cannot convincingly edit the appearance towards the target increased metallicity. We then investigate the effect of the auxiliary networks. When the latent discriminator \mathcal{LD}_2 is removed (*W/o \mathcal{LD}_2*), the generated image struggles to convey the appearance required by the target edit. Additionally, without the attribute predictor and discriminator (*W/o C/D*), the framework is only slightly able to improve the edited result from the first network \mathcal{G}_1 . Finally, we investigate the effect of the normal map information by removing them from the training (*W/o normals*). Without this information, the framework cannot reconstruct the geometry, leading to unrealistic results.

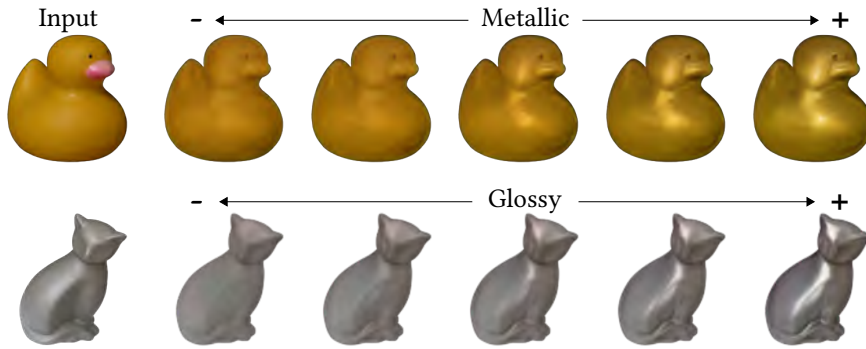


Figure 8.9: Editing results by varying the perceptual attributes *Metallic* and *Glossy*. First column is the input image, following ones show the edited image when sampling the attribute as $[-1, 0, 0.5, 0.75, 1]$ for *Metallic* and $[-1, -0.25, 0, 0.25, 1]$ for *Glossy*. Our method produces a realistic editing of the input over the whole range.

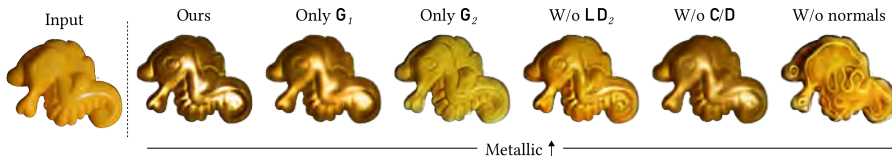


Figure 8.10: Ablation studies where we trained and tested out each of the individual components of our framework. The leftmost image shows the input photograph, followed by the target attribute (*Metallic* +1). Then, from left to right: the resulting edited image using our method, only the \mathcal{G}_1 network, only the \mathcal{G}_2 network, training without the *latent discriminator* \mathcal{LD}_2 and its associated loss function, training without the *attribute predictor and discriminator* \mathcal{C}/\mathcal{D} and its associated loss function, and training without using the normal map information of the input image. Our method qualitatively yields superior performance and allows for the creation of sharp highlights and realistic images.

8.4.2 Consistency of the Edits

We use our synthetic evaluation dataset to assess the consistency of our edits under different conditions. Figure 8.11 (a) shows edits performed when both material and geometry are the same in the input image, and only the illumination changes. Our material edits are perceptually consistent, while illumination properties are preserved within the edits. Figure 8.11 (b) shows results when only the geometry changes in the input images. Our edits yield consistent results across geometries, appearing to be all made of a similar material (within each row). Last, in Figure 8.11 (c) we evaluate the consistency of our edits using two different materials with similar reflectance properties, namely *acrylic-felt-orange* and *acrylic-felt-green*. Again our framework yields consistent, plausible results for both attributes.

8.4.3 User Study

We run an additional user study to assess the perception of the appearance in our edited images. In the study, participants were asked to rate the perceptual attribute in generated images in which such attribute had been edited with our framework. The layout of the user study is the same as the

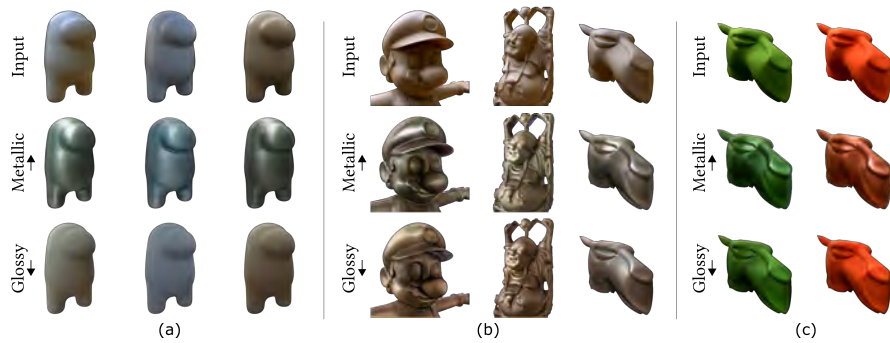


Figure 8.11: Example results illustrating the consistency of our editing framework. (a) Same object and material, but different illuminations in the input image; (b) Same illumination and material, but different geometry; (c) same geometry and illumination, but different materials with similar reflectance properties. Our framework is capable of producing compelling and consistent edits in all cases. Arrows pointing up correspond to a target attribute value of +1, while arrows pointing down correspond to a value of -1.

one used in the perceptual study (Figure 8.5) except that participants were asked to rate one attribute at a time.

STIMULI We selected three images for each attribute (*Glossy* and *Metallic*), varied in shape, illumination and material, and edited them with our method by setting the target attribute value to -1, 0 and +1. This led to two sets (one per attribute) of nine *edited images*. We also incorporated, for each attribute, nine other images from the training dataset, chosen such that they covered the whole range of attribute values; we will term them *training images*. Note that these images are unedited, and for each we have the “groundtruth” attribute value gathered through our perceptual study that was used to train our framework (Section 8.3). In Figure 8.12, we show the stimuli from the “edited images” set that we used in the validation user study. For each attribute, the top part shows the input images (synthetic) that we selected, covering different shapes, illuminations and reflectance properties. The bottom part shows the three edited images that we show in the study for each input (low attribute value, middle value and high attribute value), resulting in nine stimuli.

PROCEDURE The stimuli were shown to participants in two separate blocks, one per attribute. Each block thus consists of 18 images, for which the participants had to rate the attribute on a Likert-type 1-to-5 scale. 15 participants took part in the study, leading to 15 ratings for each image and attribute.

RESULTS For each image, we average the participants’ ratings to obtain a perceived attribute value (to which we will refer here as *collected value*). Table 8.1 shows the results of the Pearson correlation between the collected and the expected attribute values. Note that the expected attribute value is the target attribute value for the edited images, and the “groundtruth” attribute value for the training images. For both, edited and training images, there is a strong (and significant) correlation between the collected and the expected attribute values. While for the *Metallic* attribute, the correlations for edited images are on par with the ones for training images (0.90 and 0.92 respectively), correlations for the *Glossy* attribute are lower for the edited

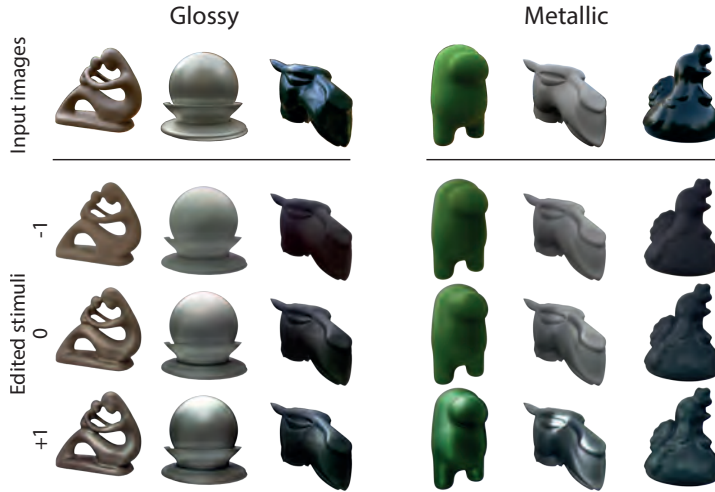


Figure 8.12: Input images and edited stimuli used in our user-study. Top: input images to our framework. Bottom: The edited images with three target attributes, leading to nine stimuli for each attribute.

Images	<i>Metallic</i>	<i>Glossy</i>
Edited	0.90, $p < 0.001$	0.86, $p = 0.003$
Training	0.92, $p < 0.001$	0.96, $p < 0.001$

Table 8.1: Pearson correlation coefficients (along with their p-value) between the expected attribute of the images shown in the user study, and the answers of the participants (collected attribute).

images than for the training images (0.86 and 0.96 respectively). This can be due to the fact that our edited images do not cover the full range of glossiness, with the most glossy images (with a target attribute set to +1) being scored between 3 and 3.7 (on a scale of 1 to 5). However, the correlations for the edited images remain high, showing that our edited images are globally well perceived.

In Figure 8.13, we show the answers that we collected for both attribute *Metallic* and *Glossy* and for the two sets of images. The blue dots show all the 15 ratings that we collected for each images, where the density of the color indicates the number of answer, while the red crosses indicates the average answer for each stimuli. While the answers for both sets of images appear to be strongly correlated, the answers collected on our edited images do not reach the full scale of the attribute, with a maximum score of 3.7 for the *Glossy* attribute, and 4 for the *Metallic* attribute. The average variances in the answers was higher for edited images than for training ones (0.42 and 0.62 respectively for *Glossy*, 0.5 and 0.84 respectively for *Metallic*).

8.5 DISCUSSION AND LIMITATIONS

We have presented an image-based framework to edit materials through the manipulation of high-level perceptual attributes. Our framework is based

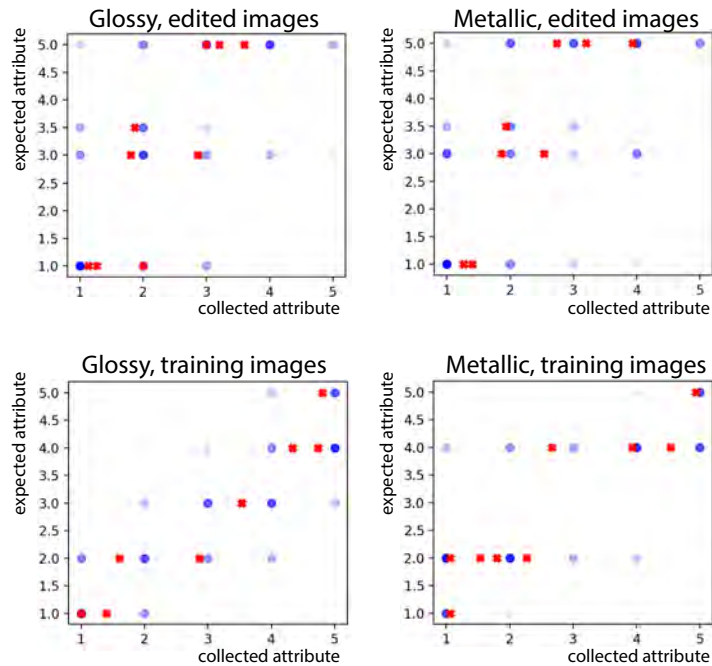


Figure 8.13: Answers collected in our validation study for both attribute Metallic and Glossy and for the two sets of images. The blue dots show all the 15 ratings that we collected for each images, where the density of the color indicates the number of answer, while the red crosses indicates the average answer for each stimuli.

on two generative networks aiming at providing an editable latent space, and reconstructing high-frequency details, respectively. We have shown the editing capabilities of our method on a variety of input images, both synthetic and real, and validated the results through a user study.

Our framework is not free of limitations, which open up several possibilities for future work. Since no normal maps are provided for real pictures, we have introduced a normal map predictor; inaccuracies in its output may lead to distortions in the edited objects, especially visible around highlights, as shown in Figure 8.14 (a); our framework would thus benefit from better models to infer normals. Besides, since our architecture does not allow for the use of skip-connections, high-frequency illumination details such as mirror-like reflections may also not be recovered properly when trying to reach high glossiness values, as shown in Figure 8.14 (b). Similarly, our framework can only create fuzzy highlights when presented with an input image depicting a diffuse material that conveys only limited information about the illumination. It would be interesting to combine our approach with recent neural rendering techniques which can create such information about the illumination [306, 187].

Our framework was trained using the dataset by Lagunas et al. [171] which contains synthetic data using the isotropic BRDFs from MERL [211]. However, MERL materials are biased in terms of albedo and reflectance. To mitigate this, we have augmented our input data with changes in hue before feeding it to our framework (see the Appendix E.1). Nevertheless, designing a dataset beyond isotropic BRDFs could allow the framework to edit a wider range of appearances. Moreover, since our dataset contains

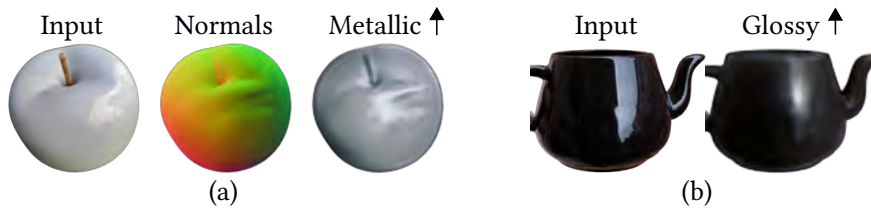


Figure 8.14: Limitations of our framework: (a) Noise in the prediction of the normals may lead to unpredicted editing results (left: input image, center: inferred normal map, right: edited image with *Metallic* +1); (b) Due to the lack of skip-connections, almost mirror-like reflections in the input image (left) are hard to model during editing when trying to reach high glossiness (right: edited image with *Glossy* +1).

single-color objects, we currently cannot edit spatially-varying reflectance (such as the duck’s beak in Figure 8.9).

We hope that our work inspires additional research and novel perceptually-based applications. We will make our data and code available for further experimentation, in order to facilitate the exploration of these possibilities.

Part VI

CONCLUSION

In this thesis we have presented contributions in three core aspects of visual appearance: how to measure appearance similarity by taking into account the subjective nature of human perception, what is the effect of confounding factors in material appearance and how they affect our perception, and, last, how to develop intuitive applications for human relighting and material editing.

MEASURING APPEARANCE SIMILARITY In this first part we have discussed two different lines of work. In the first one (Chapter 3), we have focused on the long-standing problem of measuring material similarity using single images and in such a way that it agrees with human judgements. We have presented a new dataset with carefully chosen stimuli, which are generated using physically-based rendering. From there, we have relied on the dataset to launch a set of crowdsourced experiments to collect human judgements on material appearance similarity. Then, we leveraged both the dataset and human judgements to propose a deep learning-based framework with a custom loss function capable to accurately measure material appearance similarity from images. We have validated our method, showing how it outperforms previous work, and also have proposed several applications that are enabled by our metric. However, many different potential avenues for future work remain open. For instance, we collected perceptual data by sampling a single geometry and illumination in our user studies; one could potentially extend the experiments to account for such factors, which in turn should yield a more robust metric. Besides, we rely on a synthetic dataset and, while we have carefully sampled the different variables (geometry, material, and illumination), collecting and using a dataset directly from controlled photographs would provide statistics to the neural network that better match the real world.

In the second part (Chapter 4), we have focused on measuring similarity for iconography, a non-photorealistic domain where the style or identity given by the artists are crucial to convey icon's information and message. We have relied on a dataset automatically collected from online databases where each icon is additionally paired with semantic information. We have used such semantic information and labels associated with each icon image to train a deep learning-based similarity measure, with a siamese neural network. However, the semantic information of each icon is manually added by the artists and, thus, it may be noisy. While this have allowed us to obtain plausible subjective comparisons using our metric, we had to additionally collect human judgements to validate our model. Last, we have shown several applications using our metric such as database visualizations, icon retrievals, or icon sets proposals. By means of future work, one direct approach would be to additionally curate the semantic data obtained from the dataset which would allow for better convergence of the model. In addition, we implicitly learn the separation between style and identity in the deep neural network. One could try to learn such separation explicitly, by modeling it directly through the deep learning model, using additional perceptual data, or by assuming that icons identity is directly linked with shape and using traditional computer vision techniques to disambiguate between shape (visual identity) and style.

CONFOUNDING FACTORS IN MATERIAL PERCEPTION In this part, we have focused on how our perception of materials changes when varying the physical parameters that govern the interaction between light and matter. First (Chapter 5), we have presented a comprehensive study on the joint effects of geometry and illumination on our performance for material recognition tasks. We have used a synthetic dataset to launch a set of crowdsourced experiments where users had to recognize materials given a reference and a set of candidate samples. From there, we have observed that users better recognize the visual appearance of materials when those are represented in a geometry and an illumination covering a wide frequency spectrum, and we have also observed significant first-order interactions between geometry and illumination. We additionally validated our findings by connecting them to the rendering equation through its Fourier transform. Last, we have proposed an analysis of simple statistics and complex models, such as neural networks, where we find that the latter may share similar high-level factors to humans when recognizing materials. This study represents an additional step towards understanding how such confounding factors affect human perception. As such, many potential lines of future work open up. We have done a preliminary study on human perception and deep neural networks. Given the current trends and potential exciting outcomes of such models, a more in-depth analysis explaining how deep networks work, and their relationship with human perception, remains to be done. For instance, one could continue analyzing unsupervised training schemes [295], or studying other network architectures such as transformers [321]. Besides, here we have focused on geometry and illumination for the particular case of material recognition. Observing how those affect other aspects of perception such as visual appearance similarity, or how those influence our perception on a set of material attributes is an exciting potential avenue of future work.

In the second half (Chapter 6), we have presented another user study where we analyzed the effect of motion in our perception of a set of attributes describing different aspects of material appearance. First, we have rendered a set of stimuli using different degrees of motion. Then, we have used those realistic stimuli to launch a crowdsourced experiment where users had to rate the material attributes. From there, we found that particular attributes defining how the material reflects light (e.g., *glossiness*, or *sharpness of reflections*) yield lower values as the degree of motion increases. Besides, we have launched a second user study where we analyzed the impact of the brightness attribute under different degrees of motion. From the results we have built *brightness maps* that characterized the impact of motion in our perception of the brightness attribute. By means of future work, we could extend our study to additional geometries and illuminations. In addition, other types of motion could also be explored.

