Anexos

Anexo A

Funcionamiento de PVNet

PVNet [16] es una red neuronal convolucional con la que se obtiene la *pose* real de un objeto a partir de una única imagen RGB (imagen en color), que puede estar sometida a oclusiones o truncamientos.

Dada una imagen RGB, el objetivo de la estimación de *pose* es detectar el objeto en la imagen, junto con su rotación y traslación en 3D. En el caso de PVNet, se trata de una transformación rígida (matriz de rotación-traslación) desde el sistema de coordenadas del objeto hasta el sistema de coordenadas de la cámara.

A.1. Arquitectura de PVNet

La red neuronal convolucional PVNet está basada en una red ResNet18 [9] preentrenada y con pequeñas modificaciones. Asumiendo que existen C clases de objetos a detectar y K *keypoints* para cada clase, los datos de entrada de PVNet son un vector de dimensiones $H \times W \times 3$, correspondiente a una imagen RGB estándar.

La imagen de entrada a la red se procesa por medio de una arquitectura puramente convolucional, y se obtienen como *output* dos tensores, de dimensiones $H \times W \times (K \times 2 \times C)$, correspondiente a los vectores unitarios, y $H \times W \times (C + 1)$, correspondiente a la máscara del objeto.

A.2. Estimación de la pose

PVNet divide la estimación de la *pose* en dos etapas: detección de los puntos (denominados *keypoints*) con la red neuronal y estimación de la matriz rotación-traslación (*pose*) mediante el algoritmo PnP [12]. En la Figura A.1 se representa la arquitectura de la red neuronal, además de las fases para la obtención de la *pose* del objeto.

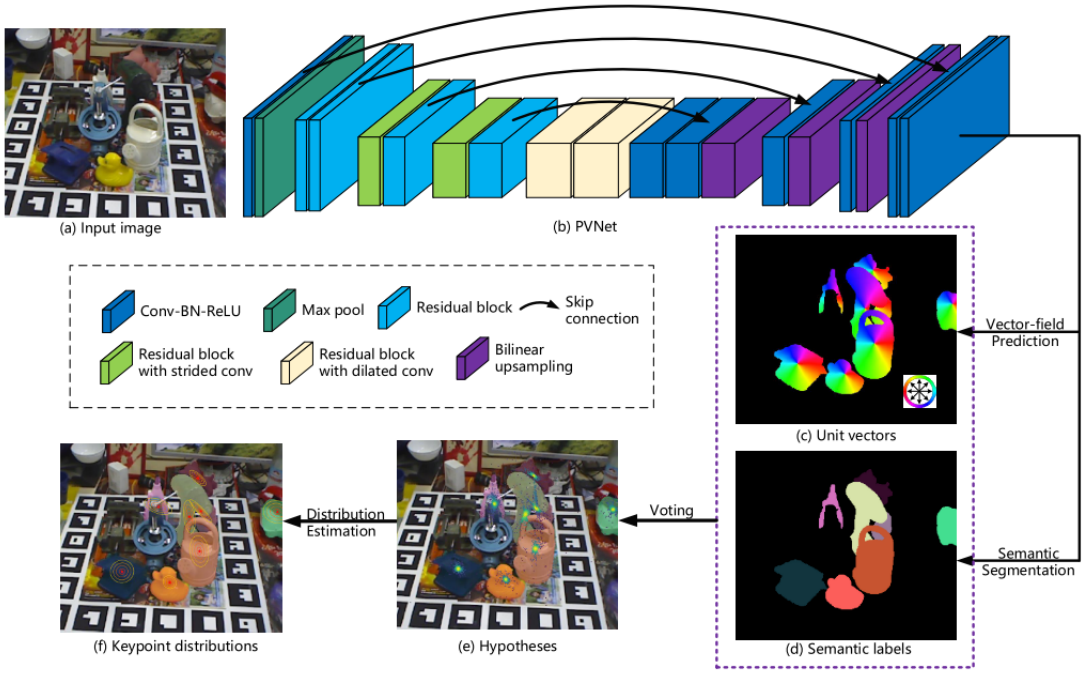

Figura A.1: Arquitectura de la red PVNet [16] y fases para la obtención de la *pose* del objeto.

A.2.1. Estimación de los keypoints

Como se puede observar en la Figura A.1, los datos de salida de la red neuronal únicamente son la segmentación o máscara del objeto, y los vectores unitarios, que representan la dirección desde cada píxel que forma el objeto hacia cada uno de los *keypoints*.

