Noname manuscript No.
(will be inserted by the editor)

Doc-Attentive-GAN: Attentive GAN for historical document denoising

Hala Neji · Mohamed Ben Halima · Javier Nogueras-Iso · Tarek M. Hamdani · Javier Lacasta · Habib Chabchoub · Adel M. Alimi

Received: date / Accepted: date

Abstract Image denoising attempts to restore images that have been degraded. Historical document denoising is specially challenging because there is considerable background noise or variation in contrast and illumination in handwritten literature and the first times of the printing press. The main objective of this work is to propose a new method for historical document denoising based on an Attentive Generative Adversarial Network (Attentive-GAN). Our proposed model for historical document denoising is named Doc-Attentive GAN , and it employs an attention map generated by a deep network to help the generator to learn and focus on the modification between the target image and its noisy version. It has been trained and tested with different historical document collections such as well-known DIBCO datasets, Arabic Historical Documents from the Tunisian National Library, and Incunabula books. The experiments demonstrate a clear improvement in the visual quality of the images obtained by Doc-Attentive-GAN with respect to the state-of-the-art.

H. Neji
University of Gabes, Tunisia
a E-mail: hala.neji@ieee.org

M. Ben Halima, T. M. Hamdani, A. M. Alimi
REsearch Groups in Intelligent Machines (REGIM Lab), University of Sfax, National Engineering School of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia

J. Nogueras-Iso, J. Lacasta
I3A, University of Zaragoza, Spain.

Habib Chabchoub
College of Business, Al Ain University, UAE

A. M. Alimi
Department of Electrical and Electronic Engineering Science, Faculty of Engineering and the Built Environment, University of Johannesburg, South Africa.

Keywords Historical document denoising · Document image denoising · Document image binarization · Generative Adversarial Networks · GAN.

1 Introduction

During the last two decades, different institutions or initiatives related to cultural heritage have launched preservation policies that have allowed the access to digitized versions of cultural material. For instance, only Europeana (open data library providing an access point to cultural access collections from different European Union member countries) and DPLA (Digital Public Library of America) provide access to more than 70 millions of digitized items [9].

An important part of the cultural heritage preserved by these repositories are digitized versions of analog-borne documents, i.e. digitized versions of historical manuscripts or any type of printing press with historical value. Logically, in order to provide users with an advanced searching and understanding of these documents, in addition to the scanned image of the original document, many of these repositories try to apply Optical Character Recognition (OCR) on these documents to facilitate also the text content of these documents.

However, the application of OCR to historical documents, even in the case of using printing press as a source, is especially complicated due to the features of the physical documents and the quality of the digitized page images [12]. With respect to the physical nature of historical documents in the first times of the printing press, Springman et al [30] enumerate as possible problems the use of mixed hand-made typefaces (including ligatures) together with varied concentration of ink, and the logical degradation of paper. The digitization mechanism, either direct scanning of original documents or the digitization of existent microfilms, may also introduce additional problems such as splicing, gutter shadow, skew, back-to-front interference, or frame/border noise [5,6, 19].

Therefore, in order to increase the accuracy of these OCR tools, it is highly relevant to pre-process the digitized documents. Apart from splitting digitized documents into page images, detecting page orientation and removing skew problems, the main focus of preprocessing is to remove as much background noise as possible. For instance, salt and pepper noise usually appears in binarized images obtained after a digitization process [33].

The objective of this paper is to propose a new method for removing background noise in the context of the digitization of historical documents. In particular, this work proposes to extend an Attentive Generative Adversarial Network (Attentive-GAN) [26] for an improved restoration of text images as an image-to-image translation task. This proposed model is called Doc-Attentive-GAN and employs an attention map generated by a deep network to help the generator to learn and focus on the modification between the target image and its noisy version. The use of attention maps makes the generator more confident to generate correct images. In addition, it helps the model to

be trained faster and have a model with different types of motion and blurry images. The main advantage of Attentive-GAN is the correction rate between the target images and the generated images.

The remainder of the paper is organized as follows. Section 2 provides a review of the related work. Section 3 presents our Doc-Attentive-GAN proposal. Section 4 shows the results of the experiments using different variants of the proposed method and various datasets for training and testing. Finally, the paper ends with some concluding remarks and and an outline for futures lines of research.

2 Related work

Historical documents suffer from several types of degradation that affect their readability and increase in size during scanning. This section compiles works related to document cleaning, i.e. removal of background noise originated by the bad condition of original documents (stains, degradation of paper) or the digitization process (gutter shadow, back to front interference).

Initial approaches were based on the segmentation of degraded historical document images in order to find subregions where thresholding algorithms could be applied to separate written text from the background [4]. Using a similar approch, Yagoubi et al.[37] proposed a method joining compression and enhancement of single-side handwritten document images. This approach presents a foreground/background segmentation algorithm, using both directional and contrast features to highlight the original information. This pretreatment step is embedded into a DjVu encoder, which is commonly used in national archives and libraries frameworks to drive the compression rate. Bag et al.[1] also proposed a new technique for restoration of faint characters in degraded documents using an adaptive–interpolative thresholding technique for binarization with stroke preservation.

Applying other types of techniques, Sun et al. [32] proposed a method based on conditional random fields (CRF) to remove the bleed-through from the scanned images of historical images. Ntirogiannis et al. [23] addressed a pixel-based binarization evaluation methodology for historical handwritten/machine-printed document images. Nafchi et al. [21] proposed a binarization model for ancient document images which consists of three standard steps: preprocessing, main binarization, and postprocessing. Hedjam et al. [15] introduced a new approach for historical document image restoration using a multispectral imaging system.

Neural networks have been also applied for denoising. For instance, Namane et al.[22] suggested a two-subnets neural network called complementary similarity measure method (CSM) for degraded printed character optical recognition. One of the subnets is a similarity layer which is based on CSM. The other one is a competitive neural network based on the winner takes all algorithm (WTA), which is used for the classification.