INTUITIVE APPLICATIONS FOR APPEARANCE EDITING The last part has dealt with developing intuitive applications whose main goal was the editing of visual appearance. We have started by introducing a framework for full-body human relighting working just on single RGB images (Chapter 7). We have generated a synthetic dataset containing more than 500 different combinations of posed humans where we explicitly modeled the visual appearance of materials with an Oren-Nayar microfacet model for the diffuse and a microfacet model with GGX distribution for the specular. We have leveraged precomputed radiance transfer and spherical harmonics to introduce our image reconstruction formulation, where we have additionally introduced

a residual term aiming to overcome problems in the reconstruction. With these, we train a deep neural network capable to relight images given just a photograph with a human in it, and the target illumination to perform the relighting. We showed how our model numerically outperforms previous approaches, and showcased relighted results in various scenarios. However, spherical harmonics basis are known for struggling in the reconstruction of high-frequency information. Exploring other bases could be an interesting line of future work. Besides, we modeled the material appearance directly in our data and implicitly learned their properties using the neural network. One could learn the computational model of material appearance explicitly in the neural network, to obtain better performance.

Finally, we have also presented a framework for intuitive editing of material appearance just from RGB images (Chapter 8). We have first created a dataset consisting of a diverse set of stimuli with different materials, geometries, illuminations, and viewpoints. Using this dataset, we have launched several crowdsourced experiments where we aimed at collecting information on several high-level attributes describing material appearance. We relied on the images in the dataset, the ratings on the material attributes, and on generative neural networks to propose an intuitive framework for material editing, where the user just needs an input RGB image together with a value representing the desired change in appearance. We have demonstrated the applicability of our method with an additional user study. Last, we have also shown how our framework works on a diverse set of inputs, including also real photographs. However, our framework relies on a *latent discriminator* to remove all the information from the attribute in the generative part of the network; this way, since all information about the attribute was removed, we are forcing the user to provide a new attribute value thus allowing for the editing. Unfortunately, this method does not allow to use the traditional skip connections in the architecture. This, in turn, removes high-frequency information from the final edited image. To alleviate this problem we additionally relied on a normal map estimation module. One potential avenue of future work could be to explore more complex architectures: Instead of having a single *latent discriminator* for the whole architecture, the use of smaller, individual *latent discriminators* for each layer could be a potential solution to adding skip connections. Moreover, this has two benefits: First, it would allow us to remove the normal estimation module (and therefore one potential source of bias in the reconstruction); and second, it would allow to better reconstruct high-frequency information in the final edited image.

PERSONAL CONCLUSIONS I would also like to add a personal note here. I started this thesis just after graduating with a Master in applied mathematics. My expertise was, mostly, in the fields of machine learning and deep learning. During the thesis, I have been working on many different projects that span also different fields. This has allowed me to learn new concepts and ideas from other domains such as perception, physically-based rendering, and online rendering; and also expand my knowledge regarding deep learning methods. Broadening my technical expertise has been both, a challenging but at the same time rewarding experience.

Throughout this thesis, I have not only broadened my technical expertise. Some projects involved international collaborations, sometimes I had to present our work, or I had to give an invited talk explaining some of the projects contained in this PhD thesis. This has also allowed me to develop a

range of soft skills regarding how to approach a technical conversation with people with other expertise, how to collaborate in different environments, how to work when there is a nine hour gap between collaborators, how to better manage my time, or on how to give a talk. I believe these soft skills are also key in the future that lies ahead, and sometimes it is something hard to learn unless you see yourself in a situation that allows for it. Luckily, during the thesis, there have been plenty of these moments where I had to challenge myself, go out of the comfort zone, and learn something new, not necessarily technical.

I also have to mention the students I have supervised, either because they were doing their final degree project or because they were interns in our group. Supervising people is a task that made me be "on the other side". Now, I could better understand my supervisors since I saw my own mistakes reflected on the students I was supervising. This is also an enlightening experience that adds extra value and weight to what my supervisors have tried to explain and taught to me during these years. I definitely learned a lot from my supervisors and from each of the students. I tried to understand the way they were thinking, and try to put myself in their situation. I tried my best to teach them everything I knew, and also tried my best to teach them not to make the same mistakes I could be doing when I was in their place.

Also, I want to add a personal note about the research internships. I believe this has also been one of the most rewarding experiences of the thesis. You travel abroad, you see a new culture, have a new temporary supervisor, and have to adapt to the rhythm and pace of an industry which could differ from what happens back at home. My two research internships have also taught me a lot of technical and soft skills. There you have to talk with people from many different backgrounds, you have a new supervisor with different workflows than what you are used to, and you have a different environment. This teaches you to be flexible, adapt, and to be reactive. You are in a new situation and you need to learn fast in order to collaborate and support your new colleagues.

We say here, in Spain, that if you work in what you like, you never have to work again. I think this is a good summary of the thesis. I was working on projects I like, learning new technical skills, developing soft skills, expanding my knowledge, and also expanding the network of people I know. I want to say thanks again to my supervisors, Belén and Diego, for giving me this huge opportunity. This years have been a challenging experience, but the more challenging it is, the more rewarding it becomes once it is done.

Part VII
APPENDIX

A.1 ADDITIONAL LOSS TERMS

We describe here the two additional loss terms that we evaluate in our ablation study (refer to Part III, Section 3.5 for details).

A.1.1 Cross-entropy Term \mathcal{L}_{CE}

This term accounts for the soft-label cross entropy [302]. It aims at learning a soft classification task by penalizing samples which do not belong to the same class. In our case, each material represented in the dataset can constitute a class, and the set of classes in the dataset is \mathcal{K} . Given an image r included in a training batch \mathcal{B} , the probability of r belonging to a certain class $k \in \mathcal{K}$ is given by $p_k(r)$. The cross-entropy loss term is given by:

$$\mathcal{L}_{CE} = \frac{1}{|\mathcal{B}|} \sum_{r \in \mathcal{B}} s(r) \tag{A.1}$$

$$s(r) = - \sum_{k \in \mathcal{K}} [(1 - \epsilon) \log p_k(r) l_k(r) + \epsilon \log p_k(r) u(k)] \tag{A.2}$$

where $l(r)$ is the one-hot encoding of the groundtruth label, and $l_k(r)$ is the value of the vector for label k (note that our training image data can be labeled, since it comes from the materials dataset presented in Section 3.2). The value of ϵ is set to 0.1, and we use the uniform distribution $u(k) = \frac{1}{|\mathcal{K}|}$. Both ϵ and $u(k)$ work as regularization parameters so that a wrong prediction does not penalize the cost function aggressively, while preventing overfitting.

A.1.2 Batch-mining Triplet Loss Term \mathcal{L}_{BTL}

In learned models for classification or recognition, a batch-mining triplet loss has been proposed in combination with a soft-label cross entropy term such as the one we use to improve the model’s generalization capabilities and accuracy [93]. It is modeled as:

$$\mathcal{L}_{BTL} = \frac{1}{|\mathcal{B}|} \sum_{r \in \mathcal{B}} [\operatorname{argmax}_{x_i^+} (\|f(r) - f(x_i^+)\|_2^2) - \operatorname{argmin}_{x_i^-} (\|f(r) - f(x_i^-)\|_2^2) + \mu]_+ \tag{A.3}$$

where x_i^+ designates images of the training batch \mathcal{B} belonging to the same class as r , and x_i^- images belonging to a different class than r . Intuitively, this loss mines and takes into consideration the hardest examples within each batch, improving the learning process.

A.2 QUERIES AND AGREEMENT WITH HUMANS

We show queries to our method and agreement with humans’ majority response in Figures A.1 and A.2. For each reference material (left) we show

two candidate materials. The number below each candidate indicates the number of human votes it received. Numbers in green indicate that the candidate judged by our model as closer to a given reference agrees with humans majority response, while numbers in red represent cases where our model does not agree with humans' majority response. As reported in Section 3.5 our model agrees with humans around 80% of the time.

A.3 MATERIAL SUGGESTION EXAMPLES

In Figures A.3 and A.4 we show additional material suggestions. Queries (left) and results for the closest materials in the Extended MERL dataset [276].



Figure A.1: Queries to our method and agreement with humans' majority response (I).



Figure A.2: Queries to our method and agreement with humans' majority response (II).

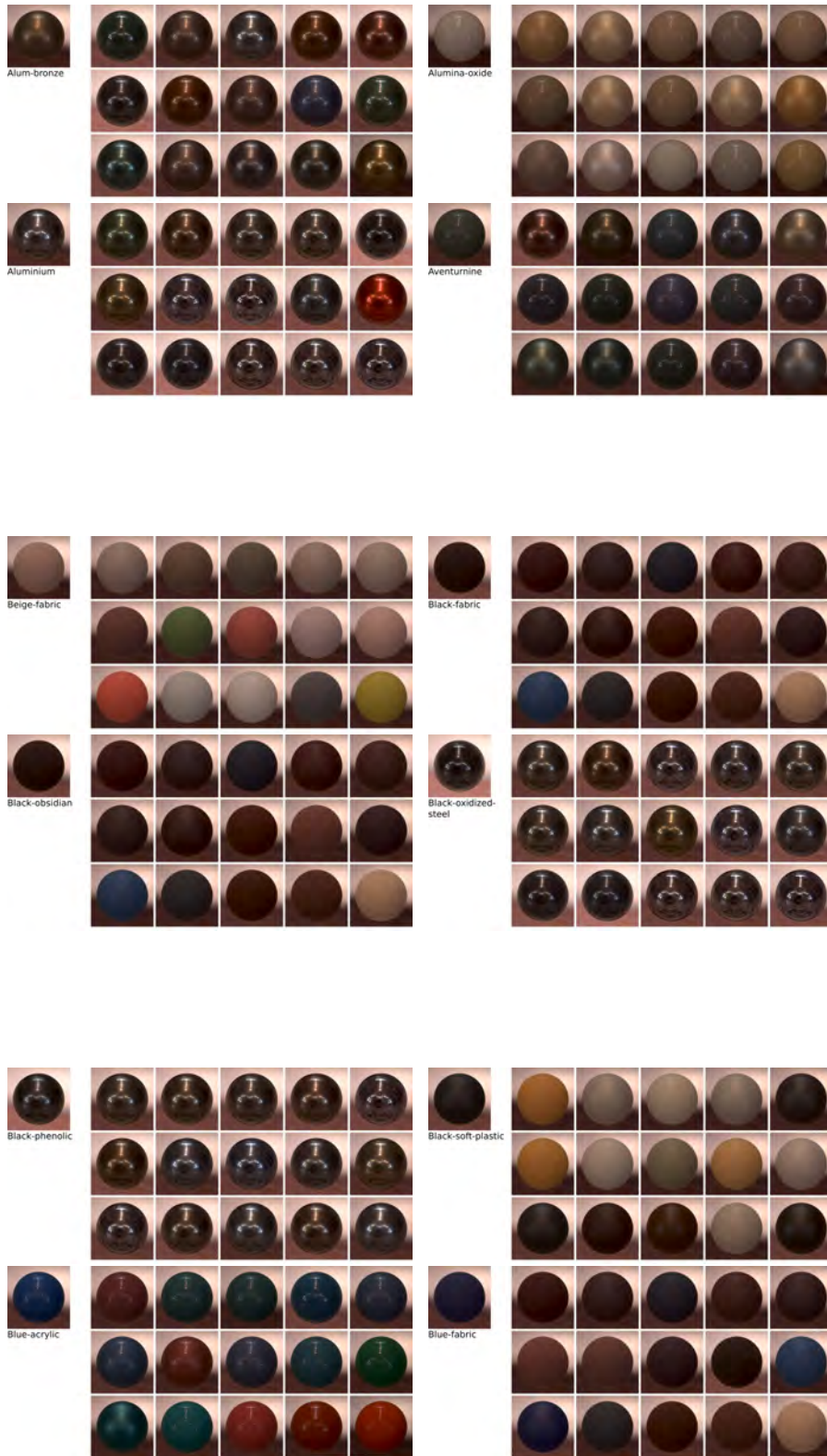


Figure A.3: Queries using our measure for materials in the Extended MERL dataset (I).



Figure A.4: Queries using our measure for materials in the Extended MERL dataset (II).

B.1 ADDITIONAL DETAILS ON IMAGE STATISTICS

To measure the correlation between image statistics and users' performance we employ a Pearson \mathcal{P} and Spearman \mathcal{S} correlation test with a significance value (p -value) of 0.05. The value \mathcal{P}^n represents the Pearson correlation for the n^{th} statistical moment (same applies for the Spearman \mathcal{S}^n correlation). Please, also refer to Part IV, Section 5.3.

LUMINANCE We analyze if the moments of the luminance of each material image have a direct influence on users' performance. We found that the moments of the luminance are not correlated with users' performance: $\mathcal{P}^1 = -0.14$ ($p = 0.17$), $\mathcal{S}^1 = -0.15$ ($p = 0.15$), $\mathcal{P}^2 = 0.02$ ($p = 0.83$), $\mathcal{S}^2 = -0.03$ ($p = 0.78$), $\mathcal{P}^3 = 0.03$ ($p = 0.77$), $\mathcal{S}^3 = 0.03$ ($p = 0.78$), $\mathcal{P}^4 = 0.01$ ($p = 0.94$), $\mathcal{S}^4 = 0.05$ ($p = 0.65$).

RGB IMAGE We analyze if the moments of the joint RGB intensity of each material image have a direct influence on users' performance. We found that the moments of the joint RGB intensity have little to no correlation with users' performance except for the standard deviation: $\mathcal{P}^1 = -0.02$ ($p = 0.79$), $\mathcal{S}^1 = -0.06$ ($p = 0.51$), $\mathcal{P}^2 = 0.43$ ($p < 0.001$), $\mathcal{S}^2 = 0.50$ ($p < 0.001$), $\mathcal{P}^3 = 0.16$ ($p = 0.09$), $\mathcal{S}^3 = 0.22$ ($p = 0.02$), $\mathcal{P}^4 = -0.1$ ($p = 0.30$), $\mathcal{S}^4 = -0.06$ ($p = 0.52$).

We also tested out the correlation for each channel and found out that for all the channels there is no correlation for any of the first 4 statistical moments.

RED CHANNEL On the red channel there seems to be a slight positive linear correlation between the fourth moment (kurtosis) and users' performance. All the other statistics show no significant correlation: $\mathcal{P}^1 = -0.10$ ($p = 0.29$), $\mathcal{S}^1 = -0.08$ ($p = 0.42$), $\mathcal{P}^2 = 0.03$ ($p = 0.60$), $\mathcal{S}^2 = -0.02$ ($p = 0.87$), $\mathcal{P}^3 = 0.07$ ($p = 0.46$), $\mathcal{S}^3 = 0.07$ ($p = 0.51$), $\mathcal{P}^4 = 0.20$ ($p = 0.04$), $\mathcal{S}^4 = 0.15$ ($p = 0.13$).

GREEN CHANNEL There is no correlation between any statistics on the green channel: $\mathcal{P}^1 = -0.04$ ($p = 0.66$), $\mathcal{S}^1 = -0.0$ ($p = 0.74$), $\mathcal{P}^2 = 0.03$ ($p = 0.55$), $\mathcal{S}^2 = 0.04$ ($p = 0.67$), $\mathcal{P}^3 = 0.05$ ($p = 0.64$), $\mathcal{S}^3 = 0.06$ ($p = 0.53$), $\mathcal{P}^4 = 0.05$ ($p = 0.63$), $\mathcal{S}^4 = 0.01$ ($p = 0.94$).

BLUE CHANNEL Similar to the green channel, the blue does not show any correlation for the first 4 statistical moments: $\mathcal{P}^1 = 0.03$ ($p = 0.72$), $\mathcal{S}^1 = -0.004$ ($p = 0.93$), $\mathcal{P}^2 = 0.06$ ($p = 0.52$), $\mathcal{S}^2 = 0.01$ ($p = 0.95$), $\mathcal{P}^3 = 0.13$ ($p = 0.19$), $\mathcal{S}^3 = 0.10$ ($p = 0.30$), $\mathcal{P}^4 = 0.16$ ($p = 0.11$), $\mathcal{S}^4 = -0.05$ ($p = 0.61$).

MOTION IN MATERIAL PERCEPTION: STIMULI AND ADDITIONAL RESULTS

C.1 STIMULI USED IN THE FIRST EXPERIMENT

We choose a set of 72 different realistic materials, that span six different material categories, and three degrees of motion: 0, 45, and 95 with a significant difference between them in order to provide a notable change in the appearance of the stimuli. Here, we provide all the rendering stimuli in the first experiment, for different motion degrees and materials; including glass materials (see Figure C.1), metallic materials (see Figure C.2), paint materials (see Figure C.3), plastic materials (see Figure C.4), rubber materials (see Figure C.5), and stone materials (see Figure C.6).

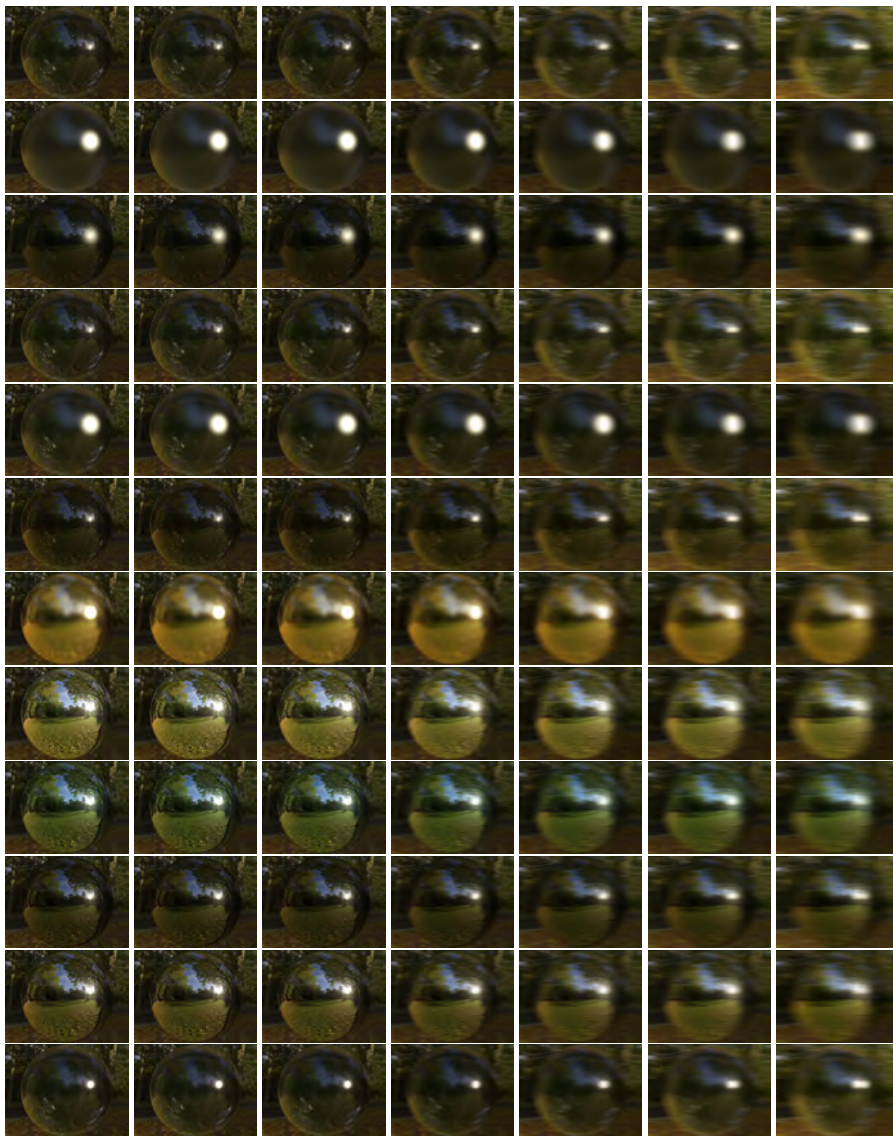


Figure C.1: Stimuli rendered with a glass material.