Posteriormente, dadas las direcciones desde cada píxel hacia cada *keypoint*, se generan hipótesis de la localización 2D de cada uno de los *keypoints*, además de sus intervalos de confianza, obtenidos a partir del algoritmo RANSAC [15]. Basándose en los intervalos de confianza, se estima la media y la covarianza de la probabilidad de la distribución espacial para cada *keypoint*.

La estimación de los *keypoints* a partir de los intervalos de confianza permite que la red se centre más en las características locales de los objetos, y disminuye la influencia de una escena saturada sobre la detección del objeto. El uso de un campo vectorial permite también la estimación de *keypoints* que estén ocluidos o truncados.

A.2.2. Estimación de la matriz rotación-traslación

Los *keypoints* estimados por la red neuronal corresponden a 8 puntos de la malla 3D del objeto, generados mediante el algoritmo *Farthest Point Sampling* (FPS) [27] de forma aleatoria antes de comenzar a entrenar la red neuronal. Además, a este conjunto

de *keypoints* se le añade el centroide del objeto.

En total, el número de puntos estimados por la red neuronal es de 9. Conociendo su posición verdadera sobre el modelo 3D, y mediante el algoritmo *Perspective-n-Point* (PnP), se puede estimar la matriz de rotación-traslación del objeto, obteniendo la *pose* verdadera del mismo.

La matriz de rotación-traslación del objeto corresponde a la matriz de transformación entre el sistema de referencia local de la cámara y el del objeto, y para la nomenclatura usada en este proyecto, se denomina $^{\mathbf{C}}\mathbf{T}_{\mathbf{O}}$.

Anexo B

Validación del modelo

En este anexo se describe el proceso de validación del modelo llevado a cabo tras el primer entrenamiento.

Para la validación del modelo, se ha seguido el método de validación cruzada [37], y se ha utilizado el mismo conjunto de datos que para el primer entrenamiento: 2666 capturas del hígado con fondo homogéneo tomadas desde todas las perspectivas posibles. Las condiciones de entrenamiento también son las mismas que para el primer entrenamiento, utilizando un tamaño de lote de 5, y realizando 40 repeticiones sobre cada entrenamiento de forma independiente. La tasa de aprendizaje parte de un valor de 0,001, disminuyendo un 50 % cada 5 repeticiones.

El objetivo de realizar una validación cruzada del modelo es comprobar si se ha producido un sobreentrenamiento del modelo (es decir, que las estimaciones se ajustan demasiado a los datos de entrenamiento, pero el modelo es incapaz de generalizar resultados).

Se ha elegido el método de validación cruzada porque el tiempo de entrenamiento del modelo no es muy alto (5 horas), y el conjunto de datos es relativamente pequeño (2666 imágenes).

Conforme a lo definido en la Sección 3.1, la proporción de imágenes utilizadas para el entrenamiento del conjunto de datos actual es del 80 %, mientras que la proporción de imágenes utilizadas para el test es del 20 %. Por tanto, se utilizan 4 veces más imágenes para el entrenamiento que para el test.

Para realizar la validación cruzada del modelo, se han creado 5 subdivisiones del conjunto de datos (533 imágenes por cada subdivisión), y se han realizado 5 entrenamientos independientes, utilizando como test cada una de las 5 subdivisiones en los 5 entrenamientos, y las otras 4 partes como datos de entrenamiento. Para cada entrenamiento, se ha partido del mismo punto, que corresponde al modelo preentrenado del gato. En la Figura B.1 se representa gráficamente el proceso de validación cruzada utilizado.

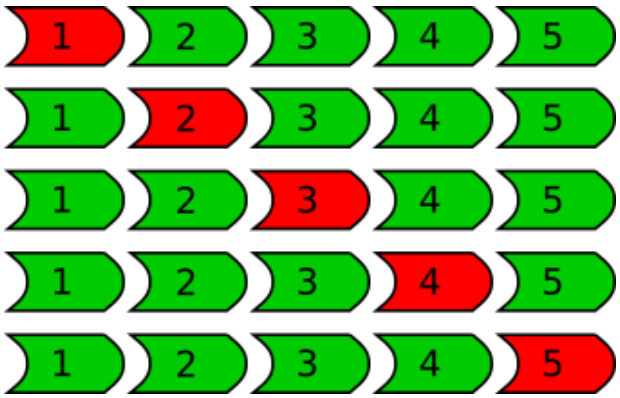

Figura B.1: Validación cruzada del modelo. El conjunto de datos se divide en 5 subconjuntos de igual tamaño, y se entrena 5 veces distintas de forma independiente. En verde, se representan las subdivisiones del conjunto de datos que se utilizan para el entrenamiento en cada uno de los entrenamientos independientes. En rojo, se representan las subdivisiones del conjunto utilizadas como test para cada uno de los 5 entrenamientos independientes.

