Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

Sistemas Urbanos Inteligentes

Segmentación Semántica

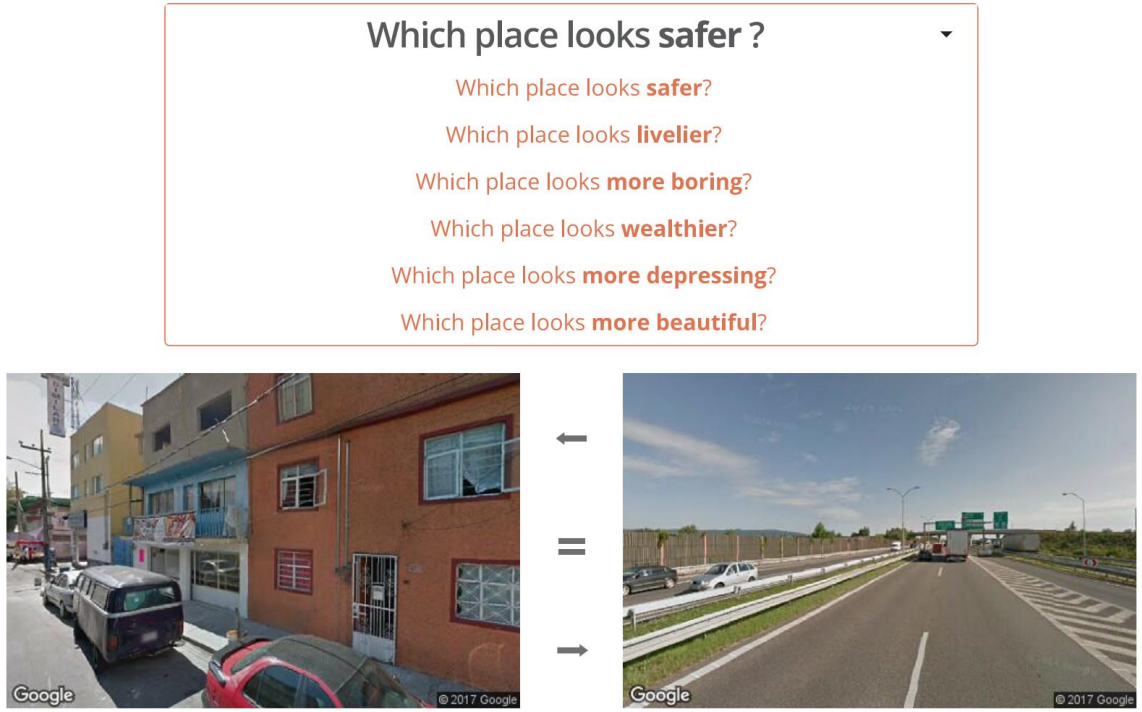
Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación

¿Qué utilidad nos pueden entregar las CNN en contextos urbanos?

¿Por qué nos gustaría cuantificar la percepción visual?