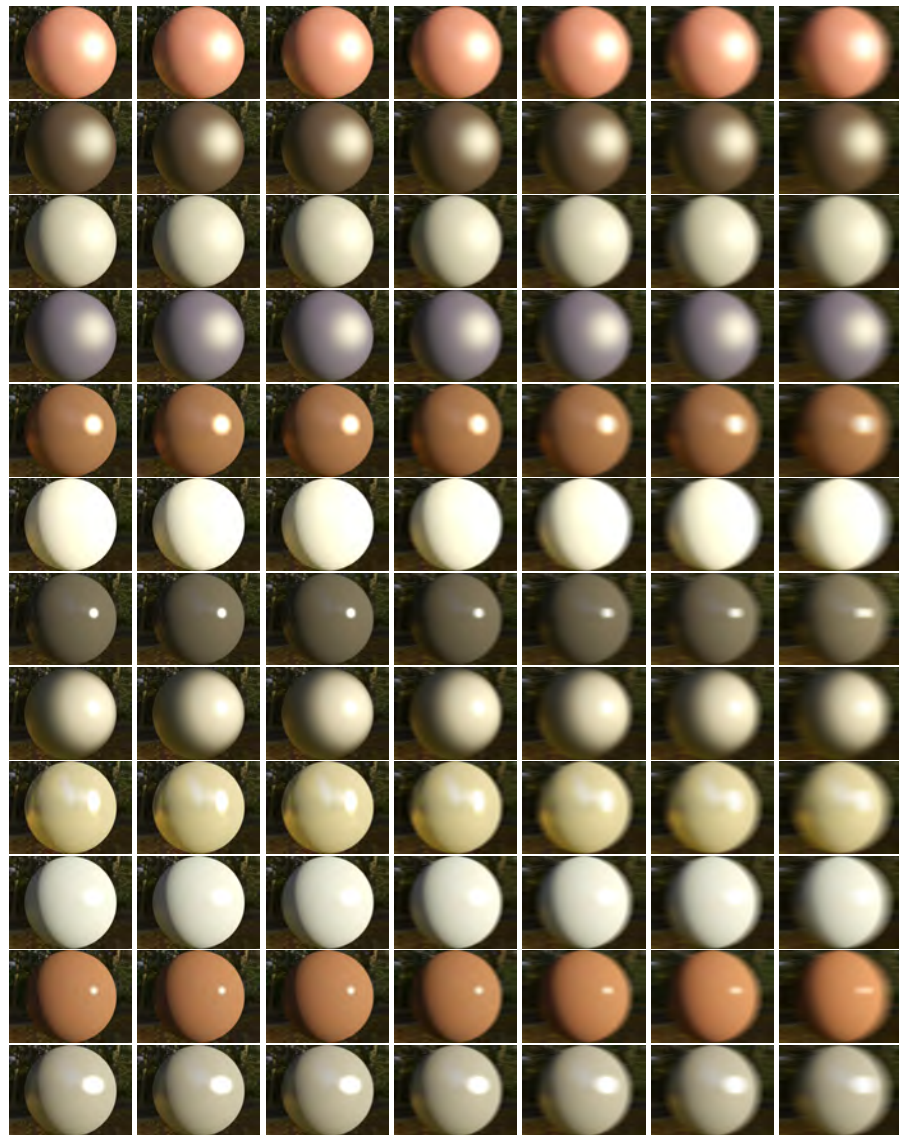


Figure C.2: Stimuli rendered with a metallic material.

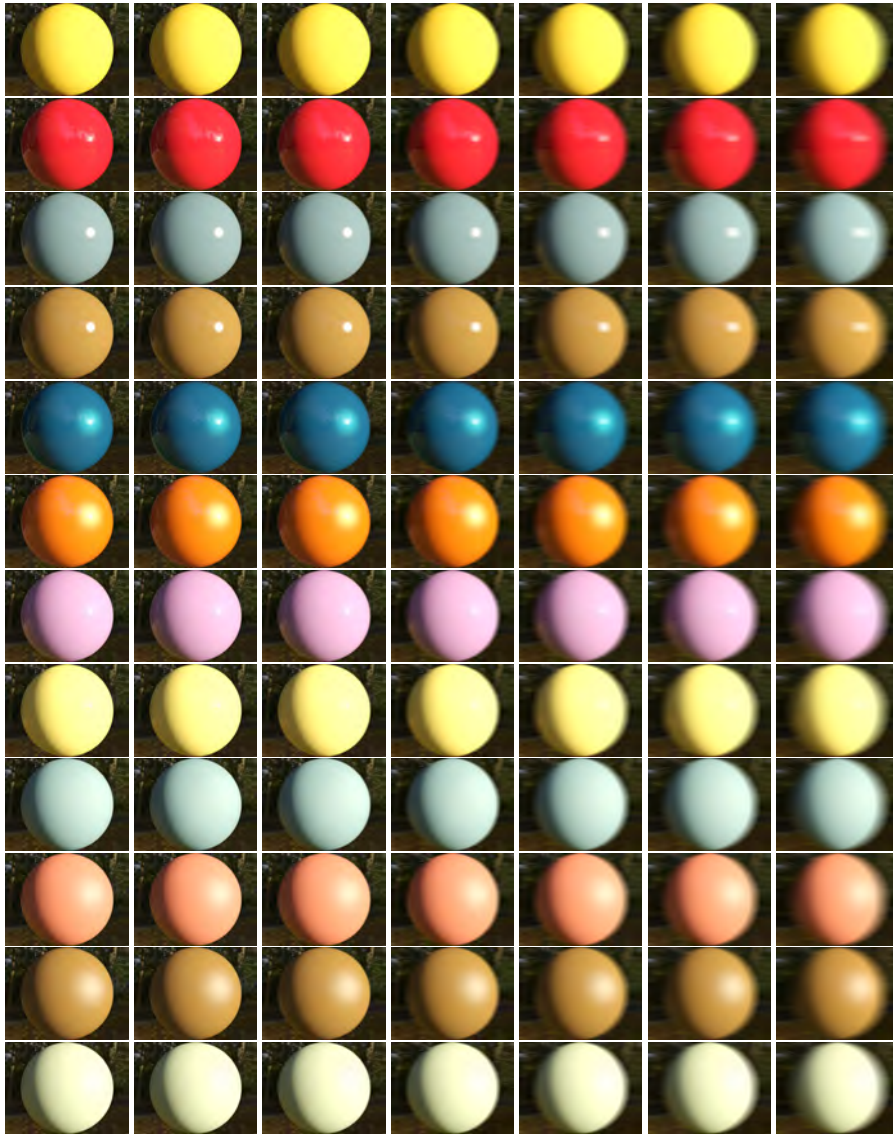


Figure C.3: Stimuli rendered with a paint material



Figure C.4: Stimuli rendered with a plastic material.

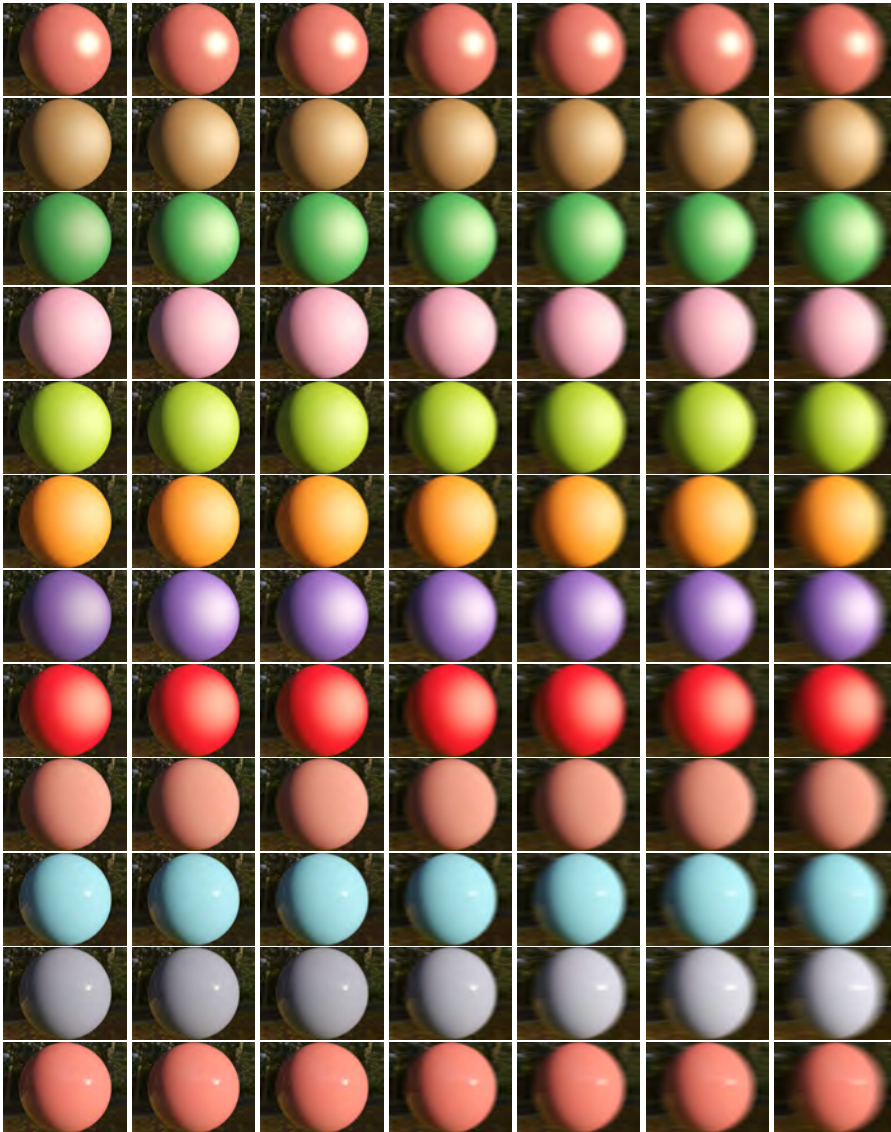


Figure C.5: Stimuli rendered with a rubber material.

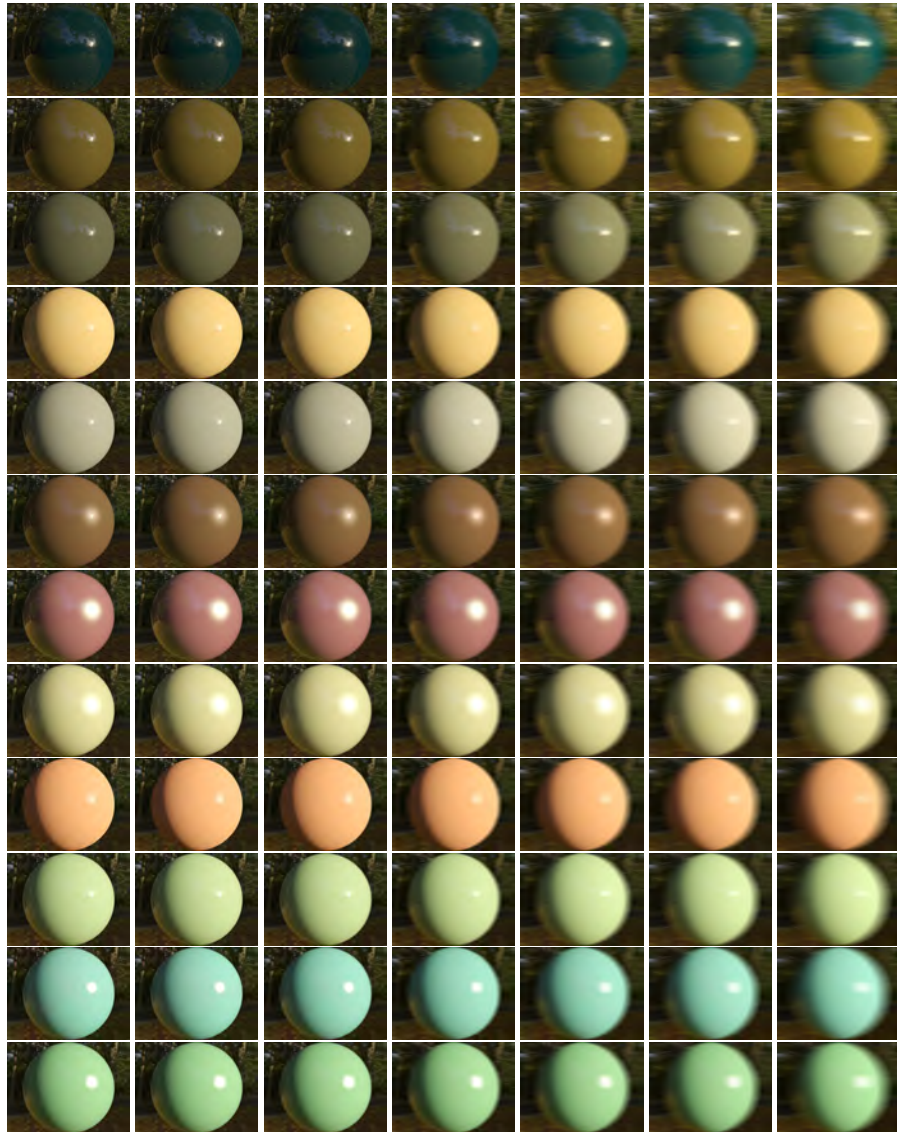


Figure C.6: Stimuli rendered with a stone material.

C.2 STIMULI USED IN THE SECOND EXPERIMENT

Here, we provide all the rendering stimuli, for the different motion degrees used for the second experiment, and rendered with a plastic (see Figure C.7) and rubber (see Figure C.8) material.

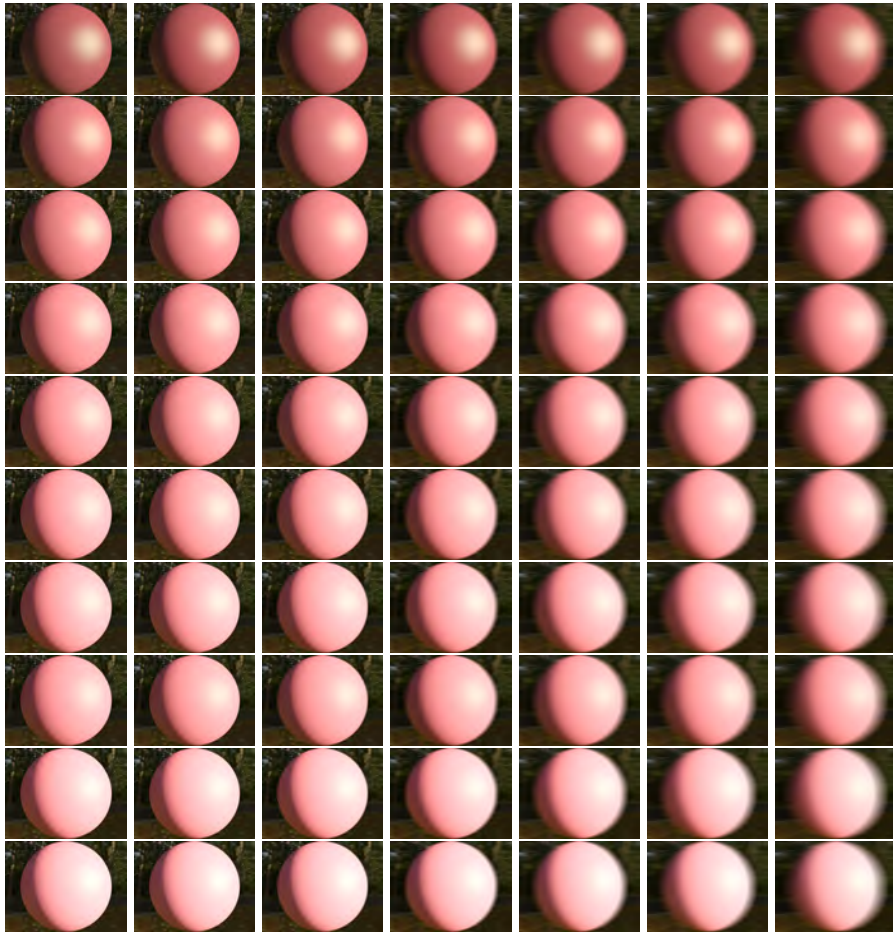


Figure C.7: Stimuli rendered with a plastic material.