In recent years several works have been focused on the applicability of Generative Adversarial Networks (GAN) for image restoration and document binarization. Regarding restoration, Cao et al. [3] proposed a novel fast GAN model for complex masked image restoration that includes a neighbouring network, a generator network, a discriminator network, and two parsing networks. Cao et al. [2] also studied a novel method for images restoration of damaged ancient murals using a consistency-enhanced generative adversarial network. Similarly, Khamekhem et al. [18] proposed an end-to-end architecture based on Generative Adversarial Networks (GAN) to recover the degraded documents into a clean and readable form. As far as document binarization is concerned, Zhao et al. [38] proposed a method for document image binarization with cascaded generators of conditional generative adversarial networks. The generator contains two sub-generators called G1 and G2. G1 extracts text pixels from an input image and G2 learns a combination of results at different scales from the first sub-generator and produces the final binary result. De et al. [7] proposed a new approach for document image binarization using dual discriminator Generative Adversarial Network (DD-GAN) which uses Focal Loss as generator loss. The model consists of two discriminator networks: one looks for the global similarity (i.e. on the whole image), and another one explores the image in small patches (i.e. local similarity). In addition, Souibgui and Kessentini [29] have used conditional GANs (cGANs) to restore severely degraded document images. They call their system DEGAN (Document Enhancement conditional Generative Adversarial Network). They have tested DEGAN for two main purposes: document cleaning and binarization (1); and water mark removal (2). Last, it must be noted that Souibgui et al. [28] also proposed a Text-Degradation Invariant Auto Encoder (Text-DIAE), a self-supervised model designed to tackle two tasks, text recognition (handwritten or scene-text) and document image enhancement.

In summary, several methods have been proposed to deal with image degradation using GAN-based approaches. As already indicated in the introduction, in this work we propose the adaptation of Attentive-GAN [26] for document denoising injecting an attention map into both generative and discriminative networks to improve the quality of document restoration. Next section describes our proposed methodology based on Attentive-GAN for text image processing.

3 Methodology

The purpose of a document enhancement process is to recover a clean image from a degraded one as an image-to-image translation task using a generative adversarial network. In this paper we study the performance of Attentive GAN for the denoising challenge. The main idea attempts to extend Attentive GAN for an improved restoration of text images called Doc-Attentive-GAN. The goal of Doc-Attentive-GAN is to use an attention map generated by a deep network which is transmitted to the generator. The main idea of an attention

map is to help the generator to learn and focus on the modification between the target image and its noisy version. The information existing in the attention maps guides the generator to generate an effective version from a noisy image and obtain a sharp version.

The main elements of our Doc-Attentive-GAN model are: an attentive-recurrent network, a contextual autoencoder, and a fully connected layer. This section explains the adaptation of Attentive GAN [26] for the reconstruction of historical documents, where our main contribution lies in the definition of the appropriate number of attention maps (see Table 2 in Section 4.2) and the loss functions part of the networks. Section 3.1 describes the generative network containing the attentive-recurrent network and the contextual autoencoder. Section 3.2 completes the description of the fully connected network explaining the discriminative part. Generally, the generative outputs should be evaluated by the discriminator network to be sure that our outputs look like the real images.

3.1 Generative Network

Usually, the role of the generator is to generate an image from an image. However, with Attentive GAN, the generator will focus in the first step to produce an attention map from input images containing noise. The main idea to use an attention map is to guide the generator to focus on the existing noise to be removed from the image. In the second step, a skip connection network will be used to produce the final outputs that will be directed to the discriminator. The architecture of the generator is shown in Figure 1. Its design consists of two subnets: an attentive-recurrent network and a contextual autoencoder.

3.1.1 Attentive-recurrent network

The attentive recurrent network is used to find regions in the input image that need to get attention. These regions are necessary for the contextual autoencoder and the discriminative network. On the one hand, the autoencoder can generate better the local image restoration. On the other hand, the discriminator can focus the assessment on specific regions.

Therefore, the first subnetwork employs a recurrent network to generate visual attention. It consists of five layers of ResNet, a convolutional LSTM unit and convolutional layers for generating the 2D attention maps. The attention map is a matrix ranging from 0 to 1, increased with each time step. The values of the attention map are initialized to 0.5 in the training process and for each step the current attention map is concatenated with the input image and then they are fed into the next block of the recurrent network. The loss function in each recurrent block is defined as the mean squared error (MSE) between the output attention map and the binary mask.

Fig. 1: Generator architecture design used in this study.