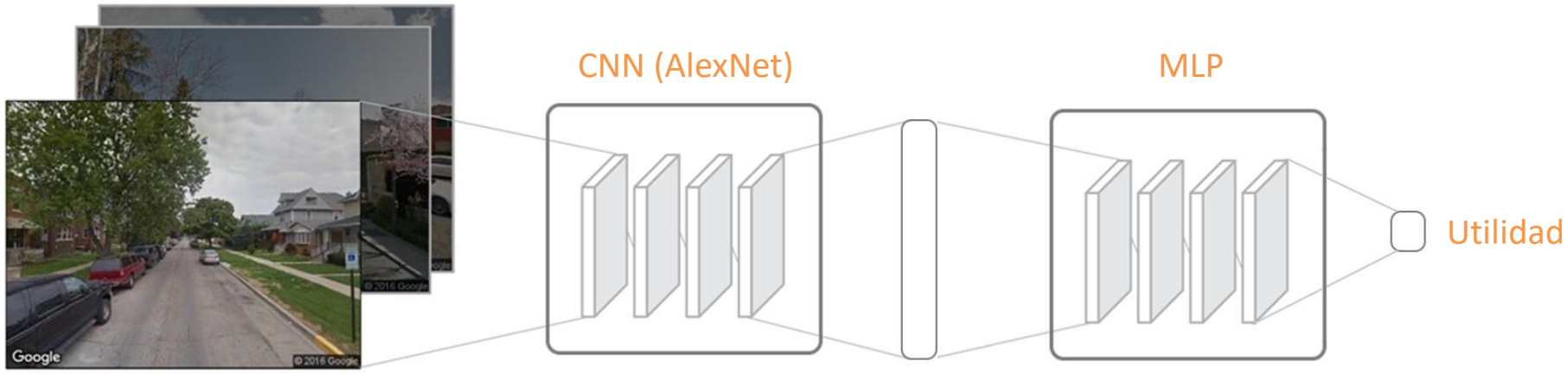
- El entorno urbano es percibido principalmente de forma visual.
- Esta percepción puede influenciar su intensidad de uso.
- Puede fomentar el uso del transporte público.
- Cuantificar esta percepción a escala nos permitiría identificar lugares candidatos para intervención.

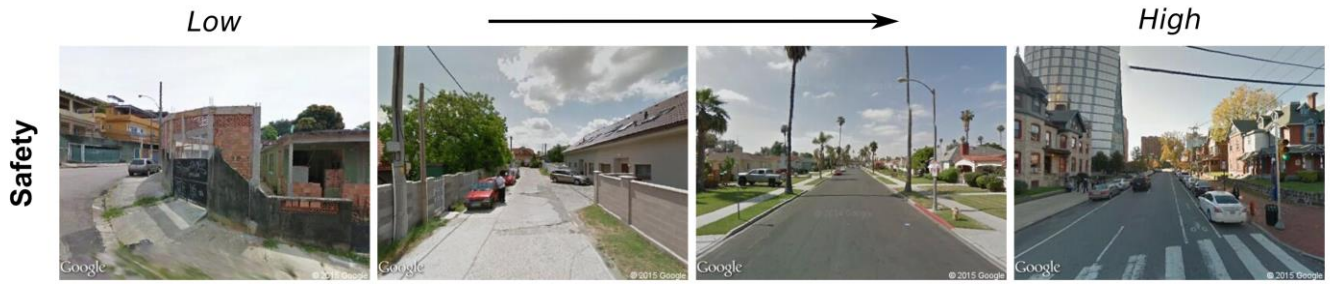
¿Cómo medir la percepción?



MIT Place Pulse data collection platform - http://pulse.media.mit.edu/

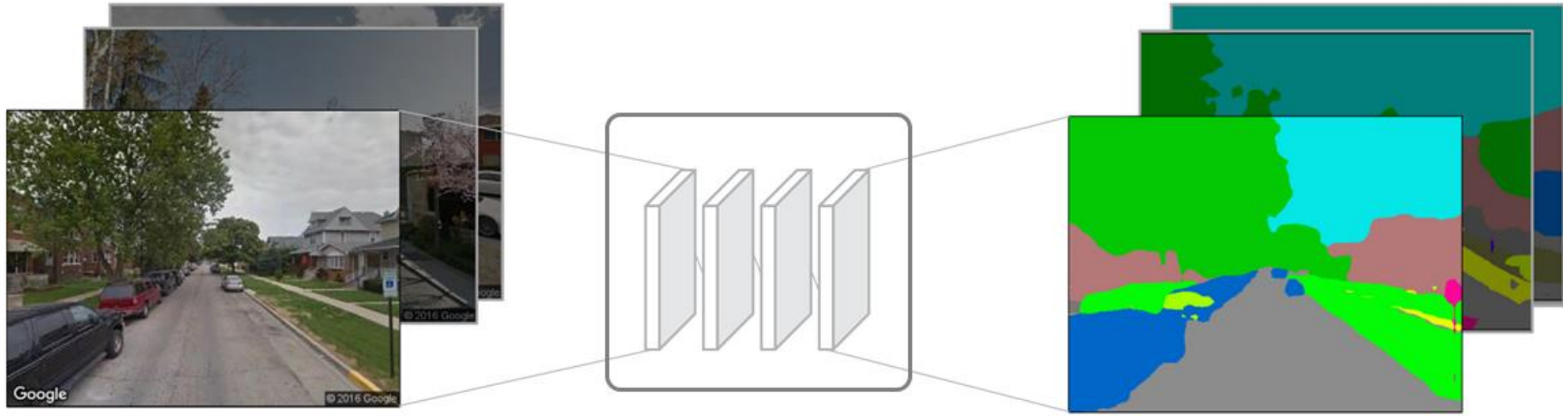
¿Cómo medir la percepción?





Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C. (2016). Deep Learning the City: Quantifying Urban Perception at a Global Scale. European Conference on Computer Vision (ECCV)

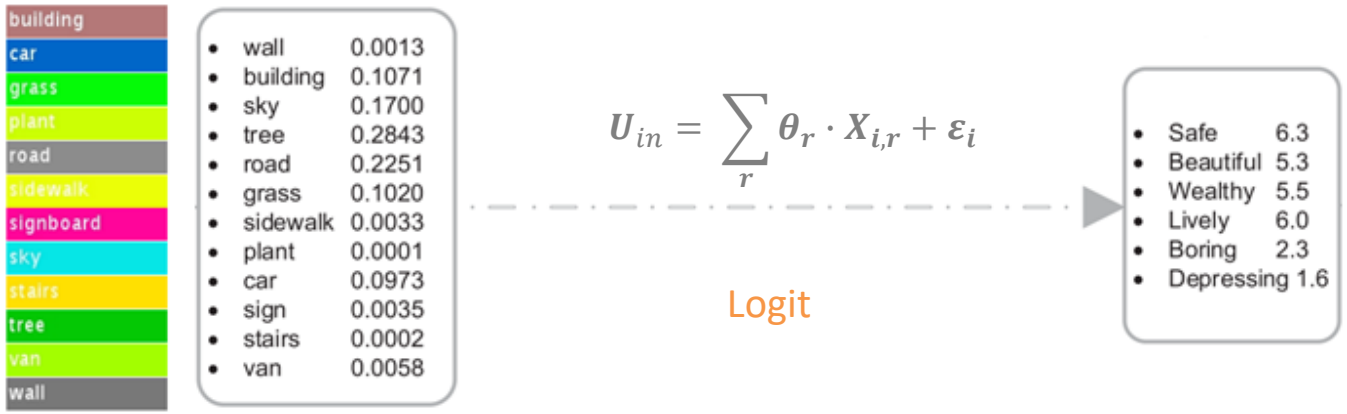
¿Cómo explicar la percepción?