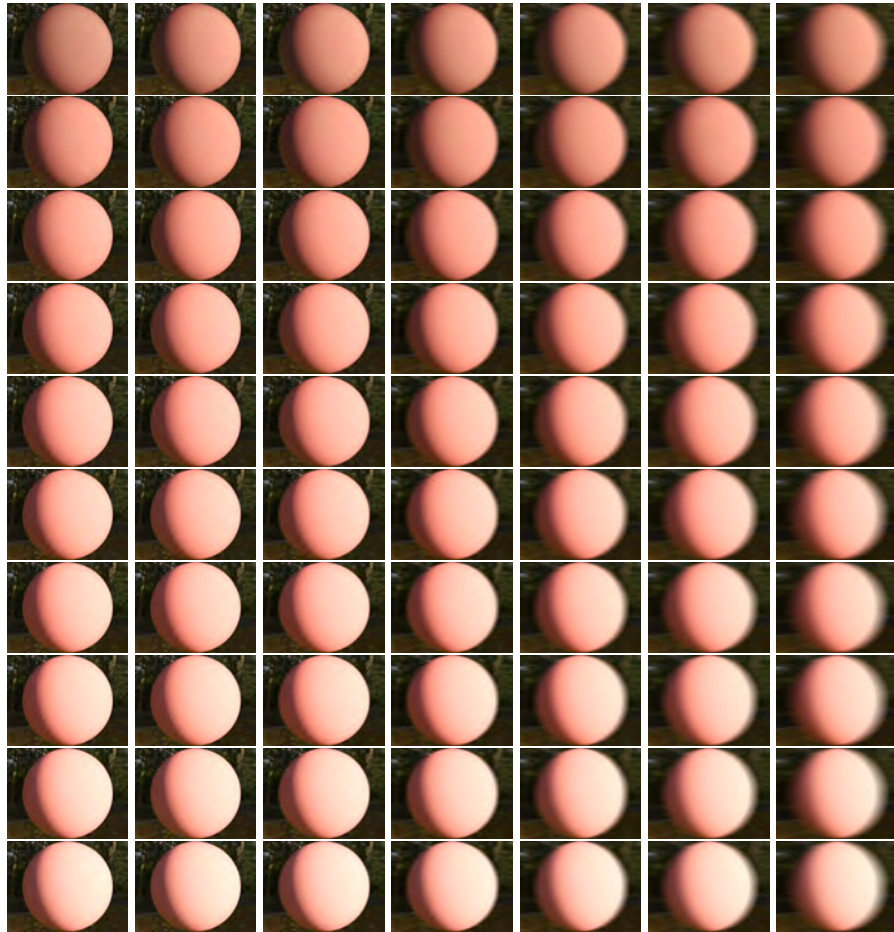


Figure C.8: Stimuli rendered with a rubber material.

C.3 ATTRIBUTE PLOTS FOR THE FIRST EXPERIMENT

Figures C.10 and C.9 show the trends of participants' ratings for each of the six perceptual properties and motion degrees. All the 14 material attributes are included here.

C.4 TABLES FOR THE STATISTICAL TESTS

In Table C.1 we show a summary with all the p -values for each rated attribute and material category can be found here. The last 3 columns are the results of Nemenyi post hoc test.

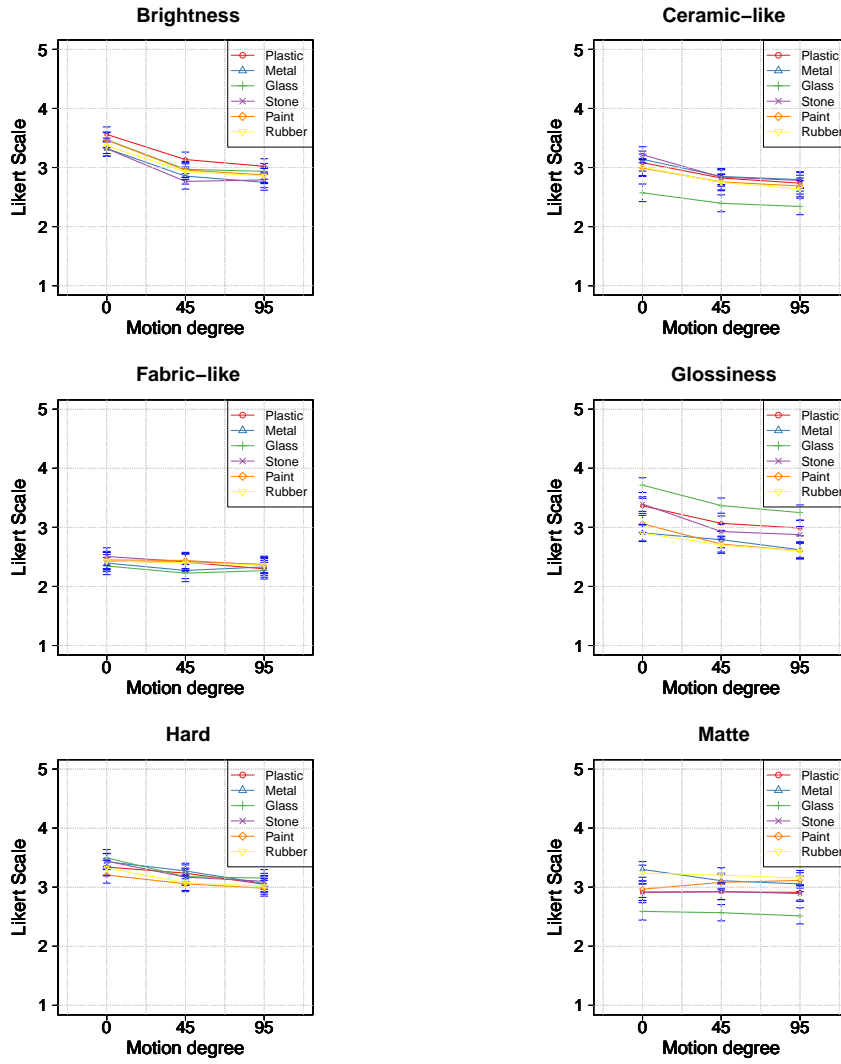


Figure C.9: Participants' ratings for each attribute (I).

Attribute	Material	df	χ^2	p-value	Motion Degree 1 - Motion Degree 2		
					0-45	0-95	45-95
Brightness	Glass	2	62.026906	0	0	0	0.987653
Brightness	Metal	2	52.210877	0	1.9e-05	0	0.648557
Brightness	Paint	2	71.273543	0	0	0	0.644612
Brightness	Plastic	2	70.490219	0	2e-06	0	0.380282
Brightness	Rubber	2	59.049369	0	5e-06	0	0.782593
Brightness	Stone	2	72.239824	0	0	0	0.94575
Ceramic-like	Glass	2	9.140673	0.010354	0.475144	0.101555	0.651724
Ceramic-like	Metal	2	21.307143	2.4e-05	0.05292	0.001302	0.459785
Ceramic-like	Paint	2	18.231847	0.00011	0.071882	0.006437	0.66809
Ceramic-like	Plastic	2	18.100125	0.000117	0.056734	0.005347	0.693741
Ceramic-like	Rubber	2	21.055416	2.7e-05	0.091045	0.002296	0.417871
Ceramic-like	Stone	2	37.925481	0	0.000496	4.6e-05	0.840614
Fabric-like	Glass	2	2.886894	0.236112	0.568434	0.710349	0.972435
Fabric-like	Metal	2	2.739726	0.254142	0.494135	0.838124	0.838124
Fabric-like	Paint	2	0.904832	0.63609	0.993151	0.812427	0.870146

MOTION IN MATERIAL PERCEPTION: STIMULI AND ADDITIONAL RESULTS

Fabric-like	Plastic	2	3.536547	0.170627	0.916398	0.670046	0.423228
Fabric-like	Rubber	2	0.28752	0.866096	0.999811	0.946764	0.952696
Fabric-like	Stone	2	4.293578	0.116859	0.570232	0.368992	0.939384
Glossiness	Glass	2	52.165517	0	0.00038	0	0.20944
Glossiness	Metal	2	19.079951	7.2e-05	0.268817	0.002579	0.178387
Glossiness	Paint	2	29.109813	0	0.017141	0.000136	0.373076
Glossiness	Plastic	2	24.957811	4e-06	0.004147	8e-04	0.892242
Glossiness	Rubber	2	19.648216	5.4e-05	0.07028	0.00339	0.552995
Glossiness	Stone	2	64.044543	0	2e-06	0	0.665083
Hard	Glass	2	26.414634	2e-06	0.003813	0.003813	1
Hard	Metal	2	23.213582	9e-06	0.235903	0.000322	0.063757
Hard	Paint	2	6.218859	0.044626	0.515371	0.133948	0.691377
Hard	Plastic	2	12.37123	0.002059	0.795545	0.021732	0.110017
Hard	Rubber	2	14.349603	0.000766	0.051119	0.017717	0.919921
Hard	Stone	2	17.769406	0.000138	0.023907	0.006472	0.902958
Matte	Glass	2	0.832642	0.659469	0.976786	0.902439	0.799138
Matte	Metal	2	13.941793	0.000939	0.103199	0.012899	0.707589
Matte	Paint	2	10.033413	0.006626	0.278732	0.045366	0.66809
Matte	Plastic	2	0.301337	0.860133	0.908633	0.987406	0.961934
Matte	Rubber	2	2.897714	0.234839	0.912467	0.417871	0.670312
Matte	Stone	2	1.23991	0.537969	0.700157	0.995186	0.756951
Metallic-like	Glass	2	17.558333	0.000154	0.08209	0.011406	0.744577
Metallic-like	Metal	2	6.260749	0.043701	0.505754	0.178387	0.786393
Metallic-like	Paint	2	7.007215	0.030089	0.526955	0.161849	0.737057
Metallic-like	Plastic	2	12.434659	0.001995	0.202993	0.038501	0.740177
Metallic-like	Rubber	2	6.317841	0.042472	0.308387	0.299321	0.999811
Metallic-like	Stone	2	18.854111	8.1e-05	0.039767	0.008738	0.859626
Plastic-like	Glass	2	6.132203	0.046602	0.463817	0.281493	0.939053
Plastic-like	Metal	2	21.538275	2.1e-05	0.126544	0.001131	0.243873
Plastic-like	Paint	2	14.7925	0.000614	0.144638	0.013738	0.621018
Plastic-like	Plastic	2	10.920943	0.004252	0.175771	0.062276	0.883633
Plastic-like	Rubber	2	16.758801	0.00023	0.386356	0.008506	0.225067
Plastic-like	Stone	2	10.328733	0.005717	0.711709	0.048253	0.257891
Roughness	Glass	2	2.233487	0.327344	0.687084	0.995159	0.627958
Roughness	Metal	2	1.769022	0.412916	0.984223	0.636611	0.742081
Roughness	Paint	2	1.327128	0.515013	0.822556	0.977167	0.702924
Roughness	Plastic	2	2.923754	0.231801	0.502798	0.622119	0.980393
Roughness	Rubber	2	0.002567	0.998717	0.999243	0.999811	0.999811
Roughness	Stone	2	1.347339	0.509834	0.962876	0.85023	0.700157
Rubber-like	Glass	2	1.832787	0.399959	0.766827	0.687084	0.990533
Rubber-like	Metal	2	3.742087	0.153963	0.564794	0.304457	0.893111
Rubber-like	Paint	2	3.308789	0.191208	0.362965	0.621018	0.903979
Rubber-like	Plastic	2	0.160305	0.922975	0.998219	0.956428	0.971889
Rubber-like	Rubber	2	6.883663	0.032006	0.290414	0.158344	0.940514
Rubber-like	Stone	2	2.132231	0.344343	0.500008	0.868791	0.81054
Sharpness	Glass	2	101.799789	0	0	0	0.592216
Sharpness	Metal	2	51.283286	0	0.003144	2e-06	0.191732
Sharpness	Paint	2	38.978552	0	0.001787	7.1e-05	0.702924
Sharpness	Plastic	2	45.868047	0	0.000238	8e-06	0.728714
Sharpness	Rubber	2	34.242595	0	0.005591	0.000341	0.727691
Sharpness	Stone	2	48.995227	0	2.7e-05	4e-06	0.918417
Soft	Glass	2	2.915209	0.232793	0.639853	0.463817	0.957268
Soft	Metal	2	1.564286	0.457425	0.917082	0.600677	0.838124
Soft	Paint	2	1.993103	0.36915	0.609202	1	0.609202

C.4 TABLES FOR THE STATISTICAL TESTS

Soft	Plastic	2	0.832524	0.659507	0.923864	0.944387	0.762746
Soft	Rubber	2	0.212865	0.899036	0.987957	0.977353	0.933952
Soft	Stone	2	5.212337	0.073817	0.218616	0.932702	0.389632
Strength	Glass	2	66.824201	0	6e-06	0	0.356824
Strength	Metal	2	42.993498	0	0.013656	4e-06	0.107576
Strength	Paint	2	18.884768	7.9e-05	0.029072	0.011595	0.946334
Strength	Plastic	2	41.848293	0	0.003415	5e-06	0.274439
Strength	Rubber	2	14.235911	0.00081	0.056087	0.040289	0.990766
Strength	Stone	2	29.444317	0	0.003038	0.000215	0.767967
Tint	Glass	2	33.292627	0	0.021045	2.2e-05	0.162949
Tint	Metal	2	27.795756	1e-06	0.069825	0.000348	0.228108
Tint	Paint	2	17.913259	0.000129	0.022423	0.012988	0.981092
Tint	Plastic	2	32.998658	0	0.000925	0.000691	0.996836
Tint	Rubber	2	22.810596	1.1e-05	0.027021	0.00318	0.771872
Tint	Stone	2	41.545562	0	0.000199	1.6e-05	0.840614

Table C.1: All p -values for each attribute and material category. The last 3 columns are the results of the Nemenyi post hoc test.

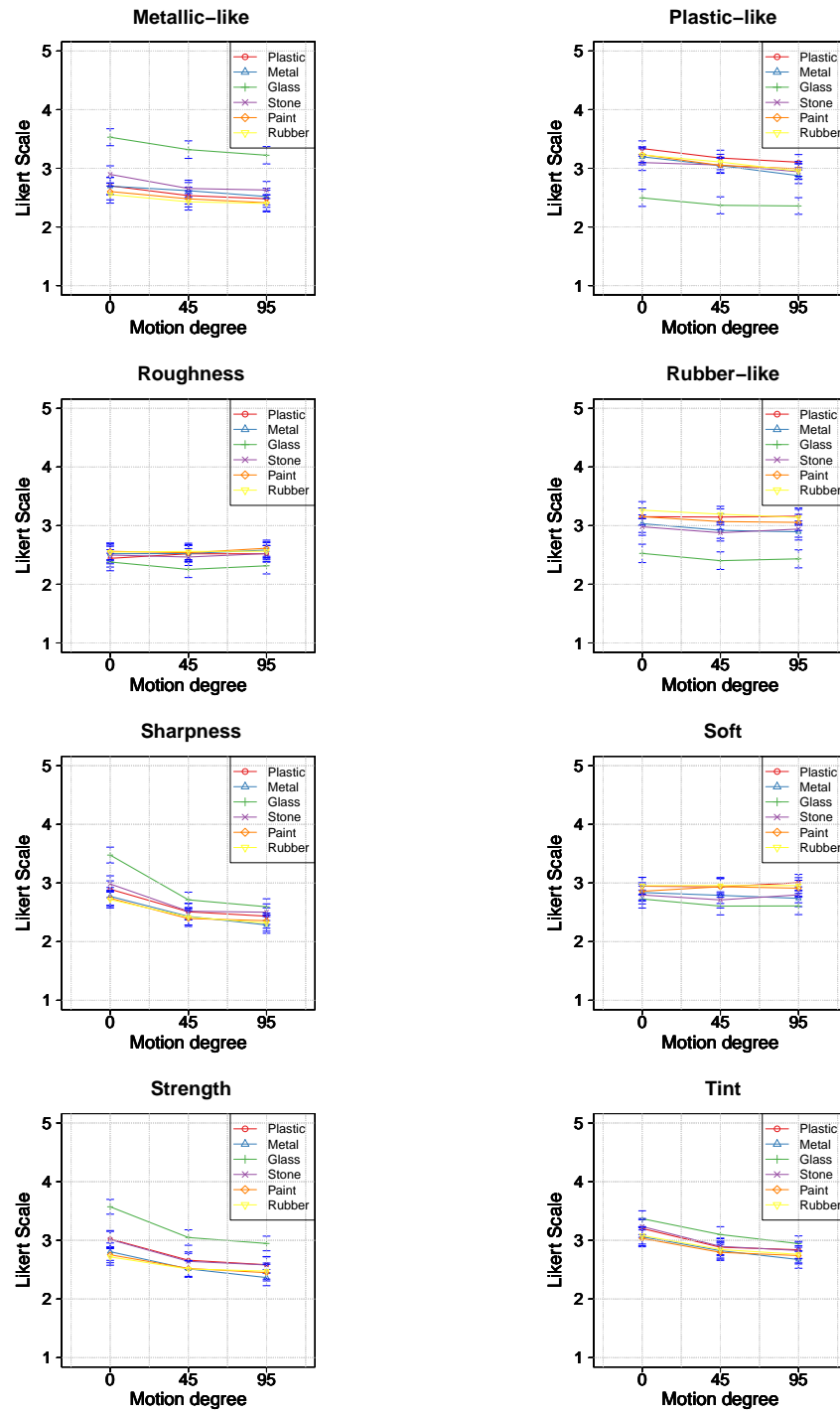


Figure C.10: Participants' ratings for each attribute (II).

FULL-BODY HUMAN RELIGHTING: ADDITIONAL RESULTS

D.1 ADDITIONAL RESULTS

In this section we show additional results for the real photographs in the test dataset. Figure [D.1](#) show a variety of relighting results under different illuminations. For each input photo and illumination map we show the final relighted image, and the reconstructed shading and residual terms. We observe how our model is capable of faithfully relighting the input photographs.



Figure D.1: Relighted results using real photographs for two different illuminations (*ennis*, and *pisa*) and five different input images. In each case, we show the relighted image, and the reconstructed shading and residual terms scaled for visualization purposes. Last column shows the results under two rotations of the same illumination.

E.1 ADDITIONAL DETAILS ON THE FRAMEWORK

Our framework is composed of two encoder-decoder networks \mathcal{G}_1 and \mathcal{G}_2 , the auxiliary latent discriminator networks $\mathcal{LD}_1, \mathcal{LD}_2$ and the auxiliary attribute predictor and discriminator \mathcal{C}/\mathcal{D} only used by means of a loss function during training.

GENERATIVE NETWORK Both generative networks \mathcal{G}_1 and \mathcal{G}_2 are composed of an encoder made of a series of convolutional blocks that reduce the spatial dimensions of the input by a factor of two, a set of residual blocks that transform the bottleneck features, and a decoder made of a series of convolutional blocks followed by bilinear upsampling layers. The target perceptual attribute is spatially replicated to match the size of the latent code and concatenated to it at the beginning of the decoder.

Let C_k denote a 4×4 Convolution layer with k filters and stride 2, then followed by a Rectified Linear Unit (ReLU), R_k denotes a residual block that contains two 3×3 convolution with k filters. D_k denotes a convolutional block (3×3 convolution with k filters - leaky Rectified Linear Unit [336]) followed by a bilinear upsampling layer. Reflection padding is used in all convolutions.

\mathcal{G}_1 takes input images at the resolutions 128×128 and contains six layers both in the encoder and decoder and two residual blocks, resulting in the following architecture:

Encoder: C32-C64-C128-C256-C512-C512-
Bottleneck: -R512-R512-
Decoder: -(b)D512-D256-(n)D128-(n)D64-(n)D32-(n)D8

where (b) indicates the concatenation of the target attribute and (n) indicates the concatenation of the normal map.