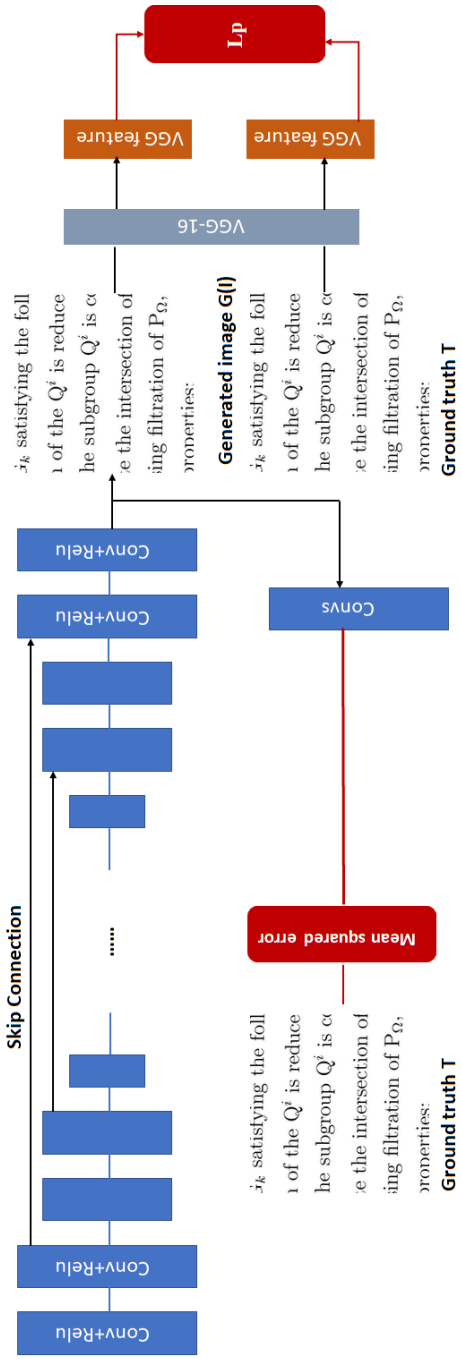


Fig. 2: Architecture of the contextual autoencoder.

3.1.2 Contextual autoencoder

The second subnetwork is a contextual autoencoder used to generate a sharp image (see Figure 2). The input of this network is the concatenation between the final attention map from the attentive-recurrent network and the input image. It consists of 16 Conv-ReLU blocks, and skip connections are added to prevent blurred outputs. There are two loss functions in the autoencoder: perceptual loss and multi-scale loss.

The perceptual loss measures the global discrepancy between the features of the autoencoder output and those of the corresponding ground-truth clean image. We use a VGG16 pretrained on the ImageNet dataset [8] to extract these features. The perceptual loss function is defined as:

$$L_P(O,T) = L_{MSE}(VGG(O), VGG(T)) \tag{1}$$

where O is the output image and T is the ground-truth image.

With respect to the multi-scale loss, the original function in Attentive-GAN is a multi-scale mean squared error loss that is computed based on a pixel-by-pixel operation with the features extracted from different decoder layers to form outputs in different sizes. The loss function is expressed as:

$$L_M(\{S\},\{T\}) = \sum_{i=1}^{M} \lambda_i L_{MSE}(S_i, T_i) \tag{2}$$

where S_i is the i th output extracted from the 1^{st}, 3^{rd} and 5^{th} of the decoder whose sizes are $1/4$, $1/2$ and 1 of the original size, respectively. T_i is the ground truth. S_i and T_i has the same scale i. $\{\lambda_i\}_{i=1}^{M}$ are the weights for different scales. The value of λ is fixed to 0.6, 0.8, 1.0.

As an alternative to this multi-scale loss function, we developed first a mean squared error noise loss function for the last output of the convolutional layer. Therefore, we drop the two skip connection between the loss function and the output scales 2 and 4. Our mean squared noise loss function is expressed as:

$$L_N(I,T,O) = L_{MSE}(DIFF(O-T), DIFF(I-T)) \tag{3}$$

where L_N is the noise loss, I is the input image, O is the output, T is the ground truth and $DIFF$ is the subtract operation.

Second, we also investigated the potential use of a multi-scale mean squared error noise loss. It is defined as:

$$L_N(\{I\},\{T\},\{S\}) = \sum_{i=1}^{M} \lambda_i L_{MSE}(DIFF(S_i - T_i), DIFF(I_i - T_i)) \tag{4}$$

where $\{\lambda_i\}_{i=1}^{M}$ are the weights for different scales.

3.2 Discriminative Network

The discriminative network consists of 7 convolution layers with the kernel of (3, 3), a fully connected layer of 1024 and a single neuron with a sigmoid activation function. The output is a 2D matrix containing probabilities of the generated image being real. This matrix contains probabilities that should be, to the discriminator, close to 1 if the clean image represents the ground truth and 0 if the image is generated by the generator.

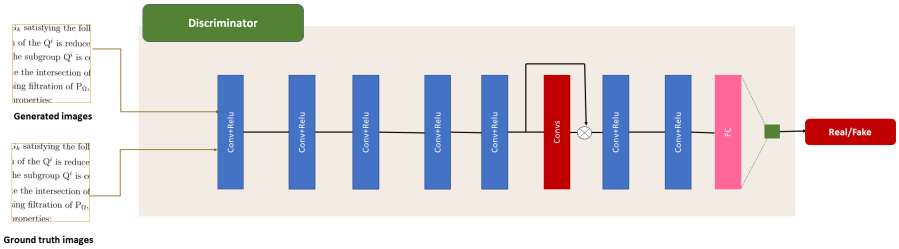


Fig. 3: Discriminator architecture design used in this study

The model is presented in Figure 3. The discriminator receives two input images which are the generated image and its ground truth version. The whole loss function of the discriminator can be expressed as:

$$L_D(O, R, A_N) = -log(D(R)) - log(1 - D(O)) + \gamma L_{map}(O, R, A_N) \quad (5)$$

where L_{map} is the loss between the features extracted from interior layers of the discriminator and the final attention map:

$$L_{map}(O, R, A_N) = L_{MSE}(D_{map}(O), A_N) + L_{MSE}(D_{map}(R), 0) \quad (6)$$

where γ is set to 0.05, L_{map} represents the process of producing a 2D map by the discriminative network. 0 represents a map containing only 0 values and R is a sample image drawn from a pool of real and clean images.

4 Experimental results

4.1 Datasets and environment configuration

We used five different datasets for the purpose of our experiments. Figure 4 shows a representative example of the training and testing datasets . A brief description of these datasets is as follows:

Fig. 4: Example of the training and testing datasets.