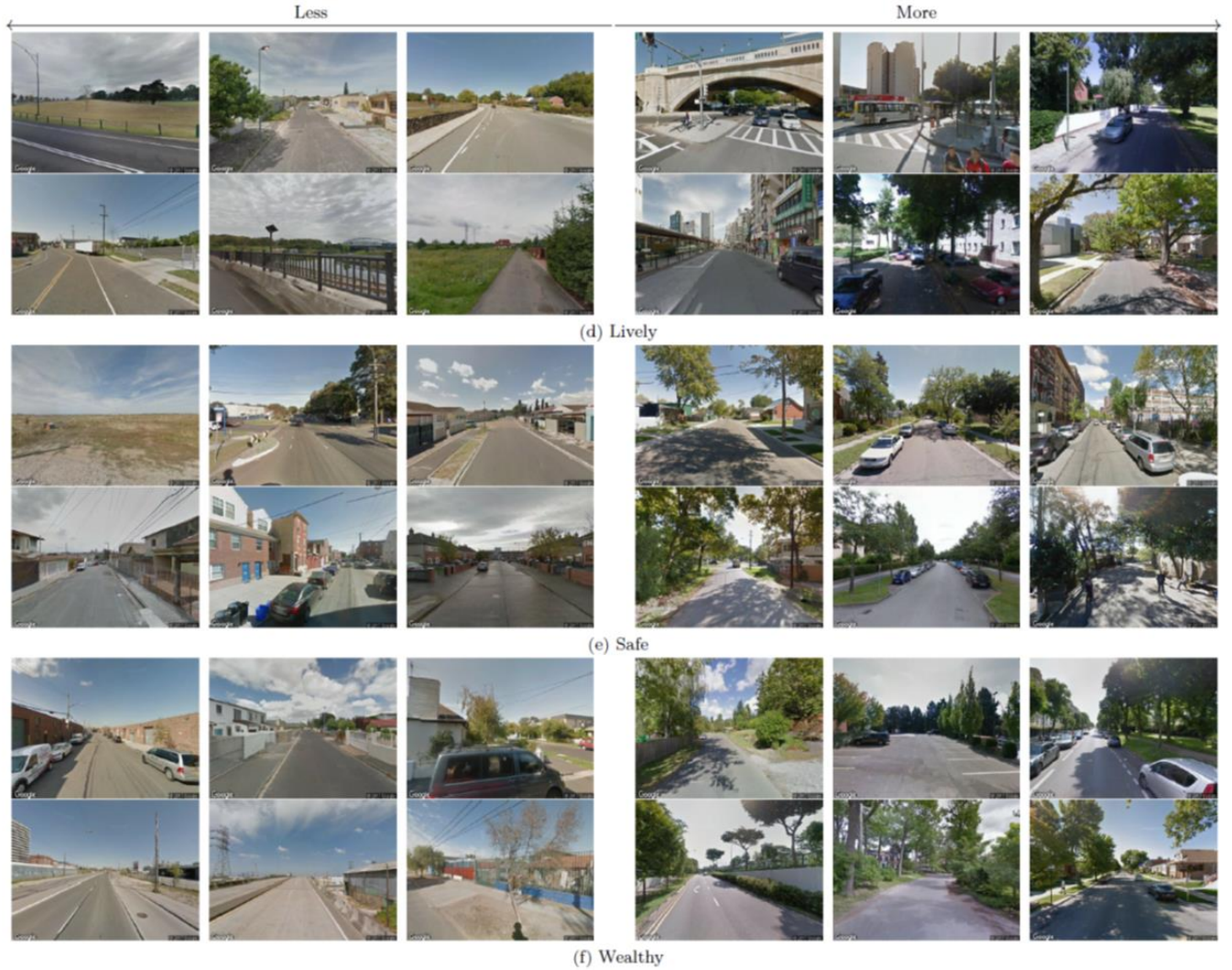
Segmentación semántica de imágenes

Rossetti, T., Lobel, H., Rocco, V., Hurtubia, R. (2019). Explaining subjective perceptions of public spaces as a function of the built environment: A massive data approach. Landscape & Urban Planning, 181, 169-178