\mathcal{G}_2 takes as input images at the resolution 256×256 and contains four layers in the encoder, three in the decoder and three residual blocks, resulting in the following architecture:

Encoder: C32s1k7-C64-C128-C256-
Bottleneck: -R256-R256-R256-
Decoder: -(b)(n)D128-(n)D64-(n)D8

where C32s1k7 indicates a 7×7 Convolution-ReLU layer with 32 filters and stride 1. This first convolution allows us to reduce the number of spatial resolution of the image while keeping the same receptive field.

Each network ends with a last convolutional block with stride 1 and 8 filters followed by a single convolutional layer with three output filters (corresponding to the RGB channels) and a hyperbolic tangent function (\tanh) to bring the values into the range $[-1, +1]$.

LATENT DISCRIMINATOR The latent discriminators, \mathcal{LD}_1 and \mathcal{LD}_2 take the features in the bottleneck of \mathcal{G}_1 and \mathcal{G}_2 , respectively, and use them to predict the attribute a of the input image. The architecture of the latent discriminators \mathcal{LD}_1 is as follows:

\mathcal{LD}_1 : Cd512-FC256-FC1

\mathcal{LD}_2 : Cd512-Cd512-Cd512-Cd512-pool-FC256-FC1

where Cdk represent a convolutional block (4×4 convolution, leakyReLU, and dropout with probability 0.3), Fck refers to a fully connected layer with k features, and pool represent an average pooling operation. At the end, the output of the latent discriminators goes through a *tanh* layer that outputs the attribute prediction \hat{a} in the range $[-1, +1]$.

ATTRIBUTE PREDICTOR AND DISCRIMINATOR The attribute predictor and discriminator \mathcal{C}/\mathcal{D} take the image as input and outputs an attribute prediction \hat{b} . The image goes first through an encoder. The features from such encoder then go to the discriminator, and the attribute predictor. The architecture is as follows:

Encoder: C32-C64-C128-C256-C512-

Discriminator: -C1

Attribute predictor: -pool-FC256-FC1

WGAN-GP LOSS FORMULATION Generative Adversarial Networks (GANs) are complex to train. This is partially due to the instability of the loss function proposed in the original formulation [109]. WGAN-GP [112] aims to alleviate such problems by introducing a new loss function that relies on the Wasserstein distance between distributions and a gradient penalty term \mathcal{L}_{GP} .

Intuitively, the discriminator is trained to give a high score to real images and a low score to generated ones, aiming at disambiguate them:

$$\mathcal{L}_{adv}(\mathcal{D}) = -\|\mathcal{D}(\mathcal{I})\|_2 + \|\mathcal{D}(\mathcal{G}(\mathcal{I}, n, b))\|_2 + \mathcal{L}_{GP} \quad (\text{E.1})$$

while the generator is trained such that the the discriminator believe that generated images are actually real (giving them a high score):

$$\mathcal{L}_{adv}(\mathcal{G}) = -\|\mathcal{D}(\mathcal{G}(\mathcal{I}, n, b))\|_2 \quad (\text{E.2})$$

We refer the reader to the original manuscript for additional information [112].

DATA AUGMENTATION To have a more diverse set of input images and help the model generalize better, we perform a set of random data augmentation routines. First, input images are scaled to have size 512×512 px and we perform random flips, 90-degree rotations, and a random crop with size 480×480 px. Then, to account for the bias in the BRDFs from the training dataset, we perform random changes in the saturation and the hue. Finally, the image is scaled to 256×256 and fed to the networks.

E.2 ADDITIONAL DETAILS ON THE NORMAL PREDICTION

Our normal map prediction module uses as input single-views of RGBA images. The architecture is based on the Pix2Pix network [138], which has been shown to perform reasonably well in normal prediction tasks [261, 220, 92]. Our goal is to maintain as much geometrical detail as possible, while making the normal predictions invariant to changes in material and illumination conditions in the input images.

E.2.1 Architecture

Our network takes RGBA images as input (RGB + background mask), and follows an encoder-decoder architecture, with 4 downsampling blocks in the encoder and 4 upsampling blocks in the decoder. In each block we repeat twice the following structure: Convolution with kernel 4×4 , a batch-normalization layer, and a leakyReLU [336]. This is done in order to reduce the impact of specular reflections in the final predictions, putting more space between the skip connection and the final output of the network. We also included residual connections within each block, as proposed by ResNet[124]. Residual connections stabilizes the network and reduces the amount of high variance noise present in the predictions. In contrast to Pix2Pix, which uses transposed convolutions, we use bilinear upsampling in order to reduce the risk of checkerboard artifacts. The final architecture is the following one:

Encoder: R64-ER64-ER128-ER256-ER512-

Bottleneck: -R512-

Decoder: -DR512-DR256-DR128-DR64-R64

where ER indicates an encoder block (downsampler) with residual connections, DR a decoder block (upsampler) with residual connections, and R a convolutional block with residual connections. The number that follows them indicates the number of filters used in the convolutions. The output uses a hyperbolic tangent function (\tanh), bounding the results of the predictions to $[-1, 1]$, which are then scaled to have unit length, and normalized to the range $[0, 1]$. The network's weights are initialized with a zero-mean normal distribution and a standard deviation of 0.02.

E.2.2 Losses

Our loss function is described in Equation E.3 and it is composed of three different losses: an adversarial loss \mathcal{L}_{adv} , a perceptual loss \mathcal{L}_{vgg} , and a reconstruction loss \mathcal{L}_{rec} .

ADVERSARIAL LOSS To infer normal maps similar to their groundtruth distribution we rely on an adversarial loss \mathcal{L}_{adv} with a binary cross entropy (BCE) function. We rely on the same discriminator model as the one proposed in Pix2Pix [138].

PERCEPTUAL LOSS To keep high-frequency details in the inferred normals we include a perceptual loss [144] \mathcal{L}_{vgg} . To extract image features we employ the VGG16 [289] model pretrained on ImageNet [61] and compute feature differences with an L_1 loss.

RECONSTRUCTION LOSS To directly supervise the prediction of each normal we rely on a Mean Squared Error (MSE) function \mathcal{L}_{rec} . Since normal vectors have unit-norm, the MSE is equivalent to a cosine distance, which has additional geometric properties.

To obtain our final loss we set the different weights to $\lambda_{adv} = 0.25$, $\lambda_{rec} = 10$, and $\lambda_{vgg} = 1$. Our final loss function is:

$$\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{vgg}\mathcal{L}_{vgg}. \quad (\text{E.3})$$

E.2.3 Training

The model was trained on synthetic data with paired groundtruth normal maps. The synthetic dataset was composed of 12 different geometries, with 5 different viewpoints, 6 different illumination conditions, and 100 different materials each; accounting for a total of 42000 images of size 128×128 px. We implemented several data augmentation techniques, including random 90 degree rotations, flips, and random gamma, hue, saturation, and brightness changes. Adam optimizer [163] is used with an initial learning rate of 0.0007, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Our network is implemented using Pytorch [244] and Pytorch Lightning [73] as our frameworks. The model was trained until evaluation losses plateaued for more than 10 epochs, which usually occurred after around 70 epochs. Overall, training took 7 hours in a single NVIDIA RTX 3080 and an AMD Ryzen 9 5900x.

E.3 ADDITIONAL RESULTS

In Figure E.1 and E.2, we show results when editing real or synthetic images with the attribute *Metallic* and *Glossy* respectively, sampling the attributes at different values along their range.



Figure E.1: Editing results by varying the perceptual attributes *Metallic*. First column is the input image, following ones show the edited image when sampling the attribute as $[-1, 0, 0.25, 0.5, 1]$.



Figure E.2: Editing results by varying the perceptual attributes *Glossy*. First column is the input image, following ones show the edited image when sampling the attribute as $[-1, -0.25, 0, 0.5, 1]$.

BIBLIOGRAPHY

- [1] Temporal properties of material categorization and material rating: visual vs non-visual material features. *Vision Research* 115 (2015), 259 – 270.
- [2] ADELSON, E. H. Lightness perception and lightness illusions. the new cognitive neurosciences, 2000.
- [3] ADELSON, E. H. On seeing stuff: the perception of materials by humans and machines. In *Human Vision and Electronic Imaging* (2001), vol. 4299, pp. 1–13.
- [4] ADELSON, E. H. Image statistics and surface perception. In *Human Vision and Electronic Imaging* (2008), vol. 6806, p. 680602.
- [5] AGARWAL, S., WILLS, J., CAYTON, L., LANCKRIET, G., KRIEGMAN, D., AND BELONGIE, S. Generalized non-metric multidimensional scaling. In *Artificial Intelligence and Statistics* (2007), pp. 11–18.
- [6] AIREY, D. *Logo Design Love: A Guide to Creating Iconic Brand Identities*. Peachpit Press, 2016.
- [7] ALIAGA, C., CASTILLO, C., GUTIERREZ, D., OTADUY, M. A., LOPEZ-MORENO, J., AND JARABO, A. An appearance model for textile fibers. *Eurographics Symposium on Rendering* 36, 4 (2017).
- [8] ALIAGA, C., O’SULLIVAN, C., GUTIERREZ, D., AND TAMSTORF, R. Sackcloth or silk?: the impact of appearance vs dynamics on the perception of animated cloth. In *ACM Symposium on Applied Perception* (2015), ACM, pp. 41–46.
- [9] ANDERSON, B. L. Visual perception of materials and surfaces. *Current Biology* 21, 24 (2011), R978–R983.
- [10] ANDERSON, B. L., AND KIM, J. Image statistics do not explain the perception of gloss and lightness. *Journal of Vision* 9, 11 (2009), 10–10.
- [11] BAGHER, M. M., SOLER, C., AND HOLZSCHUCH, N. Accurate fitting of measured reflectances using a shifted gamma micro-facet distribution. In *Computer Graphics Forum* (2012), vol. 31, Wiley Online Library, pp. 1509–1518.
- [12] BAI, X., YANG, X., LATECKI, L. J., LIU, W., AND TU, Z. Learning context-sensitive shape similarity by graph transduction. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 32, 5 (2010), 861–874.
- [13] BANERJEE, A., AND DAVE, R. N. Validating clusters using the hopkins statistic. In *Proc. of IEEE International Conference on Fuzzy Systems* (2004), vol. 1, pp. 149–153.
- [14] BARLA, P., PACANOWSKI, R., AND VANGORP, P. A composite brdf model for hazy gloss. In *Computer Graphics Forum* (2018), vol. 37, pp. 55–66.
- [15] BARNARD, M. *Graphic design as communication*. Routledge, 2013.

- [16] BARRON, J. T., AND MALIK, J. Color constancy, intrinsic images, and shape estimation. In *Proc. European Conference on Computer Vision* (2012), pp. 57–70.
- [17] BARRON, J. T., AND MALIK, J. Shape, illumination, and reflectance from shading. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 37, 8 (2014), 1670–1687.
- [18] BATES, C. L., CRAGUN, B. J., AND DAY, P. R. Automatic icon generation, Sept. 24 2002. US Patent 6,456,307.
- [19] BECK, J., AND PRAZDNY, S. Highlights and the perception of glossiness. *Attention, Perception, & Psychophysics* 30, 4 (1981), 407–410.
- [20] BEIGPOUR, S., SHEKHAR, S., MANSOURYAR, M., MYSZKOWSKI, K., AND SEIDEL, H.-P. Light-field appearance editing based on intrinsic decomposition. *Journal of Perceptual Imaging* 1, 1 (2018), 10502–1.
- [21] BELL, S., AND BALA, K. Learning Visual Similarity for Product Design with Convolutional Neural Networks. *ACM Trans. on Graphics (Proc. SIGGRAPH)* 34, 4 (2015).
- [22] BELL, S., UPCHURCH, P., SNAVELY, N., AND BALA, K. Opensurfaces: A richly annotated catalog of surface appearance. *ACM Trans. on Graphics* 32, 4 (2013), 111.
- [23] BELL, S., UPCHURCH, P., SNAVELY, N., AND BALA, K. Material recognition in the wild with the materials in context database. In *Proc. Computer Vision and Pattern Recognition* (2015), pp. 3479–3487.
- [24] BELONGIE, S., MALIK, J., AND PUZICHA, J. Shape matching and object recognition using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24, 4 (2002), 509–522.
- [25] BEN-ARTZI, A., EGAN, K., DURAND, F., AND RAMAMOORTHI, R. A precomputed polynomial representation for interactive brdf editing with global illumination. *ACM Trans. on Graphics* 27, 2 (2008), 1–13.
- [26] BERNSTEIN, G. L., AND LI, W. Lillicon: using transient widgets to create scale variations of icons. *ACM Trans. on Graphics* 34, 4 (2015), 144.
- [27] BIERON, J., AND PEERS, P. An adaptive brdf fitting metric. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 59–74.
- [28] BLAKE, A., AND BÜLTHOFF, H. Does the brain know the physics of specular reflection? *Nature* 343, 6254 (1990), 165.
- [29] BLOJ, M. G., KERSTEN, D., AND HURLBERT, A. C. Perception of three-dimensional shape influences colour perception through mutual illumination. *Nature* 402, 6764 (1999), 877.
- [30] BOBER, M. Mpeg-7 visual shape descriptors. *IEEE Transactions on Circuits and Systems for Video Technology* 11, 6 (2001), 716–719.
- [31] BOSS, M., BRAUN, R., JAMPANI, V., BARRON, J. T., LIU, C., AND LENSCH, H. Nerd: Neural reflectance decomposition from image collections. *arXiv preprint arXiv:2012.03918* (2020).

-
- [32] BOUSSEAU, A., CHAPOULIE, E., RAMAMOORTHI, R., AND AGRAWALA, M. Optimizing environment maps for material depiction. In *Computer Graphics Forum* (2011), vol. 30, pp. 1171–1180.
- [33] BOYADZHIEV, I., BALA, K., PARIS, S., AND ADELSON, E. Band-sifting decomposition for image-based material editing. *ACM Trans. on Graphics* 34, 5 (2015), 1–16.
- [34] BRADSKI, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).
- [35] BRADY, T., AND OLIVA, A. Spatial frequency integration during active perception: perceptual hysteresis when an object recedes. *Frontiers in Psychology* 3 (2012), 462.
- [36] BROMLEY, J., GUYON, I., LECUN, Y., SÄCKINGER, E., AND SHAH, R. Signature verification using a "siamese" time delay neural network. In *NIPS Proc* (1994).
- [37] BROSSIER, P., BELLO, J. P., AND PLUMBLEY, M. D. Real-time temporal segmentation of note objects in music signals. In *Proc. of ICMC 2004, the 30th Annual International Computer Music Conference* (2004).
- [38] BURT, P., AND ADELSON, E. The laplacian pyramid as a compact image code. *IEEE Transactions on communications* 31, 4 (1983), 532–540.
- [39] CAVDAN, M., DREWING, K., AND DOERSCHNER, K. Materials in action: The look and feel of soft. *bioRxiv* (2021).
- [40] CHADWICK, A., AND KENTRIDGE, R. The perception of gloss: A review. *Vision Research* 109 (2015), 221 – 235.
- [41] CHAMPION, R. A., AND WARREN, P. A. Contrast effects on speed perception for linear and radial motion. *Vision Research* 140 (2017), 66–72.
- [42] CHANDRAKER, M., AND RAMAMOORTHI, R. What an image reveals about material reflectance. In *Proc. International Conference on Computer Vision* (2011).
- [43] CHEN, B., WANG, C., PIOVARČI, M., SEIDEL, H.-P., DIDYK, P., MYSZKOWSKI, K., AND SERRANO, A. The effect of geometry and illumination on appearance perception of different material categories. *The Visual Computer* (2021), 1–13.
- [44] CHEN, K., XU, K., YU, Y., WANG, T.-Y., AND HU, S.-M. Magic decorator: automatic material suggestion for indoor digital scenes. *ACM Trans. on Graphics* 34, 6 (2015), 232.
- [45] CHENG, D., GONG, Y., ZHOU, S., WANG, J., AND ZHENG, N. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proc. Computer Vision and Pattern Recognition* (2016), pp. 1335–1344.
- [46] CHOI, Y., UH, Y., YOO, J., AND HA, J.-W. Stargan v2: Diverse image synthesis for multiple domains. In *Proc. Computer Vision and Pattern Recognition* (2020), pp. 8188–8197.
- [47] CHRISTOU, C. G., AND KOENDERINK, J. J. Light source dependence in shape from shading. *Vision Research* 37, 11 (1997).

- [48] COLLET, A., CHUANG, M., SWEENEY, P., GILLET, D., EVSEEV, D., CALABRESE, D., HOPPE, H., KIRK, A., AND SULLIVAN, S. High-quality streamable free-viewpoint video. *ACM Trans. on Graphics* 34, 4 (2015).
- [49] CRANDALL, D. J., AND HUTTENLOCHER, D. P. Weakly supervised learning of part-based spatial models for visual object recognition. In *Proc. European Conference on Computer Vision* (2006), A. Leonardis, H. Bischof, and A. Pinz, Eds., pp. 16–29.
- [50] CUNNINGHAM, D. W., WALLRAVEN, C., FLEMING, R. W., AND STRASSER, W. Perceptual reparameterization of material properties. In *Proc. Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging* (2007), pp. 89–96.
- [51] DANA, K. J., VAN GINNEKEN, B., NAYAR, S. K., AND KOENDERINK, J. J. Reflectance and texture of real-world surfaces. *ACM Trans. on Graphics* 18, 1 (Jan. 1999), 1–34.
- [52] DAZ.
- [53] DEBEVEC, P.
- [54] DEBEVEC, P., HAWKINS, T., TCHOU, C., DUIKER, H.-P., SAROKIN, W., AND SAGAR, M. Acquiring the reflectance field of a human face. In *ACM Trans. on Graphics (Proc. SIGGRAPH)* (2000).
- [55] DEBEVEC, P., YU, Y., AND BORSHUKOV, G. Efficient view-dependent image-based rendering with projective texture-mapping. In *Eurographics Symposium on Rendering* (1998).
- [56] DELANOY, J., LAGUNAS, M., CONDOR, J., MASIA, B., AND GUTIERREZ, D. A generative framework for image-based editing of material appearance using perceptual attributes. *Computer Graphics Forum Accepted with major revisions* (2021).
- [57] DELANOY, J., LAGUNAS, M., GALVE, I., GUTIERREZ, D., SERRANO, A., FLEMING, R., AND MASIA, B. The role of objective and subjective measures in material similarity learning. In *ACM SIGGRAPH Posters* (2020).
- [58] DELANOY, J., SERRANO, A., MASIA, B., AND GUTIERREZ, D. Perception of material appearance: A comparison between painted and rendered images. *Journal of Vision* 21, 5 (05 2021), 16–16.
- [59] DEMIRALP, Ç., BERNSTEIN, M. S., AND HEER, J. Learning perceptual kernels for visualization design. *IEEE Trans. on Visualization and Computer Graphics* 20, 12 (2014), 1933–1942.
- [60] DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., AND FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In *Proc. Computer Vision and Pattern Recognition* (2009), Ieee, pp. 248–255.
- [61] DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., AND FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In *Proc. Computer Vision and Pattern Recognition* (2009), pp. 248–255.
- [62] DI CICCO, F., WIJNTJES, M. W., AND PONT, S. C. Understanding gloss perception through the lens of art: Combining perception, image analysis, and painting recipes of 17th century painted grapes. *Journal of Vision* 19, 3 (2019), 7–7.