(a) DIBCO dataset
This dataset refers to the datasets provided in the annual editions of a competition called Document Image Binarization Contest (DIBCO), organized in conjunction with ICDAR (International Conference on Document Analysis and Recognition) and ICFHR (International Conference on Frontiers in Handwriting Recognition) conferences.[1]
The documents of these datasets originate from collections of historical documents in well-known national libraries. In each edition, the conference organizers select images that contain representative degradations which appear frequently (e.g. variable background intensity, shadows, smear, smudge, low contrast, bleed-through and show-through) [10]. Table 1 presents the number of documents on each DIBCO dataset. For the purpose of our experiments, we have used different combinations of the datasets for training and testing.

(b) Incunabula dataset
This dataset consists of different book page images obtained from the incunabula collection available at the digital repository of the University of Zaragoza, called Zaguan. This is a remarkable collection containing 384 different incunabula items with 413 digitized documents (some books have several digitized exemplars), which were originated by 165 different printing offices in Europe. From the set of Zaguan incunabula documents we

[1] https://vc.ee.duth.gr/dibco2019/

Table 1: The number of documents on each DIBCO dataset.

Data set	Handwritten	Printed	Total
DIBCO'09	5	5	10
H-DIBCO'10	10	-	10
DIBCO'11	8	8	16
H-DIBCO'12	14	-	14
DIBCO'13	8	8	16
H-DIBCO'14	10	-	10
H-DIBCO'16	10	-	10
DIBCO'17	20	-	20
H-DIBCO'18	10	-	10
Total	95	21	116

have selected one page of 58 distinct books to create a test dataset that contains text images with real noise originated by either the digitalization process or the physical degradation of the paper.

(c) Salt and pepper dataset
In order to create a synthetic dataset with salt and pepper noise, we have used the Typenrepertorium der Wiegendrucke [13], a comprehensive catalogue of font types and printing offices in the incunabula period that is accessible in electronic format through a web site [31]. This catalogue includes sample pages of the books printed by different printers and their particular font types. From this dataset of sample pages, we chose 109 pages. First, we up-sampled by 3 all the pages. Then, we added a salt and pepper noise with 0.07 factor. Finally, we cut these pages into images of size 528x264. As a result, we have obtained a database which contains 1000 pairs of images. Although the initial objective of this dataset was to train and test models for removing salt and pepper noise, it was also useful for building models that can denoise Arabic historical documents.

(d) TuLiPa dataset
This dataset was produced as a result of the TuLiPa project, whose objective is to provide a platform for automatic restoration, dating and translation of Arabic historical documents preserved in the National Library of Tunisia. The most common degradation problems found in images of ancient Arabic documents are the effect of transparency, the presence of stains of moisture absorbed by the paper, variations in the color of the paper, the presence of wrinkles and tears, and deformations due to the natural curvature of the pages. In order to create a TuLiPa dataset corpus of Arabic historical documents for the purpose of this work, we chose 137 pages from the National Library of Tunisia. We use this dataset only for the testing part.

(e) Noisy Office dataset
The Kaggle platform includes a competition on denoising dirty documents.[2] The data associated with this competition contain 144 pairs of images

[2] https://www.kaggle.com/c/denoising-dirty-documents/

(noisy image and clean image) with modern printed text, which can be used for training and testing.

With respect to the environment configuration, it must be noted that all the computation works were executed on an Ubuntu server with an NVIDIA Quadro P6000 GPU. In addition, we have created a Fighshare repository[3] with the datasets (used as input or generated as output) and pre-trained models in the experiments reported at this section. The code is based on the software provided with Attentive-GAN.[4] Our contribution relies on the use of new loss functions, the identification of the most appropriate number of attention maps, and the selection of the datasets for the training of each scenario of historical document denoising.

4.2 Denoising of historical documents

We started with the binarization of DIBCO datasets. Our contribution is focused on the selection of the loss function to be used by the contextual autoencoder and the number of attention maps required by the attentive-recurrent network in order to generate results with appropriate visual quality.

First, we trained the network using DIBCO 2009, 2010, 2012, 2013, 2014, 2016, 2017, 2018. For this training, we used 6746 pairs of images (patches of size 256 × 256). Then, we evaluated the performance of the binarization in our method using DIBCO 2011 as the testing dataset using two metrics: the Peak Signal to Noise Ratio (PSNR)[35] and the Structural Similarity Index (SSIM)[36]. As presented in Table 2, we tried to change the number of attention maps and the loss function. The best PSNR value was 18.66 and SSIM= 0.91 with four attention maps in the attentive recurrent network and the multiscale mean squared error loss as the autoencoder loss function. We also compared our result with the state of the art (see Table 3) and our result is competitive with the state of the art.

Table 2: Comparative results using different loss functions and different numbers of attention maps with Doc-Attentive-GAN and DIBCO 2011 dataset

	Multiscale MSE		MSE noise		MSE	
Attention map number	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
2	17.81	0.89	16.77	0.87	17.66	0.88
3	17.65	0.88	17.13	0.87	17.91	0.89
4	**18.66**	**0.91**	18.17	0.89	18.35	0.90
5	17.63	0.88	17.89	0.89	17.79	0.88
6	17.84	0.89	18.20	0.89	17.45	0.88

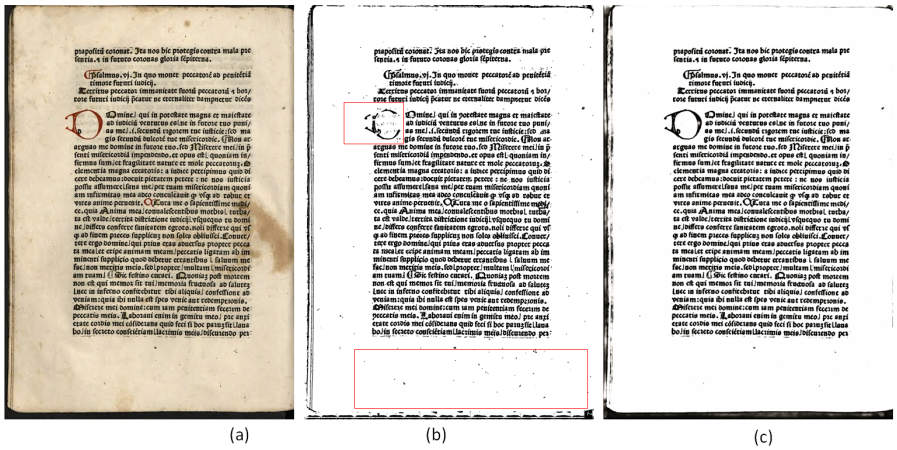
[3] https://figshare.com/s/d2d7155f42ed6fa12c7e

[4] https://github.com/MaybeShewill-CV/attentive-gan-derainnet

Table 3: Comparative results on H-DIBCO 2011.