En la Tabla B.1 se exponen los resultados de la validación cruzada realizada sobre los datos de entrenamiento. Como se puede observar, los resultados son muy homogéneos entre subdivisiones (*splits*, en inglés). Además, la estimación de la máscara del objeto (*segmentation loss*) produce un error menor que en el caso de la estimación de los vértices (*vertex loss*).

	Train			
	Precision %	**Recall %**	**Seg. Loss %**	**Ver. Loss %**
Split 1	99,955	99,978	0,008583	0,075
Split 2	99,964	99,978	0,008779	0,102
Split 3	99,957	99,983	0,008812	0,087
Split 4	99,959	99,981	0,008	0,071
Split 5	99,959	99,983	0,008213	0,089

Tabla B.1: Resultados de la validación cruzada en los datos de entrenamiento para los 5 entrenamientos realizados.

En la Tabla B.2 se exponen los resultados de la validación cruzada sobre los datos de test. Como se puede observar, los resultados son muy homogéneos, y con valores muy similares a los datos de entrenamiento. Por tanto, se puede afirmar que no se ha producido sobreestimación con el primer entrenamiento realizado, ya que los resultados para el entrenamiento y el test son muy similares. En el caso del test, la estimación de los vértices (*vertex loss*) es algo peor que en el caso de entrenamiento, pero los valores se pueden considerar aceptables, con un error por debajo del $0,5\,\%$.

Por último, se han calculado los errores en la estimación de la traslación y la rotación del hígado, conforme a la métrica expuesta en el Anexo C. En la Tabla B.3 se exponen

	Test			
	Precision %	**Recall %**	**Seg. Loss %**	**Ver. Loss %**
Split 1	99,96	99,977	0,008748	0,237
Split 2	99,961	99,98	0,008333	0,275
Split 3	99,958	99,98	0,008687	0,236
Split 4	99,959	99,98	0,008271	0,255
Split 5	99,959	99,979	0,008409	0,24

Tabla B.2: Resultados de la validación cruzada en los datos de test para los 5 entrenamientos realizados.

los resultados en traslación y rotación del sistema de referencia calculado con respecto a la pose verdadera. Para obtener estos resultados, se ha calculado la media y la varianza de la traslación y la rotación para los datos del test, una vez obtenida la función de estimación en el entrenamiento. Como se puede observar para los 5 entrenamientos, los datos en traslación y rotación son muy homogéneos, aunque se estima con más precisión la traslación que la rotación, donde el error es considerable (25 º) pero mejorable con entrenamientos posteriores, como los que se realizan en la Sección 3.2.

	Pose estimation			
	$\mathbf{d_{RMS}}(cm)$	$\theta_{\mathbf{RMS}}(^o)$	$\mathbf{d_{s^2}}$	$\theta_{\mathbf{s^2}}$
Split 1	0,16822	25,8429	0,02679	631,15905
Split 2	0,49979	26,73338	0,24598	677,71824
Split 3	0,20582	23,00947	0,04108	503,86028
Split 4	0,15595	23,05769	0,02306	504,34913
Split 5	0,13877	22,61551	0,01837	487,83626

Tabla B.3: Resultados de la validación cruzada en la estimación de la pose para los 5 entrenamientos realizados.

En conclusión, tras realizar la validación cruzada del modelo se puede afirmar que no se ha producido un sobreentrenamiento, y que el conjunto de datos elaborado para el primer entrenamiento es correcto y válido para entrenar a la red neuronal.

Anexo C

Métrica de evaluación de la red neuronal

Se ha creado una nueva métrica para la evaluación de las estimaciones de pose realizadas por la red neuronal, comparando el error en la estimación de las poses predichas con la red neuronal con respecto a las poses de referencia contenidas en el conjunto de datos.

Se ha creado una métrica para la traslación y otra para la rotación. No se han considerado las métricas de estimación de pose que utiliza la red PVNet, puesto que esta nueva métrica aporta una mayor información, y se puede detectar con más facilidad dónde falla la red, así como el origen de los errores.