-
- [63] DOERSCHNER, K., FLEMING, R. W., YILMAZ, O., SCHRATER, P. R., HARTUNG, B., AND KERSTEN, D. Visual motion and the perception of surface material. *Current Biology* 21, 23 (2011), 2010–2016.
- [64] DOERSCHNER, K., FLEMING, R. W., YILMAZ, O., SCHRATER, P. R., HARTUNG, B., AND KERSTEN, D. Visual motion and the perception of surface material. *Current Biology* 21, 23 (2011), 2010 – 2016.
- [65] DORSEY, J., RUSHMEIER, H., AND SILLION, F. *Digital modeling of material appearance*. 2010.
- [66] DOULAMIS, A. D., AND DOULAMIS, N. D. Generalized nonlinear relevance feedback for interactive content-based retrieval and organization. *IEEE Transactions on Circuits and Systems for Video Technology* 14, 5 (2004), 656–671.
- [67] DROR, R. O., ADELSON, E. H., AND WILLSKY, A. S. Estimating surface reflectance properties from images under unknown illumination. In *Human Vision and Electronic Imaging* (2001), vol. 4299, pp. 231–243.
- [68] DROR, R. O., ADELSON, E. H., AND WILLSKY, A. S. Surface reflectance estimation and natural illumination statistics.
- [69] DU, S.-P., MASIA, B., HU, S.-M., AND GUTIERREZ, D. A metric of visual comfort for stereoscopic motion. *ACM Trans. on Graphics* 32, 6 (2013), 222.
- [70] DUPUY, J., AND JAKOB, W. An adaptive parameterization for efficient material acquisition and rendering. *ACM Trans. on Graphics* 37, 6 (2018), 1–14.
- [71] DURAND, F., HOLZSCHUCH, N., SOLER, C., CHAN, E., AND SILLION, F. X. A frequency analysis of light transport. *ACM Trans. on Graphics* 24, 3 (2005), 1115–1126.
- [72] EL-NAQA, I., YANG, Y., GALATSANOS, N. P., NISHIKAWA, R. M., AND WERNICK, M. N. A similarity learning approach to content-based image retrieval: application to digital mammography. *IEEE Transactions on Medical Imaging* 23, 10 (Oct 2004), 1233–1244.
- [73] FALCON, WA, E. A. Pytorch lightning. *GitHub. Note: <https://github.com/PyTorchLightning/pytorch-lightning>* (2019).
- [74] FARNUNG-LAURSEN, L., KOYAMA, Y., CHEN, H.-T., GARCES, E., GUTIERREZ, D., HARPER, R., AND IGARASHI, T. Icon Set Selection via Human Computation. In *Pacific Graphics Short Papers* (2016).
- [75] FAUL, F. The influence of fresnel effects on gloss perception. *Journal of Vision* 19, 13 (2019), 1–1.
- [76] FERWERDA, J. A., PELLACINI, F., AND GREENBERG, D. P. Psychophysically based model of surface gloss perception. In *Human Vision and Electronic Imaging VI* (2001), vol. 4299, pp. 291–302.
- [77] FILIP, J., CHANTLER, M. J., GREEN, P. R., AND HAINDL, M. A psychophysically validated metric for bidirectional texture data reduction. *ACM Trans. on Graphics* 27, 5 (2008), 138–1.
- [78] FILIP, J., AND KOLAFOVÁ, M. Perceptual attributes analysis of real-world materials. *ACM Trans. on Applied Perception (TAP)* 16, 1 (2019), 1.
-

- [79] FILIP, J., AND VÁVRA, R. Template-based sampling of anisotropic brdfs. In *Computer Graphics Forum* (2014), vol. 33, pp. 91–99.
- [80] FISCHER, S., ŠROUBEK, F., PERRINET, L., REDONDO, R., AND CRISTÓBAL, G. Self-invertible 2d log-gabor wavelets. *International Journal of Computer Vision* 75, 2 (2007), 231–246.
- [81] FLEMING, R. W. Human perception: Visual heuristics in the perception of glossiness. *Current Biology* 22, 20 (2012), R865–R866.
- [82] FLEMING, R. W. Visual perception of materials and their properties. *Vision Research* 94 (2014), 62 – 75.
- [83] FLEMING, R. W. Material perception. *Annual Review of Vision Science* 3 (2017), 365–388.
- [84] FLEMING, R. W., AND BÜLTHOFF, H. H. Low-level image cues in the perception of translucent materials. *ACM Trans. on Applied Perception (TAP)* 2, 3 (2005), 346–382.
- [85] FLEMING, R. W., DROR, R. O., AND ADELSON, E. H. How do humans determine reflectance properties under unknown illumination?
- [86] FLEMING, R. W., DROR, R. O., AND ADELSON, E. H. Real-world illumination and the perception of surface reflectance properties. *Journal of Vision* 3, 5 (2003), 3–3.
- [87] FLEMING, R. W., GEGENFURTNER, K. R., AND NISHIDA, S. Visual perception of materials: the science of stuff. *Vision Research*, 109 (2015), 123–124.
- [88] FLEMING, R. W., NISHIDA, S., AND GEGENFURTNER, K. R. Perception of material properties. *Vision Research* 115 (2015), 157 – 162.
- [89] FLEMING, R. W., AND STORRS, K. R. Learning to see stuff. *Current Opinion in Behavioral Sciences* 30 (2019), 100–108.
- [90] FLEMING, R. W., WIEBEL, C., AND GEGENFURTNER, K. Perceptual qualities and material classes. *Journal of Vision* 13, 8 (2013), 9–9.
- [91] FORES, A., FERWERDA, J., AND GU, J. Toward a perceptually based metric for brdf modeling. In *Color and Imaging Conference* (2012), vol. 2012, Society for Imaging Science and Technology, pp. 142–148.
- [92] GABEUR, V., FRANCO, J.-S., MARTIN, X., SCHMID, C., AND ROGEZ, G. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proc. International Conference on Computer Vision* (2019).
- [93] GAO, J., AND NEVATIA, R. Revisiting temporal modeling for video-based person reid. *arXiv preprint arXiv:1805.02104* (2018).
- [94] GARCES, E., AGARWALA, A., GUTIERREZ, D., AND HERTZMANN, A. A Similarity Measure for Illustration Style. *ACM Trans. on Graphics (Proc. SIGGRAPH)* 33, 4 (2014).
- [95] GARCES, E., AGARWALA, A., HERTZMANN, A., AND GUTIERREZ, D. Style-based Exploration of Illustration Datasets. *Multimedia Tools and Applications* 76, 11 (2017).

-
- [96] GARCES, E., ECHEVARRIA, J. I., ZHANG, W., WU, H., ZHOU, K., AND GUTIERREZ, D. Intrinsic light field images. In *Computer Graphics Forum* (2017), vol. 36.
- [97] GARCES, E., MUNOZ, A., LOPEZ-MORENO, J., AND GUTIERREZ, D. Intrinsic images by clustering. *Computer Graphics Forum* 31, 4 (2012).
- [98] GARDNER, M.-A., SUNKAVALLI, K., YUMER, E., SHEN, X., GAMBARETTO, E., GAGNÉ, C., AND LALONDE, J.-F. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090* (2017).
- [99] GED, G., OBEIN, G., SILVESTRI, Z., LE ROHELLEC, J., AND VIÉNOT, F. Recognizing real materials from their glossy appearance. *Journal of Vision* 10, 9 (2010), 18–18.
- [100] GEISLER, W. S. Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology* 59 (2008), 167–192.
- [101] GEORGOULIS, S., REMATAS, K., RITSCHER, T., GAVVES, E., FRITZ, M., VAN GOOL, L., AND TUYTELAARS, T. Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 40, 8 (2017).
- [102] GEORGOULIS, S., VANWEDDINGEN, V., PROESMANS, M., AND VAN GOOL, L. Material classification under natural illumination using reflectance maps. In *IEEE Winter Conference on Applications of Computer Vision* (2017), IEEE, pp. 244–253.
- [103] GIESEL, M., AND ZAIDI, Q. Frequency-based heuristics for material perception. *Journal of Vision* 13, 14 (2013), 7–7.
- [104] GKIOULEKAS, I., WALTER, B., ADELSON, E. H., BALA, K., AND ZICKLER, T. On the appearance of translucent edges. In *Proc. Computer Vision and Pattern Recognition* (2015), pp. 5528–5536.
- [105] GKIOULEKAS, I., WALTER, B., ADELSON, E. H., BALA, K., AND ZICKLER, T. On the appearance of translucent edges. In *Proc. Computer Vision and Pattern Recognition* (2015), pp. 5528–5536.
- [106] GKIOULEKAS, I., XIAO, B., ZHAO, S., ADELSON, E. H., ZICKLER, T., AND BALA, K. Understanding the role of phase function in translucent appearance. *ACM Trans. on Graphics* 32, 5 (2013), 147.
- [107] GLOROT, X., AND BENGIO, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (2010), vol. 9 of *Proceedings of Machine Learning Research*, PMLR, pp. 249–256.
- [108] GOODFELLOW, I., BENGIO, Y., AND COURVILLE, A. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [109] GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDEFARLEY, D., OZAIR, S., COURVILLE, A., AND BENGIO, Y. Generative adversarial networks. *arXiv preprint arXiv:1406.2661* (2014).
- [110] GRAMAZIO, C. C., LAIDLAW, D. H., AND SCHLOSS, K. B. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Trans. on Visualization and Computer Graphics* 23, 1 (2017), 521–530.

- [111] GUARNERA, D., GUARNERA, G. C., TOSCANI, M., GLENCROSS, M., LI, B., HARDEBERG, J. Y., AND GEGENFURTNER, K. R. Perceptually validated analytical brdfs parameters remapping. In *ACM Trans. on Graphics (Proc. SIGGRAPH)* (2018).
- [112] GULRAJANI, I., AHMED, F., ARJOVSKY, M., DUMOULIN, V., AND COURVILLE, A. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028* (2017).
- [113] GUO, J., GUO, Y., PAN, J., AND LU, W. Brdf analysis with directional statistics and its applications. *IEEE Trans. on Visualization and Computer Graphics* (2018).
- [114] GUO, K., LINCOLN, P., DAVIDSON, P., BUSCH, J., YU, X., WHALEN, M., HARVEY, G., ORTS-ESCOLANO, S., PANDEY, R., DOURGARIAN, J., ET AL. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Trans. on Graphics* 38, 6 (2019).
- [115] GUTIERREZ, D., SERON, F. J., LOPEZ-MORENO, J., SANCHEZ, M. P., FANDOS, J., AND REINHARD, E. Depicting procedural caustics in single images. *ACM Trans. on Graphics* 27, 5 (2008), 1–9.
- [116] HABER, T., FUCHS, C., BEKAER, P., SEIDEL, H.-P., GOESELE, M., AND LENSCH, H. P. Relighting objects from image collections. In *Proc. Computer Vision and Pattern Recognition* (2009), pp. 627–634.
- [117] HARTUNG, B., AND KERSTEN, D. Distinguishing shiny from matte. *Journal of Vision* 2, 7 (2002), 551–551.
- [118] HAVRAN, V., FILIP, J., AND MYSZKOWSKI, K. Perceptually Motivated BRDF Comparison using Single Image. *Computer Graphics Forum* (2016).
- [119] HAWKEN, M. J., AND PARKER, A. J. Spatial properties of neurons in the monkey striate cortex. *Proc. of the Royal society of London. Series B.* 231, 1263 (1987), 251–288.
- [120] HAYMAN, E., CAPUTO, B., FRITZ, M., AND EKLUNDH, J.-O. On the significance of real-world conditions for material classification. In *Proc. European Conference on Computer Vision* (2004), Springer, pp. 253–266.
- [121] HDRIHAVEN.
- [122] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. *CoRR abs/1512.03385* (2015).
- [123] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *Proc. Computer Vision and Pattern Recognition* (2016), pp. 770–778.
- [124] HE, K., ZHANG, X., REN, S., AND SUN, J. Deep residual learning for image recognition. In *Proc. Computer Vision and Pattern Recognition* (2016).
- [125] HEDMAN, P., RITSCHER, T., DRETTAKIS, G., AND BROSTOW, G. Scalable inside-out image-based rendering. *ACM Trans. on Graphics* 35, 6 (2016).
- [126] HEER, J., AND BOSTOCK, M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proc. SIGCHI* (2010), ACM, pp. 203–212.

-
- [127] HERTZMANN, A. Why do line drawings work? a realism hypothesis. *Perception* 49, 4 (2020), 439–451.
- [128] HO, Y.-X., LANDY, M. S., AND MALONEY, L. T. How direction of illumination affects visually perceived surface roughness. *Journal of Vision* 6, 5 (2006), 8–8.
- [129] HOLD-GEOFFROY, Y., ATHAWALE, A., AND LALONDE, J.-F. Deep sky modeling for single image outdoor lighting estimation. In *Proc. Computer Vision and Pattern Recognition* (2019).
- [130] HOLD-GEOFFROY, Y., SUNKAVALLI, K., HADAP, S., GAMBARETTO, E., AND LALONDE, J.-F. Deep outdoor illumination estimation. In *Proc. Computer Vision and Pattern Recognition* (2017).
- [131] HORTON, W. K. *The icon book: Visual symbols for computer systems and documentation*. John Wiley & Sons, Inc., 1994.
- [132] HOU, K.-C., HO, C.-H., ET AL. A preliminary study on aesthetic of apps icon design. In *IASDR 2013 5th International Congress of International Association of Societies of Design Research* (2013), pp. 1–12.
- [133] HU, M.-K. Visual pattern recognition by moment invariants. *IRE transactions on information theory* 8, 2 (1962), 179–187.
- [134] HUANG, G., LIU, Z., VAN DER MAATEN, L., AND WEINBERGER, K. Q. Densely connected convolutional networks. In *Proc. Computer Vision and Pattern Recognition* (2017), vol. 1, p. 3.
- [135] HUNTER, R. S., ET AL. Methods of determining gloss. *NBS Research paper RP 958* (1937).
- [136] IKEUCHI, K., AND HORN, B. K. Numerical shape from shading and occluding boundaries. *Artificial Intelligence* 17, 1-3 (1981), 141–184.
- [137] IOFFE, S., AND SZEGEDY, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR abs/1502.03167* (2015).
- [138] ISOLA, P., ZHU, J.-Y., ZHOU, T., AND EFROS, A. A. Image-to-image translation with conditional adversarial networks. In *Proc. Computer Vision and Pattern Recognition* (July 2017).
- [139] JAKOB, W. Mitsuba renderer. <http://www.mitsuba-renderer.org>.
- [140] JAKOB, W. Mitsuba renderer, 2010. <http://www.mitsuba-renderer.org>.
- [141] JARABO, A., MASIA, B., BOUSSEAU, A., PELLACINI, F., AND GUTIERREZ, D. How do people edit light fields. *ACM Trans. on Graphics* 33, 4 (2014), 4.
- [142] JARABO, A., WU, H., DORSEY, J., RUSHMEIER, H., AND GUTIERREZ, D. Effects of approximate filtering on the appearance of bidirectional texture functions. *IEEE Trans. on Visualization and Computer Graphics* 20, 6 (2014).
- [143] JIMENEZ, J., ZSOLNAI, K., JARABO, A., FREUDE, C., AUZINGER, T., WU, X.-C., VON DER PAHLEN, J., WIMMER, M., AND GUTIERREZ, D. Separable subsurface scattering. *Computer Graphics Forum* 34, 6 (2015).

- [144] JOHNSON, J., ALAHI, A., AND FEI-FEI, L. Perceptual losses for real-time style transfer and super-resolution. In *Proc. European Conference on Computer Vision* (2016), Springer, pp. 694–711.
- [145] JOHNSON, M. K., AND ADELSON, E. H. Shape estimation in natural illumination. In *Proc. Computer Vision and Pattern Recognition* (2011).
- [146] JULESZ, B. Visual pattern discrimination. *IRE transactions on Information Theory* 8, 2 (1962), 84–92.
- [147] KALIDEO. Remove.bg.
- [148] KANAMORI, Y., AND ENDO, Y. Relighting humans: occlusion-aware inverse rendering for full-body human images. In *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)* (2018).
- [149] KAWABE, T., MARUYA, K., FLEMING, R. W., AND NISHIDA, S. Seeing liquids from visual motion. *Vision research* 109 (2015), 125–138.
- [150] KAWATO, M., HAYAKAWA, H., AND INUI, T. A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network: Computation in Neural Systems* 4, 4 (1993), 415–422.
- [151] KENDALL, M., AND GIBBONS, J. D. *Rank Correlation Methods*, 5 ed. A Charles Griffin Title, September 1990.
- [152] KENDALL, M. G., AND BABINGTON-SMITH, B. On the method of paired comparisons. *Biometrika* 31 (1940), 324–345.
- [153] KERR, W. B., AND PELLACINI, F. Toward evaluating material design interface paradigms for novice users. In *ACM Trans. on Graphics* (2010), vol. 29, ACM, p. 35.
- [154] KERSTEN, D., MAMASSIAN, P., AND YUILLE, A. Object perception as bayesian inference. *Annual Review of Psychology* 55 (2004), 271–304.
- [155] KETTNER, L., RAAB, M., SEIBERT, D., JORDAN, J., AND KELLER, A. The Material Definition Language. In *Workshop on Material Appearance Modeling* (2015), R. Klein and H. Rushmeier, Eds., The Eurographics Association.
- [156] KHAN, E. A., REINHARD, E., FLEMING, R. W., AND BÜLTHOFF, H. H. Image-based material editing. *ACM Trans. on Graphics* 25, 3 (2006), 654–663.
- [157] KHOLGADE, N., SIMON, T., EFROS, A., AND SHEIKH, Y. 3d object manipulation in a single photograph using stock 3d models. *ACM Trans. on Graphics* 33, 4 (2014).
- [158] KHOTANZAD, A., AND HONG, Y. H. Invariant image recognition by zernike moments. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 12, 5 (1990), 489–497.
- [159] KIM, J., AND ANDERSON, B. L. Image statistics and the perception of surface gloss and lightness. *Journal of Vision* 10, 9 (2010), 3–3.
- [160] KIM, K., GU, J., TYREE, S., MOLCHANOV, P., NIESSNER, M., AND KAUTZ, J. A lightweight approach for on-the-fly reflectance estimation. In *Proc. International Conference on Computer Vision* (2017), pp. 20–28.