Method	PSNR
Otsu et al. [25]	15.7
Sauvola et al. [27]	15.6
Howe et al. [16]	19.30
Vo et al. [34]	20.10
Guo et al. [11]	16.50
He et al.[14]	19.9
Zhao et al.[38]	20.3
Competition winner	16.1
Kang et al.[17]	19.9
Doc-Attentive-GAN	18.66

In addition, we tested this model (trained with DIBCO datasets) with the incunabula dataset. The results were good in qualitative comparisons with the DEGAN method[29]. As shown in Figure 5, our method clearly improves the quality of image specially in the details of the page margins and the letters. However, our model did not work very well for Arabic samples. As shown in Figure 6, although it clearly cleans the background, it deletes some parts of the page. Notwithstanding this, our result is better than DEGAN [29].



(a) (b) (c)

Fig. 5: Qualitative binarization results produced with some examples of the Incunabula dataset, comparison between Doc-Attentive-GAN and DEGAN. (a) original image (b) DEGAN (c) Doc-Attentive-GAN

Second, we used DIBCO 2014 for testing and we trained the model with the rest of DIBCO datasets. We selected 5734 pairs of images for training and 1013 for the validation. The model was trained with the same configuration as the previous model: we just changed the loss function derived for the autoencoder network to compare the effect of the loss function. As presented in Table 4,

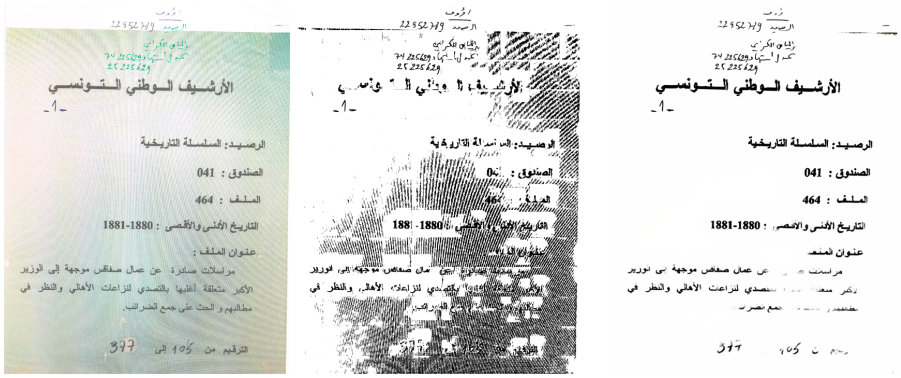


Fig. 6: Qualitative binarization results produced with some examples of the TuLiPa dataset, comparison between Doc-Attentive-GAN and DEGAN. (a) original image (b) DEGAN (c) Doc-Attentive-GAN

the best loss function used in the autoencoder was the MSE noise loss : the obtained PSNR and SSIM values were 19.90 and 0.92 respectively.

Table 4: Comparative results using Attentive GAN, DIBCO 2014 dataset for testing, 4 attention maps, and different loss functions

Loss	PSNR	SSIM
Multiscale Mean squared error	18.43	0.90
Mean squared error	18.16	0.90
Mean squared error noise	**19.90**	**0.92**
Multiscale Mean squared error noise	19.28	0.91

We also compared our result with the state of the art. As presented in Table 5, our result is competitive. Our method removes the background degradation from different images and visually we can deduce the capability to improve the quality of image as shown in Figure 7, where we present the comparative results of the output images generated using different loss functions (the figure also includes the degraded input and the ground truth image).

Then, we trained our Doc-Attentive-GAN using four attention map, the multiscale loss function and different combinations of the DIBCO datasets for training and testing. For each DIBCO dataset mentioned in Table 6, we used it for testing the model and the remaining DIBCO datasets were used for training the model. Table 6 compares the results of our model with the different DIBCO competition winners. Figure 8 shows and illustrative comparison between our output and the output obtained with DEGAN [29].

Last, in order to create a model for denoising historical Arabic manuscripts we have used our salt and pepper dataset. We trained the model using four

Table 5: Comparative results on H-DIBCO 2014.

Method	**PSNR**
Otsu et al. [25]	18.72
Sauvola et al. [27]	17.63
Howe et al. [16]	19.3
Vo et al. [34]	23.23
Guo et al. [11]	19.17
He et al.[14]	22.1
Zhao et al.[38]	22.12
Competition winner [24]	22.66
Kang et al.[17]	22.37
Doc-Attentive-GAN	19.90

Table 6: Comparison of the proposed methods with (H)DIBCO competition winners.