C.1. Métrica de la traslación

Para cuantificar el error de la traslación en la pose obtenida por la red, se calcula la distancia euclídea (en centímetros) del vector que une los dos orígenes de los dos sistemas de referencia (pose estimada y pose de referencia). Siendo ${}^{\mathbf{C}}\mathbf{T}_{\mathbf{O,n}}$ la pose obtenida a partir de la red neuronal, y ${}^{\mathbf{C}}\mathbf{T}_{\mathbf{O,r}}$ la pose verdadera para una posición determinada del hígado:

$$
\begin{aligned}
{}^{\mathbf{C}}\mathbf{T}_{\mathbf{O,n}} &= \begin{pmatrix} \mathbf{R}_{\mathbf{n},(\mathbf{3x3})} & \mathbf{p}_{\mathbf{n},(\mathbf{1x3})} \\ \mathbf{0}_{(\mathbf{3x1})} & 1 \end{pmatrix} \\
{}^{\mathbf{C}}\mathbf{T}_{\mathbf{O,r}} &= \begin{pmatrix} \mathbf{R}_{\mathbf{r},(\mathbf{3x3})} & \mathbf{p}_{\mathbf{r},(\mathbf{1x3})} \\ \mathbf{0}_{(\mathbf{3x1})} & 1 \end{pmatrix}
\end{aligned}
\tag{C.1}
$$

Donde $\mathbf{R_n}$ y $\mathbf{R_r}$ representan las matrices de rotación de la pose obtenida a partir de la red neuronal y de la pose verdadera, respectivamente, de dimensiones 3x3. $\mathbf{p_n}$ y $\mathbf{p_r}$ representan el vector de traslación de la pose obtenida con la red neuronal y de la verdadera pose, respectivamente, de dimensiones 3x1.

Los vectores de traslación están formados por las tres componentes XYZ que determinan su posición en el espacio:

$$\mathbf{p_n} = \begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix}$$
$$\mathbf{p_r} = \begin{pmatrix} x_r \\ y_r \\ z_r \end{pmatrix} \tag{C.2}$$

Por tanto, la distancia euclídea entre los dos sistemas de referencia, conocidas las componentes XYZ de cada uno de ellos será:

$$d_p = \sqrt{(x_n - x_r)^2 + (y_n - y_r)^2 + (z_n - z_r)^2} \tag{C.3}$$

C.2. Métrica de la rotación

La métrica elegida para la estimación del error en la rotación es el cálculo del ángulo (en grados) que forman los dos sistemas de referencia. Para ello, hay que calcular el producto escalar de las matrices de rotación de la pose estimada por la red y de la pose verdadera, respectivamente. En el caso de dos matrices, el producto escalar se define como:

$$\mathbf{R_n} \cdot \mathbf{R_r} = tr(\mathbf{R_n^T}.\mathbf{R_r}) \tag{C.4}$$

Partiendo de la definición de producto escalar y de C.4, se puede obtener el ángulo que forman dos sistemas de referencia θ_p como:

$$\theta_p = arccos\left(\frac{tr(\mathbf{R_n^T}.\mathbf{R_r}) - 1}{2}\right) \tag{C.5}$$

C.3. Media y varianza de los resultados

Lo expuesto en las secciones anteriores corresponde a la comparación uno a uno de las distintas poses estimadas con la red con respecto a las poses verdaderas. Para poder tener una visión general de las estimaciones de pose realizadas en el test, es necesario calcular la media y la varianza del conjunto de errores.

La media cuadrática para la traslación d_{RMS} y para la rotación θ_{RMS} se definen como:

$$d_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_{pi}^2}$$
$$\theta_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \theta_{pi}^2} \tag{C.6}$$

Donde N es el número de muestras, d_{pi} el error en la traslación para cada una de las poses, y θ_{pi} el error en la rotación para cada una de las poses. En este caso, será el 20 % de todas las muestras que contiene el conjunto de datos, y que se han asignado al test.

La varianza para la traslación d_{s^2} y para la rotación θ_{s^2} se definen como:

$$d_{s^2} = \frac{1}{N}\sum_{i=1}^{N}(d_{pi} - \overline{d_p})^2$$
$$\theta_{s^2} = \frac{1}{N}\sum_{i=1}^{N}(\theta_{pi} - \overline{\theta_p})^2 \tag{C.7}$$

Donde $\overline{d_p}$ es la media aritmética del error en la traslación para todas las muestras, y $\overline{\theta_p}$ es la media aritmética del error en la rotación para todas las muestras.