-
- [161] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [162] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2014).
- [163] KINGMA, D. P., AND BA, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [164] KLEIMAN, Y., VAN KAICK, O., SORKINE-HORNUNG, O., AND COHEN-OR, D. Shed: shape edit distance for fine-grained shape similarity. *ACM Trans. on Graphics* 34, 6 (2015), 235.
- [165] KŘIVÁNEK, J., FERWERDA, J. A., AND BALA, K. Effects of global illumination approximations on material appearance. *ACM Trans. on Graphics (Proc. SIGGRAPH)* 29, 4 (2010), 112:1–112:10.
- [166] KULKARNI, T. D., WHITNEY, W., KOHLI, P., AND TENENBAUM, J. B. Deep convolutional inverse graphics network. *arXiv preprint arXiv:1503.03167* (2015).
- [167] KWAN, K. C., SINN, L. T., HAN, C., WONG, T.-T., AND FU, C.-W. Pyramid of Arclength Descriptor for Generating Collage of Shapes. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)* 35, 6 (2016), 1–12.
- [168] LAB, I. V. . G.
- [169] LAGUNAS, M., GARCES, E., AND GUTIERREZ, D. Learning icons appearance similarity. *Multimedia Tools and Applications* (2018), 1–19.
- [170] LAGUNAS, M., GARCES, E., AND GUTIERREZ, D. Learning icons appearance similarity. *Multimedia Tools and Applications* 78, 8 (2019), 10733–10751.
- [171] LAGUNAS, M., MALPICA, S., SERRANO, A., GARCES, E., GUTIERREZ, D., AND MASIA, B. A similarity measure for material appearance. *ACM Trans. on Graphics (Proc. SIGGRAPH)* 38, 4 (2019).
- [172] LAGUNAS, M., SERRANO, A., GUTIERREZ, D., AND MASIA, B. The joint role of geometry and illumination on material recognition. *Journal of Vision* 21 (2021).
- [173] LAGUNAS, M., SUN, X., YANG, J., VILLEGAS, R., ZHANG, J., SHU, Z., MASIA, B., AND GUTIERREZ, D. Single-image full-body human relighting. In *Eurographics Symposium on Rendering* (2021), The Eurographics Association.
- [174] LAMPLE, G., ZEGHIDOUR, N., USUNIER, N., BORDES, A., DENOYER, L., AND RANZATO, M. A. Fader networks: manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems* (2017), vol. 30.
- [175] LAND, E. H., AND McCANN, J. J. Lightness and retinex theory. *JOSA* 61, 1 (1971).
- [176] LATECKI, L. J., LAKAMPER, R., AND ECKHARDT, T. Shape descriptors for non-rigid shapes with a single closed contour. In *IEEE Winter Conference on Applications of Computer Vision* (2000), vol. 1, IEEE, pp. 424–429.

- [177] LAWRENCE, J., BEN-ARTZI, A., DECORO, C., MATUSIK, W., PFISTER, H., RAMAMOORTHI, R., AND RUSINKIEWICZ, S. Inverse shade trees for non-parametric material representation and editing. *ACM Trans. on Graphics* 25, 3 (2006), 735–745.
- [178] LECUN, Y., BOSER, B., DENKER, J. S., HENDERSON, D., HOWARD, R. E., HUBBARD, W., AND JACKEL, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 4 (Dec. 1989), 541–551.
- [179] LEE, S., AND LEE, D. Fusion of ir and visual images based on gaussian and laplacian decomposition using histogram distributions and edge selection. *Mathematical Problems in Engineering* 2016 (2016).
- [180] LEGENDRE, C., MA, W.-C., FYFFE, G., FLYNN, J., CHARBONNEL, L., BUSCH, J., AND DEBEVEC, P. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proc. Computer Vision and Pattern Recognition* (2019), pp. 5918–5928.
- [181] LEHTINEN, J. A framework for precomputed and captured light transport. *ACM Trans. on Graphics* 26, 4 (2007), 13–es.
- [182] LOUP, F. B., POINTER, M. R., DUTRÉ, P., AND HANSELAER, P. Geometry of illumination, luminance contrast, and gloss perception. *Journal of the Optical Society of America A* 27, 9 (2010), 2046–2054.
- [183] LEMPITSKY, V., AND IVANOV, D. Seamless mosaicing of image-based texture maps. In *Proc. Computer Vision and Pattern Recognition* (2007).
- [184] LEUNG, T., AND MALIK, J. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision* 43, 1 (2001), 29–44.
- [185] LEWIS, J. P., ROSENHOLTZ, R., FONG, N., AND NEUMANN, U. VisualIDs : Automatic Distinctive Icons for Desktop Interfaces. *ACM Trans. on Graphics (Proc. SIGGRAPH)* 1, 212 (2004), 416–423.
- [186] LI, W., AND FRITZ, M. Recognizing materials from virtual examples. In *Proc. European Conference on Computer Vision* (2012), pp. 345–358.
- [187] LI, Z., SHAFIEL, M., RAMAMOORTHI, R., SUNKAVALLI, K., AND CHANDRAKER, M. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proc. Computer Vision and Pattern Recognition* (2020).
- [188] LI, Z., AND SNAVELY, N. Learning intrinsic image decomposition from watching the world. In *Proc. Computer Vision and Pattern Recognition* (2018), pp. 9039–9048.
- [189] LIN, D. C.-E., AND MARTELARO, N. Learning personal style from few examples. *arXiv preprint arXiv:2105.14457* (2021).
- [190] LIU, G., CEYLAN, D., YUMER, E., YANG, J., AND LIEN, J.-M. Material editing using a physically based rendering network. In *Proc. International Conference on Computer Vision* (Oct 2017).
- [191] LIU, M., DING, Y., XIA, M., LIU, X., DING, E., ZUO, W., AND WEN, S. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proc. Computer Vision and Pattern Recognition* (2019).

-
- [192] LIU, T., HERTZMANN, A., LI, W., AND FUNKHOUSER, T. Style Compatibility for 3D Furniture Models. *ACM Trans. on Graphics (Proc. SIGGRAPH)* 34, 4 (2015), 1–9.
- [193] LIU, Y., AGARWALA, A., LU, J., AND RUSINKIEWICZ, S. Data-driven iconification. In *Proc. Computational Aesthetics and Sketch Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering* (2016), Eurographics Association, pp. 113–124.
- [194] LOMBARDI, S., AND NISHINO, K. Reflectance and natural illumination from a single image. In *Proc. European Conference on Computer Vision* (2012), pp. 582–595.
- [195] LOMBARDI, S., AND NISHINO, K. Reflectance and illumination recovery in the wild. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 38, 1 (2015).
- [196] LOPEZ-MORENO, J., GARCES, E., HADAP, S., REINHARD, E., AND GUTIERREZ, D. Multiple light source estimation in a single image. In *Computer Graphics Forum* (2013), vol. 32.
- [197] LOPEZ-MORENO, J., HADAP, S., REINHARD, E., AND GUTIERREZ, D. Light source detection in photographs. In *Spanish Computer Graphics Conference (CEIG)* (2009), pp. 161–167.
- [198] LOSHCILOV, I., AND HUTTER, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [199] LU, F., CHEN, X., SATO, I., AND SATO, Y. Symps: Brdf symmetry guided photometric stereo for shape and light source estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 40, 1 (2018), 221–234.
- [200] LUN, Z., KALOGERAKIS, E., AND SHEFFER, A. Elements of Style: Learning Perceptual Shape Style Similarity. *ACM Trans. on Graphics* 34, 4 (2015), 84:1–14.
- [201] LUN, Z., KALOGERAKIS, E., AND SHEFFER, A. Elements of style: learning perceptual shape style similarity. *ACM Trans. on Graphics* 34, 4 (2015), 84.
- [202] LUPTON, E. Thinking with type. *Critical Guide for Designers, Writers, Editors & Students* (2004).
- [203] LUPTON, E., AND PHILLIPS, J. C. *Graphic Design: The New Basics: Revised and Expanded*. Chronicle Books, 2015.
- [204] MA, W.-C., CHU, H., ZHOU, B., URTASUN, R., AND TORRALBA, A. Single image intrinsic decomposition without a single intrinsic image. In *Proc. European Conference on Computer Vision* (2018), pp. 201–217.
- [205] MALONEY, L. T., AND BRAINARD, D. H. Color and material perception: Achievements and challenges. *Journal of Vision* 10, 9 (2010), 19–19.
- [206] MANTIUK, R., DALY, S., AND KEROFESKY, L. Display adaptive tone mapping. In *ACM Trans. on Graphics* (2008), vol. 27, ACM, p. 68.
- [207] MAO, R., LAGUNAS, M., MASIA, B., AND GUTIERREZ, D. The effect of motion on the perception of material appearance. In *ACM Symposium on Applied Perception* (2019), pp. 1–9.

- [208] MARLOW, P. J., AND ANDERSON, B. L. Motion and texture shape cues modulate perceived material properties. *Journal of Vision* 16, 1 (2016), 5–5.
- [209] MARLOW, P. J., KIM, J., AND ANDERSON, B. L. The perception and misperception of specular surface reflectance. *Current Biology* 22, 20 (2012), 1909–1913.
- [210] MATUSIK, W., AJDIN, B., GU, J., LAWRENCE, J., LENSCH, H. P., PELLACINI, F., AND RUSINKIEWICZ, S. Printing spatially-varying reflectance. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)* 28, 5 (2009).
- [211] MATUSIK, W., PFISTER, H., BRAND, M., AND McMILLAN, L. A data-driven reflectance model. *ACM Trans. on Graphics* 22, 3 (2003), 759–769.
- [212] MAXIMOV, M., LEAL-TAIXÉ, L., FRITZ, M., AND RITSCHER, T. Deep appearance maps. In *Proc. International Conference on Computer Vision* (2019), pp. 8729–8738.
- [213] McFEE, B., AND LANCKRIET, G. Learning Multi-modal Similarity. *Journal of Machine Learning Research* 12 (2011), 491–523.
- [214] McINNES, L., AND HEALY, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
- [215] MILDENHALL, B., SRINIVASAN, P. P., TANCIK, M., BARRON, J. T., RAMAMOORTHI, R., AND NG, R. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proc. European Conference on Computer Vision* (2020).
- [216] MINGOLLA, E., AND TODD, J. T. Perception of solid shape from shading. *Biological Cybernetics* 53, 3 (1986), 137–151.
- [217] MOHLER, D. S., AND VICK, J. H. Screen icon manipulation by context and frequency of use, Mar. 3 2015. US Patent 8,972,878.
- [218] MOTOYOSHI, I., NISHIDA, S., SHARAN, L., AND ADELSON, E. H. Image statistics and the perception of surface qualities. *Nature* 447, 7141 (2007), 206–209.
- [219] MYLO, M., GIESEL, M., ZAIDI, Q., HULLIN, M., AND KLEIN, R. Appearance bending: A perceptual editing paradigm for data-driven material models. *Vision, Modeling and Visualization. The Eurographics Association* (2017).
- [220] NATSUME, R., SAITO, S., HUANG, Z., CHEN, W., MA, C., LI, H., AND MORISHIMA, S. Siclope: Silhouette-based clothed people. In *Proc. Computer Vision and Pattern Recognition* (June 2019).
- [221] NEFS, H. T. On the visual appearance of objects. In *Product experience*. Elsevier, 2008, pp. 11–39.
- [222] NESTMEYER, T., LALONDE, J.-F., MATTHEWS, I., AND LEHRMANN, A. Learning physics-guided face relighting under directional light. In *Proc. Computer Vision and Pattern Recognition* (2020).
- [223] NGAN, A., DURAND, F., AND MATUSIK, W. Experimental Analysis of BRDF Models. In *Eurographics Symposium on Rendering* (2005), The Eurographics Association.

-
- [224] NGAN, A., DURAND, F., AND MATUSIK, W. Image-driven navigation of analytical brdf models. In *Rendering Techniques* (2006), pp. 399–407.
- [225] NGUYEN, C. H., SCHERZER, D., RITSCHER, T., AND SEIDEL, H.-P. Material editing in complex scenes by surface light field manipulation and reflectance optimization. In *Computer Graphics Forum* (2013), vol. 32, Wiley Online Library, pp. 185–194.
- [226] NIELSEN, J. B., JENSEN, H. W., RAMAMOORTHI, R., AND DIEGO, S. On Optimal , Minimal BRDF Sampling for Reflectance Acquisition. *ACM Trans. on Graphics* 34, 6 (2015), 1–11.
- [227] NISHIDA, S., AND SHINYA, M. Use of image-based information in judgements of surface-reflectance properties. *Journal of the Optical Society of America A* 15, 12 (1998), 2951–2965.
- [228] NORMAN, J. F., TODD, J. T., AND ORBAN, G. A. Perception of three-dimensional shape from specular highlights, deformations of shading, and other types of visual information. *Psychological Science* 15, 8 (2004), 565–570.
- [229] O. FRIED, S. AVIDAN, D. C.-O. Patch2vec: Globally consistent image patch representation. *Pacific Graphics* 36, 7 (2017).
- [230] OBEIN, G., KNOBLAUCH, K., AND VIÉOT, F. Difference scaling of gloss: Nonlinearity, binocularity, and constancy. *Journal of Vision* 4, 9 (2004), 4–4.
- [231] O'DONOVAN, P., AGARWALA, A., AND HERTZMANN, A. Color Compatibility From Large Datasets. *ACM Trans. on Graphics* 30, 4 (2011).
- [232] O'DONOVAN, P., LIBEKS, J., AGARWALA, A., AND HERTZMANN, A. Exploratory font selection using crowdsourced attributes. *ACM Trans. on Graphics* 33, 4 (2014), 92.
- [233] OKATANI, T., AND DEGUCHI, K. Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center. *Computer vision and image understanding* 66, 2 (1997).
- [234] OLIVA, A., AND TORRALBA, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision* 42, 3 (2001), 145–175.
- [235] OLKKONEN, M., AND BRAINARD, D. H. Perceived glossiness and lightness under real-world illumination. *Journal of Vision* 10, 9 (2010), 5–5.
- [236] OLKKONEN, M., AND BRAINARD, D. H. Joint effects of illumination geometry and object shape in the perception of surface reflectance. *i-Perception* 2, 9 (2011), 1014–1034.
- [237] OREN, M., AND NAYAR, S. K. Generalization of lambert's reflectance model. In *Proc. of SIGGRAPH* (1994).
- [238] OREN, M., AND NAYAR, S. K. A theory of specular surface geometry. *International Journal of Computer Vision* 24, 2 (1997), 105–124.
- [239] OSADA, R., FUNKHOUSER, T., CHAZELLE, B., AND DOBKIN, D. Shape distributions. *ACM Trans. on Graphics* 21, 4 (2002), 807–832.

- [240] O'SHEA, J. P., AGRAWALA, M., AND BANKS, M. S. The influence of shape cues on the perception of lighting direction. *Journal of Vision* 10, 12 (2010), 21–21.
- [241] PALMER, S. E. Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. *Explorations in Cognition* (1975), 279–307.
- [242] PARKER, S. G., BIGLER, J., DIETRICH, A., FRIEDRICH, H., HOBEROCK, J., LUEBKE, D., MCALLISTER, D., MCGUIRE, M., MORLEY, K., ROBISON, A., ET AL. Optix: a general purpose ray tracing engine. In *ACM Trans. on Graphics* (2010), vol. 29, ACM, p. 66.
- [243] PARKHI, O. M., VEDALDI, A., AND ZISSERMAN, A. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)* (2015).
- [244] PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N., ANTIGA, L., DESMAISON, A., KOPE, A., YANG, E., DEVITO, Z., RAISON, M., TEJANI, A., CHILAMKURTHY, S., STEINER, B., FANG, L., BAI, J., AND CHINTALA, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. 2019.
- [245] PELE, O., AND WERMAN, M. The quadratic-chi histogram distance family. In *Proc. European Conference on Computer Vision* (2010), pp. 749–762.
- [246] PELLACINI, F., FERWERDA, J. A., AND GREENBERG, D. P. Toward a psychophysically-based light reflection model for image synthesis. In *Proc. on Computer Graphics and Interactive Techniques* (2000), SIGGRAPH '00, p. 55–64.
- [247] PELLACINI, F., FERWERDA, J. A., AND GREENBERG, D. P. Toward a psychophysically-based light reflection model for image synthesis. In *Proc. on Computer Graphics and Interactive Techniques* (2000), pp. 55–64.
- [248] PEREIRA, T., AND RUSINKIEWICZ, S. Gamut mapping spatially varying reflectance with an improved BRDF similarity metric. In *Computer Graphics Forum* (2012), vol. 31, Wiley Online Library, pp. 1557–1566.
- [249] PHARR, M., JAKOB, W., AND HUMPHREYS, G. *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016.
- [250] PIZLO, Z. Perception viewed as an inverse problem. *Vision Research* 41, 24 (2001), 3145–3161.
- [251] POHLERT, T. Pmcmrplus: calculate pairwise multiple comparisons of mean rank sums extended. *R package version 1, 0* (2018).
- [252] RAMACHANDRAN, V. S. Perception of shape from shading. *Nature* 331, 6152 (1988).
- [253] RAMAMOORTHI, R. *Precomputation-based rendering*. 2009.
- [254] RAMAMOORTHI, R., AND HANRAHAN, P. An efficient representation for irradiance environment maps. In *Proc. on Computer Graphics and Interactive Techniques* (2001), pp. 497–500.
- [255] REDDI, S. J., KALE, S., AND KUMAR, S. On the convergence of adam and beyond.