Database	Method	PSNR
DIBCO'09	Best Competition System	18.66
	Doc-Attentive-GAN	18.24
H-DIBCO'11	Best Competition System	17.97
	Doc-Attentive-GAN	18.66
H-DIBCO'12	Best Competition System	21.80
	Doc-Attentive-GAN	19.71
DIBCO'13	Best Competition System	21.29
	Doc-Attentive-GAN	19.55
H-DIBCO'14	Best Competition System	22.66
	Doc-Attentive-GAN	19.56
H-DIBCO'16	Best Competition System	18.45
	Doc-Attentive-GAN	19.22
DIBCO'17	Best Competition System	18.28
	Doc-Attentive-GAN	17.60
H-DIBCO'18	Best Competition System	19.11
	Doc-Attentive-GAN	15.57

attentive maps, the multi-scale loss function, and our salt and pepper synthetic dataset. We used 850 pairs of images for training, 50 for validation and 100 for testing. This model performed well with the TuLiPa dataset. In addition, we compared qualitatively our result with DEGAN [29] and our pre-trained models using different configurations. As a result, the best output was obtained using the model trained with the salt and pepper dataset because the noise of the documents is very similar to the salt and pepper noise. Figure 9 shows the output obtained with the Doc-Attentive-GAN model (trained with the salt and pepper dataset), which improves the quality of the Arabic manuscript and behaves better than the rest of the methods.

4.3 Denoising of modern printed text images

In this part we focused on the noisy office dataset. We trained Doc-Attentive-GAN with multi-scale noise loss and four attention maps. We used 112 images for training and 32 images for testing. In addition, we compared quantitatively our method with DEGAN. The results are shown in Table 7. Doc-Attentive-

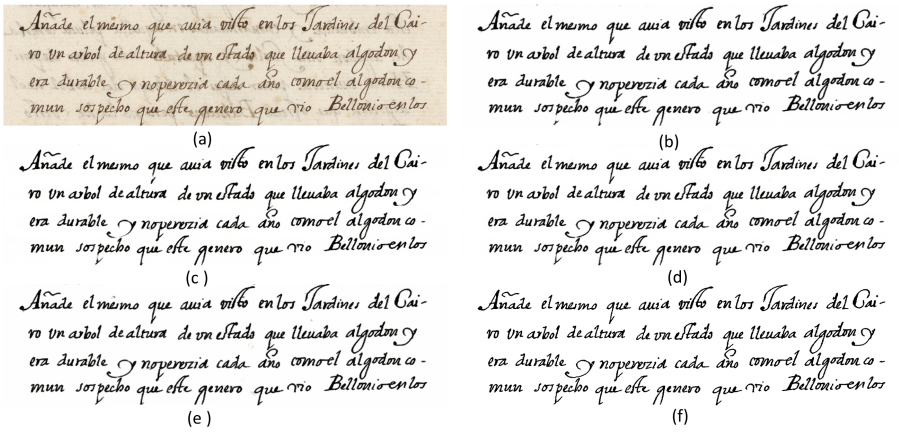


Fig. 7: Comparative results of output images using different loss functions. (a) original image (b) output image with multiscale MSE loss (c) output image with MSE loss (d) output image with multiscale MSE noise loss (e) output image with MSE noise loss (f) ground truth

GAN obtained a PSNR that equals to 32.18db and according to SSIM, we achieved 0.99 on average.

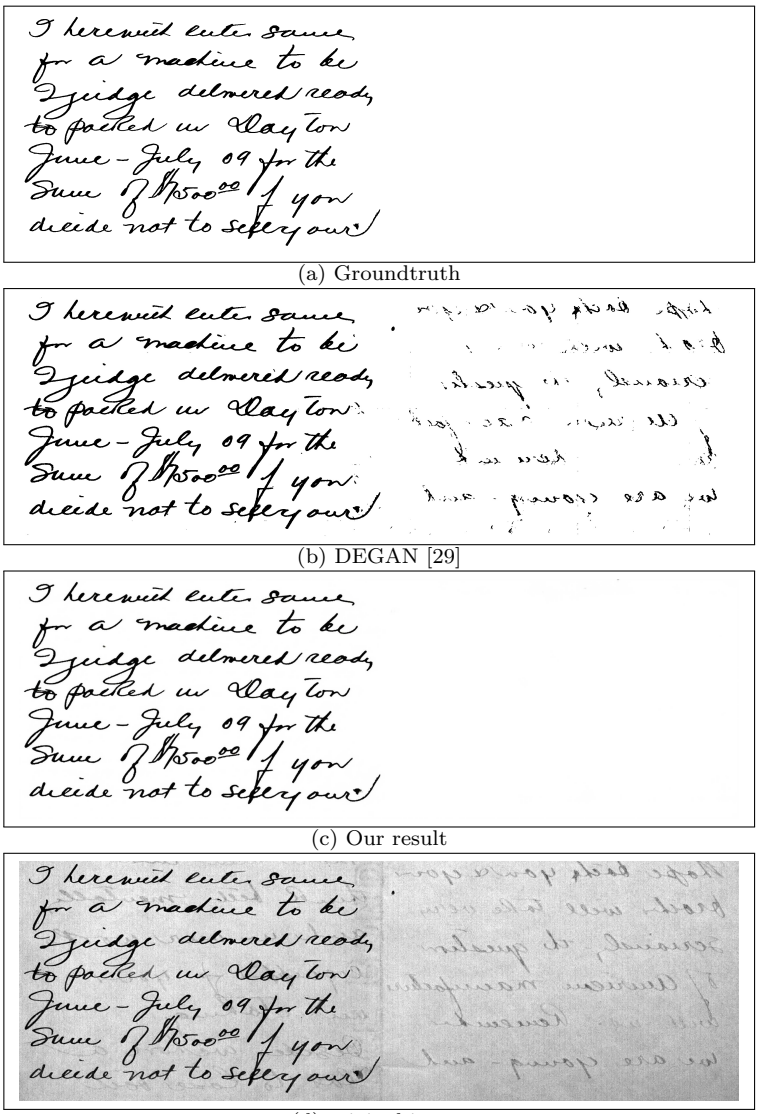
Table 7: Quantitative results of the denoising task (noisy office dataset)

Method	PSNR	SSIM
DEGAN [29]	38.12	0.99
Doc-Attentive-GAN	32.18	0.99

Figure 10 shows the qualitative results of the denoising task. We demonstrate that Attentive-GAN succeeded to reconstruct with high resolution. Although the average PSNR is lower in Table 7, the output image may have a higher quality. This is due to the fact that the PSNR metric measures the sensitivity to errors (norm of the arithmetic difference between the reference and the output image). Despite our Doc-Attentive-GAN model may not completely remove the grey background color in Figure 10, the obtained letters are sharper than those obtained by DEGAN.