-
- [256] REMATAS, K., RITSCHER, T., FRITZ, M., GAVVES, E., AND TUYTELAARS, T. Deep reflectance maps. In *Proc. Computer Vision and Pattern Recognition* (2016), pp. 4508–4516.
- [257] RONNEBERGER, O., FISCHER, P., AND BROX, T. U-net: Convolutional networks for biomedical image segmentation. In *Intl. Conf. on Medical Image Computing and Computer-assisted Intervention* (2015), pp. 234–241.
- [258] RUBINSTEIN, M., GUTIERREZ, D., SORKINE, O., AND SHAMIR, A. A comparative study of image retargeting. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)* 29, 6 (2010), 160:1–160:10.
- [259] RUBINSTEIN, M., GUTIERREZ, D., SORKINE, O., AND SHAMIR, A. A comparative study of image retargeting. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)* 29, 6 (2010), 160:1–160:10.
- [260] RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATY, A., KHOSLA, A., BERNSTEIN, M., ET AL. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.
- [261] SAITO, S., SIMON, T., SARAGIH, J., AND JOO, H. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proc. Computer Vision and Pattern Recognition* (2020).
- [262] SAKANO, Y. Effects of self-motion on gloss perception. *Perception* 37 *ECVP Abstract Supplement, 2008* 77 (2008).
- [263] SALEH, B., DONTCHEVA, M., HERTZMANN, A., AND LIU, Z. Learning style similarity for searching infographics. In *Proceedings of the 41st Graphics Interface Conference* (2015), GI '15, Canadian Information Processing Society, pp. 59–64.
- [264] SAVARESE, S., LI, F. F., AND PERONA, P. Can we see the shape of a mirror? *Journal of Vision* 3, 9 (2003), 74–74.
- [265] SCHAFFALITZKY, F., AND ZISSERMAN, A. Viewpoint invariant texture matching and wide baseline stereo. In *Proc. International Conference on Computer Vision* (2001), vol. 2, pp. 636–643.
- [266] SCHILLER, P. H., FINLAY, B. L., AND VOLMAN, S. F. Quantitative studies of single-cell properties in monkey striate cortex. i. spatiotemporal organization of receptive fields. *Journal of Neurophysiology* 39, 6 (1976), 1288–1319.
- [267] SCHLÜTER, N., AND FAUL, F. Visual shape perception in the case of transparent objects. *Journal of Vision* 19, 4 (2019), 24–24.
- [268] SCHMIDT, F., FLEMING, R. W., AND VALSECCHI, M. Softness and weight from shape: Material properties inferred from local shape features. *Journal of Vision* 20, 6 (2020), 2–2.
- [269] SCHMIDT, T.-W., PELLACINI, F., NOWROUZEZAHRAI, D., JAROSZ, W., AND DACHSBACHER, C. State of the art in artistic editing of appearance, lighting and material. In *Computer Graphics Forum* (2016), vol. 35, pp. 216–233.
- [270] SCHROFF, F., KALENICHENKO, D., AND PHILBIN, J. Facenet: A unified embedding for face recognition and clustering. In *Proc. Computer Vision and Pattern Recognition* (2015), pp. 815–823.

- [271] SCHULTZ, M., AND JOACHIMS, T. Learning a Distance Metric from Relative Comparisons. In *Advances in Neural Information Processing Systems* (2003).
- [272] SCHWARTZ, G., AND NISHINO, K. Material recognition from local appearance in global context. *arXiv preprint arXiv:1611.09394* (2016).
- [273] SCHWARTZ, G., AND NISHINO, K. Recognizing material properties from images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (2019).
- [274] SENGUPTA, S., KANAZAWA, A., CASTILLO, C. D., AND JACOBS, D. W. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild'. In *Proc. Computer Vision and Pattern Recognition* (2018), pp. 6296–6305.
- [275] SERRANO, A., CHEN, B., WANG, C., PIOVARČI, M., SEIDEL, H.-P., DIDYK, P., AND MYŠKOWSKI, K. The effect of shape and illumination on material perception: model and applications. *ACM Trans. on Graphics* 40, 4 (2021), 1–16.
- [276] SERRANO, A., GUTIERREZ, D., MYŠKOWSKI, K., SEIDEL, H.-P., AND MASIA, B. An intuitive control space for material appearance. *ACM Trans. on Graphics* 35, 6 (Nov. 2016), 186:1–186:12.
- [277] SETLUR, V., ALBRECHT-BUEHLER, C., GOOCH, A. A., ROSOFF, S., AND GOOCH, B. Semanticons: Visual metaphors as file icons. *Eurographics* 24, 3 (2005), 647–656.
- [278] SETLUR, V., AND MACKINLAY, J. D. Automatic generation of semantic icon encodings for visualizations. In *Proc. SIGCHI* (2014), CHI '14, pp. 541–550.
- [279] SÈVE, R. Problems connected with the concept of gloss. *Color Research & Application* 18, 4 (1993), 241–252.
- [280] SHARAN, L., ROSENHOLTZ, R., AND ADELSON, E. Material perception: What can you see in a brief glance? *Journal of Vision* 9, 8 (2009), 784–784.
- [281] SHARAN, L., ROSENHOLTZ, R., AND ADELSON, E. H. Eye movements for shape and material perception. *Journal of Vision* 8, 6 (2008), 219–219.
- [282] SHARIF RAZAVIAN, A., AZIZPOUR, H., SULLIVAN, J., AND CARLSSON, S. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proc. Computer Vision and Pattern Recognition (CVPR) Workshops* (2014), pp. 806–813.
- [283] SHUGRINA, M., LU, J., AND DIVERDI, S. Playful palette: an interactive parametric color mixer for artists. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 61.
- [284] SHUM, H., AND KANG, S. B. Review of image-based rendering techniques. In *Visual Communications and Image Processing 2000* (2000), vol. 4067.
- [285] SIKORA, T. The mpeg-7 visual standard for content description-an overview. *IEEE Transactions on Circuits and Systems for Video Technology* 11, 6 (2001), 696–702.

-
- [286] SIMO-SERRA, E., AND ISHIKAWA, H. Fashion Style in 128 Floats: Joint Ranking and Classification using Weak Data for Feature Extraction. In *Proc. Computer Vision and Pattern Recognition* (2016).
- [287] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [288] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2014).
- [289] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *CoRR* (2015).
- [290] SINHA, S. N., KOPF, J., GOESELE, M., SCHARSTEIN, D., AND SZELISKI, R. Image-based rendering for scenes with reflections. *ACM Trans. on Graphics* 31, 4 (2012).
- [291] SLOAN, P.-P., KAUTZ, J., AND SNYDER, J. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. *ACM Trans. on Graphics* 21, 3 (2002).
- [292] SOLER, C., SUBR, K., AND NOWROUZEZAHRAI, D. A versatile parameterization for measured material manifolds. In *Computer Graphics Forum* (2018), vol. 37, pp. 135–144.
- [293] SRINIVASAN, P. P., DENG, B., ZHANG, X., TANCIK, M., MILDENHALL, B., AND BARRON, J. T. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proc. Computer Vision and Pattern Recognition* (2021), pp. 7495–7504.
- [294] SRIVASTAVA, N., HINTON, G., KRIZHEVSKY, A., SUTSKEVER, I., AND SALAKHUTDINOV, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15, 1 (2014), 1929–1958.
- [295] STORRS, K. R., ANDERSON, B. L., AND FLEMING, R. W. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour* (2021), 1–16.
- [296] SUN, T., BARRON, J. T., TSAI, Y.-T., XU, Z., YU, X., FYFFE, G., RHEMANN, C., BUSCH, J., DEBEVEC, P. E., AND RAMAMOORTHI, R. Single image portrait relighting. *ACM Trans. on Graphics* 38, 4 (2019).
- [297] SUN, T., JENSEN, H. W., AND RAMAMOORTHI, R. Connecting measured brdfs to analytic brdfs by data-driven diffuse-specular separation. *ACM Trans. on Graphics (Proc. SIGGRAPH Asia)*, pp. 273:1–273:15.
- [298] SUN, T., JENSEN, H. W., AND RAMAMOORTHI, R. Connecting measured brdfs to analytic brdfs by data-driven diffuse-specular separation. *ACM Trans. on Graphics* 37, 6 (2018), 1–15.
- [299] SUN, T., SERRANO, A., GUTIERREZ, D., AND MASIA, B. Attribute-preserving gamut mapping of measured brdfs. *Computer Graphics Forum* 36, 4 (July 2017).
- [300] SUN, T., SERRANO, A., GUTIERREZ, D., AND MASIA, B. Attribute-preserving gamut mapping of measured brdfs. In *Computer Graphics Forum* (2017), vol. 36, pp. 47–54.

- [301] SUN, X., VILLEGAS, R., LAGUNAS, M., YANG, J., AND ZHANG, J. End-to-end relighting of a foreground object of an image, 2019. PCT international application number: 16/823,092.
- [302] SZEGEDY, C., VANHOUCKE, V., IOFFE, S., SHLENS, J., AND WOJNA, Z. Rethinking the inception architecture for computer vision. arxiv.
- [303] TAMUZ, O., LIU, C., BELONGIE, S., SHAMIR, O., AND KALAI, A. T. Adaptively learning the crowd kernel. In *Proc. International Conference on Machine Learning* (2011), pp. 673–680.
- [304] TANAKA, M., AND HORIUCHI, T. Investigating perceptual qualities of static surface appearance using real materials and displayed images. *Vision research* 115 (2015), 246–258.
- [305] TANI, Y., ARAKI, K., NAGAI, T., KOIDA, K., NAKAUCHI, S., AND KITAZAKI, M. Enhancement of glossiness perception by retinal-image motion: Additional effect of head-yoked motion parallax. *PloS one* 8, 1 (2013), e54549.
- [306] THIES, J., ZOLLHÖFER, M., AND NIESSNER, M. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. on Graphics* 38, 4 (2019), 1–12.
- [307] THOMPSON, W., FLEMING, R., CREEM-REGEHR, S., AND STEFANUCCI, J. K. *Visual Perception from a Computer Graphics Perspective*, 1st ed. A. K. Peters, Ltd., 2011.
- [308] THOMPSON, W., FLEMING, R., CREEM-REGEHR, S., AND STEFANUCCI, J. K. *Visual perception from a computer graphics perspective*. AK Peters/CRC Press, 2016.
- [309] TIEST, W. M. B. Tactual perception of material properties. *Vision Research* 50, 24 (2010), 2775–2782.
- [310] TODD, J. T., NORMAN, J. F., KOENDERINK, J. J., AND KAPPERS, A. M. Effects of texture, illumination, and surface reflectance on stereoscopic shape perception. *Perception* 26, 7 (1997), 807–822.
- [311] TORRESANI, L. *Weakly Supervised Learning*. Springer US, Boston, MA, 2014, pp. 883–885.
- [312] ULLMAN, S. *The interpretation of visual motion*. Massachusetts Inst of Technology Pr, 1979.
- [313] UNSPLASH.
- [314] UPCHURCH, P., SNAVELY, N., AND BALA, K. From A to Z: supervised transfer of style and content using deep neural network generators. *CoRR abs/1603.02003* (2016).
- [315] VAN DER MAATEN, L., AND HINTON, G. E. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research* 9 (2008), 2579–2605.
- [316] VAN DER MAATEN, L., AND WEINBERGER, K. Stochastic triplet embedding. In *IEEE International Workshop on Machine Learning for Signal Processing* (2012), pp. 1–6.

-
- [317] VAN LEEUWEN, T. Semiotics and iconography. *Handbook of visual analysis* (2001), 92–118.
- [318] VAN STRATEN, R. *An introduction to iconography: Symbols, allusions and meaning in the visual arts*. Routledge, 2012.
- [319] VANGORP, P., BARLA, P., AND FLEMING, R. W. The perception of hazy gloss. *Journal of Vision* 17, 5 (2017), 19–19.
- [320] VANGORP, P., LAURIJSSSEN, J., AND DUTRÉ, P. The influence of shape on the perception of material reflectance. *ACM Trans. on Graphics* 26, 3 (July 2007).
- [321] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., AND POLOSUKHIN, I. Attention is all you need. In *Advances in neural information processing systems* (2017), pp. 5998–6008.
- [322] VÁVRA, R., AND FILIP, J. Minimal sampling for effective acquisition of anisotropic brdfs. In *Computer Graphics Forum* (2016), vol. 35, pp. 299–309.
- [323] VIDAURRE, R., CASAS, D., GARCES, E., AND LOPEZ-MORENO, J. Brdf estimation of complex materials with nested learning. In *IEEE Winter Conference on Applications of Computer Vision* (2019).
- [324] VOULODIMOS, A., DOULAMIS, N., DOULAMIS, A., AND PROTOPAPADAKIS, E. Deep learning for computer vision: A brief review. 1–13.
- [325] WALTER, B., MARSCHNER, S. R., LI, H., AND TORRANCE, K. E. Microfacet models for refraction through rough surfaces. In *Eurographics Symposium on Rendering* (2007).
- [326] WANG, Q., WANG, Z., GENOVA, K., SRINIVASAN, P., ZHOU, H., BARON, J. T., MARTIN-BRUALLA, R., SNAVELY, N., AND FUNKHOUSER, T. Ibrnet: Learning multi-view image-based rendering. *arXiv preprint arXiv:2102.13090* (2021).
- [327] WANG, Z., BOVIK, A. C., SHEIKH, H. R., SIMONCELLI, E. P., ET AL. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.
- [328] WANG, Z., YU, X., LU, M., WANG, Q., QIAN, C., AND XU, F. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Trans. on Graphics* 39, 6 (2020), 1–13.
- [329] WEISS, Y. Deriving intrinsic images from image sequences. In *Proc. International Conference on Computer Vision* (2001).
- [330] WELINDER, P., BRANSON, S., PERONA, P., AND BELONGIE, S. J. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems* (2010), pp. 2424–2432.
- [331] WILLS, J., AGARWAL, S., KRIEGMAN, D., AND BELONGIE, S. Toward a perceptual space for gloss. *ACM Trans. on Graphics* 28, 4 (Sept. 2009), 103:1–103:15.
- [332] WU, P., HOI, S. C., XIA, H., ZHAO, P., WANG, D., AND MIAO, C. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM International Conference on Multimedia* (2013), pp. 153–162.

- [333] XIA, H., HOI, S. C. H., JIN, R., AND ZHAO, P. Online multiple kernel similarity learning for visual search. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 36, 3 (2014), 536–549.
- [334] XIAO, B., BI, W., JIA, X., WEI, H., AND ADELSON, E. H. Can you see what you feel? color and folding properties affect visual–tactile material discrimination of fabrics. *Journal of Vision* 16, 3 (2016), 34–34.
- [335] XIAO, B., ZHAO, S., GKIOULEKAS, I., BI, W., AND BALA, K. Effect of geometric sharpness on translucent material perception. *Journal of Vision* 20, 7 (2020), 10–10.
- [336] XU, B., WANG, N., CHEN, T., AND LI, M. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853* (2015).
- [337] XUEY, S., WANG, J., TONG, X., DAI, Q., AND GUO, B. Image-based material weathering. In *Computer Graphics Forum* (2008), vol. 27, pp. 617–626.
- [338] YE, G., GARCES, E., LIU, Y., DAI, Q., AND GUTIERREZ, D. Intrinsic video and applications. *ACM Trans. on Graphics* 33, 4 (2014).
- [339] YIN, W., SCHÜTZE, H., XIANG, B., AND ZHOU, B. ABCNN: attention-based convolutional neural network for modeling sentence pairs. *CoRR abs/1512.05193* (2015).
- [340] YOSINSKI, J., CLUNE, J., BENGIO, Y., AND LIPSON, H. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* (2014), pp. 3320–3328.
- [341] YU, Y., AND SMITH, W. A. Inverserendernet: Learning single image inverse rendering. In *Proc. Computer Vision and Pattern Recognition* (2019), pp. 3155–3164.
- [342] YUMER, M. E., CHAUDHURI, S., HODGINS, J. K., AND KARA, L. B. Semantic shape editing using deformation handles. *ACM Trans. on Graphics* 34, 4 (2015), 86.
- [343] ZANKER, J. M. Does motion perception follow weber’s law? *Perception* 24, 4 (1995), 363–372.
- [344] ZEILER, M. D., AND FERGUS, R. Visualizing and understanding convolutional networks. In *Proc. European Conference on Computer Vision* (2014), Springer, pp. 818–833.
- [345] ZELL, E., ALIAGA, C., JARABO, A., ZIBREK, K., GUTIERREZ, D., MCDONNELL, R., AND BOTSCH, M. To stylize or not to stylize?: the effect of shape and material stylization on the perception of computer-generated faces. *ACM Trans. on Graphics* 34, 6 (2015), 184.
- [346] ZHANG, D., AND LU, G. Shape-based image retrieval using generic fourier descriptor. *Signal Processing: Image Communication* 17, 10 (2002), 825–848.
- [347] ZHANG, F., DE RIDDER, H., AND PONT, S. The influence of lighting on visual perception of material qualities. In *Human Vision and Electronic Imaging* (2015), vol. 9394, p. 93940Q.

-
- [348] ZHANG, K., LUAN, F., WANG, Q., BALA, K., AND SNAVELY, N. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proc. Computer Vision and Pattern Recognition* (2021), pp. 5453–5462.
- [349] ZHANG, K., RIEGLER, G., SNAVELY, N., AND KOLTUN, V. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492* (2020).
- [350] ZHANG, R., ISOLA, P., EFROS, A. A., SHECHTMAN, E., AND WANG, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. Computer Vision and Pattern Recognition* (2018), pp. 586–595.
- [351] ZHANG, X., FANELLO, S., TSAI, Y.-T., SUN, T., XUE, T., PANDEY, R., ORTS-ESCOLANO, S., DAVIDSON, P., RHEMANN, C., DEBEVEC, P., ET AL. Neural light transport for relighting and view synthesis. *ACM Trans. on Graphics* 40, 1 (2021).
- [352] ZHANG, X., SRINIVASAN, P. P., DENG, B., DEBEVEC, P., FREEMAN, W. T., AND BARRON, J. T. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *arXiv preprint arXiv:2106.01970* (2021).
- [353] ZHAO, N., KIM, N. W., HERMAN, L. M., PFISTER, H., LAU, R. W., ECHEVARRIA, J., AND BYLINSKII, Z. Iconate: Automatic compound icon generation and ideation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–13.
- [354] ZHOU, H., HADAP, S., SUNKAVALLI, K., AND JACOBS, D. W. Deep single-image portrait relighting. In *Proc. International Conference on Computer Vision* (2019).
- [355] ZSOLNAI-FEHÉR, K., WONKA, P., AND WIMMER, M. Gaussian material synthesis. *ACM Trans. on Graphics* 37, 4 (July 2018), 76:1–76:14.
- [356] ZSOLNAI-FEHÉR, K., WONKA, P., AND WIMMER, M. Photorealistic material editing through direct image manipulation. In *Computer Graphics Forum* (2020), vol. 39, pp. 107–120.