5 Conclusions

The problem of historical document enhancement is complex. On the one hand, historical documents present very different types of degradations. On the other hand, there are few datasets available to be used for training or testing and containing as well a ground truth version of these documents. The purpose of this paper has been to study different alternatives to remove the noise and



(a) Groundtruth

(b) DEGAN [29]

(c) Our result

(d) original image

Fig. 8: Qualitative comparison of our model with DEGAN using a DIBCO 2013 example.

generate a clean version of documents with readable letters to improve the performance of OCR processes.

In this work we have proposed the use of an Attentive GAN architecture with new loss functions and different numbers of attention maps. In particular, we have integrated a Mean Squared Error (MSE) noise loss and multi-scale MSE noise loss. With respect to the obtained results, we can conclude

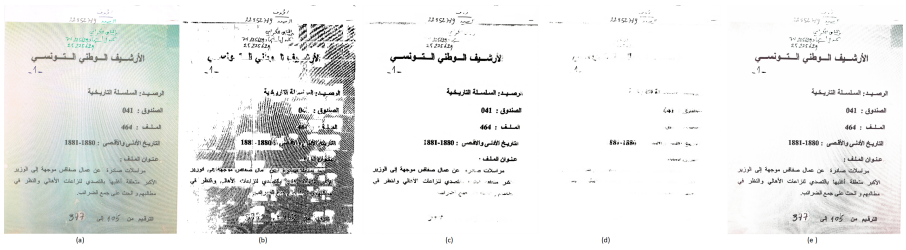
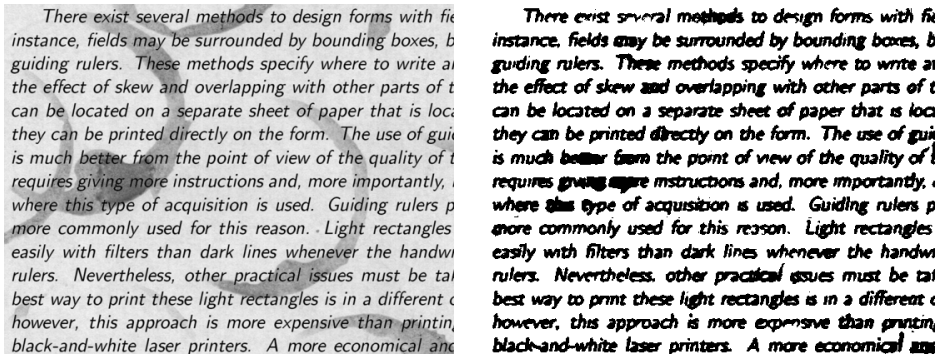


Fig. 9: Qualitative binarization results produced with an Arabic sample in TuLiPa dataset, comparison between Doc-Attentive-GAN pre-trained models and DEGAN. (a) original image (b) DEGAN [29] (c) Doc-Attentive-GAN model trained with all DIBCO datasets (d) Doc-Attentive-GAN model trained with the noisy office dataset (e) Doc-Attentive-GAN model trained with the salt and pepper dataset



(a) original image

(b) DEGAN [29]

(c) Our result

(d) Groundtruth

Fig. 10: Qualitative comparison of our model with state-of-the-art methods

that our models succeeded to restore historical documents independently of the types of noise. We have demonstrated that GAN-based designs of neural networks can help to solve the problems related to the degradation of histor-

ical document images. Specially, we had the chance to improve the quality of the output generated after applying our model to degraded images available in DIBCO competitions and Arabic historical document collections, which is a big challenge in the recent years.

In the future we want to investigate whether alternative combinations of GAN based networks can provide a better performance for historical document denoising than Doc-Attentive-GAN. For instance, we want to evaluate if the pyramid of GAN networks proposed by Ma et al. [20] could be applied to the case of document denoising.

Acknowledgements We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Quadro P6000 GPU used for this research.

Funding The research leading to these results has been partially supported by the Ministry of Higher Education and Scientific Research of Tunisia under the grant agreement number LR11ES48, the Spanish Regional Government of Aragon (project T59_23R), and the Spanish Ministry of Science and Innovation (project PID2020-113353RB-I00).

Declarations

Conflicts of interests The authors declare that they have no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data availability The datasets (used as input or generated as output) and pretrained models involved in the experiments reported in section 4 are available in the Figshare repository `https://figshare.com/s/d2d7155f42ed6fa12c7e`.

References

1. Bag, S., Bhowmick, P.: Adaptive–interpolative binarization with stroke preservation for restoration of faint characters in degraded documents. Journal of Visual Communication and Image Representation **31**, 266–281 (2015)
2. Cao, J., Zhang, Z., Zhao, A., Cui, H., Zhang, Q.: Ancient mural restoration based on a modified generative adversarial network. Heritage Science **8**(1), 7 (2020)
3. Cao, Z., Niu, S., Zhang, J., Wang, X.: Fast generative adversarial networks model for masked image restoration. IET Image Processing **13**(7), 1124–1129 (2019)
4. Chen, Y., Leedham, G.: Decompose algorithm for thresholding degraded historical document images. IEE Proceedings-Vision, Image and Signal Processing **152**(6), 702–714 (2005)
5. Conway, P., Chapman, S., Kenney, A.R.: Digital imaging and preservation microfilming: The future of the hybrid approach for the preservation of brittle books (1999)
6. Dale, R.L.: Rlg guidelines for microfilming to support digitization (2003)
7. De, R., Chakraborty, A., Sarkar, R.: Document image binarization using dual discriminator generative adversarial networks. IEEE Signal Processing Letters **27**, 1090–1094 (2020)

8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255. Ieee (2009)
9. Díaz-Corona, D., Lacasta, J., Latre, M.Á., Zarazaga-Soria, F.J., Nogueras-Iso, J.: Profiling of knowledge organisation systems for the annotation of linked data cultural resources. Information Systems **84**, 17–28 (2019)
10. Gatos, B., Ntirogiannis, K., Pratikakis, I.: Icdar 2009 document image binarization contest (dibco 2009). In: 2009 10th International conference on document analysis and recognition, pp. 1375–1382. IEEE (2009)
11. Guo, J., He, C., Zhang, X.: Nonlinear edge-preserving diffusion with adaptive source for document images binarization. Applied Mathematics and Computation **351**, 8–22 (2019)
12. Gupta, A., Gutierrez-Osuna, R., Christy, M., Capitanu, B., Auvil, L., Grumbach, L., Furuta, R., Mandell, L.: Automatic assessment of ocr quality in historical documents. pp. 1735–1741 (2015)
13. Haebler, K.: Typenrepertorium der Wiegendrucke, vol. 5 vols. Verlag von Rudolf Haupt, Leipzig (1905-1924)
14. He, S., Schomaker, L.: Deepotsu: Document enhancement and binarization using iterative deep learning. Pattern recognition **91**, 379–390 (2019)
15. Hedjam, R., Cheriet, M.: Historical document image restoration using multispectral imaging system. Pattern Recognition **46**(8), 2297–2312 (2013)
16. Howe, N.R.: Document binarization with automatic parameter tuning. International journal on document analysis and recognition (ijdar) **16**(3), 247–258 (2013)
17. Kang, S., Iwana, B.K., Uchida, S.: Complex image processing with less data—document image binarization by integrating multiple pre-trained u-net modules. Pattern Recognition **109**, 107577 (2021)
18. Khamekhem Jemni, S., Souibgui, M.A., Kessentini, Y., Fornés, A.: Enhance to read better: A multi-task adversarial network for handwritten document image enhancement. Pattern Recognition **123**, 108370 (2022)
19. Lins, R.D., Banergee, S., Thielo, M.: Automatically detecting and classifying noises in document images. In: Proceedings of the 2010 ACM Symposium on Applied Computing, pp. 33–39 (2010)
20. Ma, R., Zhang, B., Hu, H.: Gaussian pyramid of conditional generative adversarial network for real-world noisy image denoising. Neural Processing Letters **51**, 2669–2684 (2020)
21. Nafchi, H.Z., Moghaddam, R.F., Cheriet, M.: Phase-based binarization of ancient document images: Model and applications. IEEE transactions on image processing **23**(7), 2916–2930 (2014)
22. Namane, A., Guessoum, A., Soubari, E.H., Meyrueis, P.: Csm neural network for degraded printed character optical recognition. Journal of visual communication and image representation **25**(5), 1171–1186 (2014)
23. Ntirogiannis, K., Gatos, B., Pratikakis, I.: Performance evaluation methodology for historical document image binarization. IEEE Transactions on Image Processing **22**(2), 595–609 (2012)
24. Ntirogiannis, K., Gatos, B., Pratikakis, I.: Icfhr2014 competition on handwritten document image binarization (h-dibco 2014). In: 2014 14th International conference on frontiers in handwriting recognition, pp. 809–813. IEEE (2014)
25. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)
26. Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J.: Attentive generative adversarial network for raindrop removal from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2482–2491 (2018)
27. Sauvola, J., Pietikäinen, M.: Adaptive document image binarization. Pattern recognition **33**(2), 225–236 (2000)
28. Souibgui, M.A., Biswas, S., Mafla, A., Biten, A.F., Fornés, A., Kessentini, Y., Lladós, J., Gomez, L., Karatzas, D.: Text-diae: A self-supervised degradation invariant autoencoder for text recognition and document enhancement. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 2330–2338 (2023)

29. Souibgui, M.A., Kessentini, Y.: De-gan: A conditional generative adversarial network for document enhancement. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
30. Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A., Fink, F.: Ocr of historical printings of latin texts: problems, prospects, progress. In: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, pp. 71–75 (2014)
31. Staatbibliothek zu Berlin: Tw - typenrepertorium der wiegendrucke (2020)
32. Sun, B., Li, S., Zhang, X.P., Sun, J.: Blind bleed-through removal for scanned historical document image with conditional random fields. IEEE Transactions on Image Processing **25**(12), 5702–5712 (2016)
33. Tan, C.L., Liu, Q.H.: Extraction of newspaper headlines from microfilm for automatic indexing. Document Analysis and Recognition **6**(3), 201–210 (2003)
34. Vo, Q.N., Kim, S.H., Yang, H.J., Lee, G.: Binarization of degraded document images based on hierarchical deep supervised network. Pattern Recognition **74**, 568–586 (2018)
35. Wang, Z., Bovik, A.C.: Mean squared error: Love it or leave it? a new look at signal fidelity measures. IEEE signal processing magazine **26**(1), 98–117 (2009)
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
37. Yagoubi, M.R., Serir, A., Beghdadi, A.: Joint enhancement-compression of handwritten document images through djvu encoder. Journal of Visual Communication and Image Representation **41**, 324–338 (2016)
38. Zhao, J., Shi, C., Jia, F., Wang, Y., Xiao, B.: Document image binarization with cascaded generators of conditional generative adversarial networks. Pattern Recognition **96**, 106968 (2019)