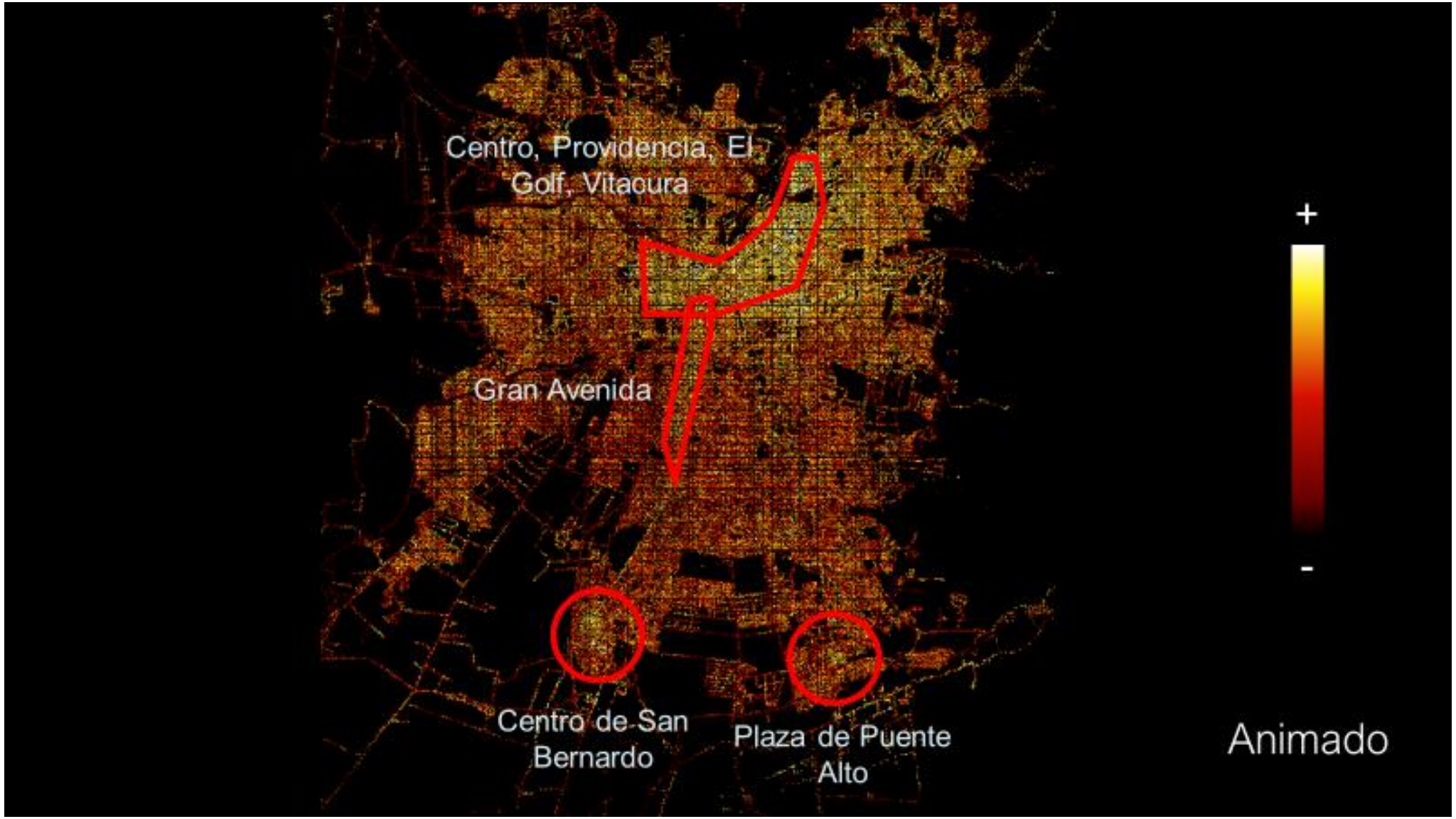
Atributos ahora son semánticos



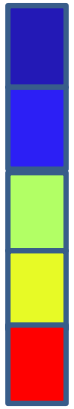
Rossetti, T., Lobel, H., Rocco, V., Hurtubia, R. (2019). Explaining subjective perceptions of public spaces as a function of the built environment: A massive data approach. Landscape & Urban Planning, 181, 169-178



Rossetti, T., Lobel, H., Rocco, V., Hurtubia, R. (2019). Explaining subjective perceptions of public spaces as a function of the built environment: A massive data approach. Landscape & Urban Planning, 181, 169-178


Centro, Providencia, El Golf, Vitacura
Gran Avenida
Centro de San Bernardo
Plaza de Puente Alto
+
-
Animado

¿Cuál es la tarea asociada a la segmentación semántica?



person

grass

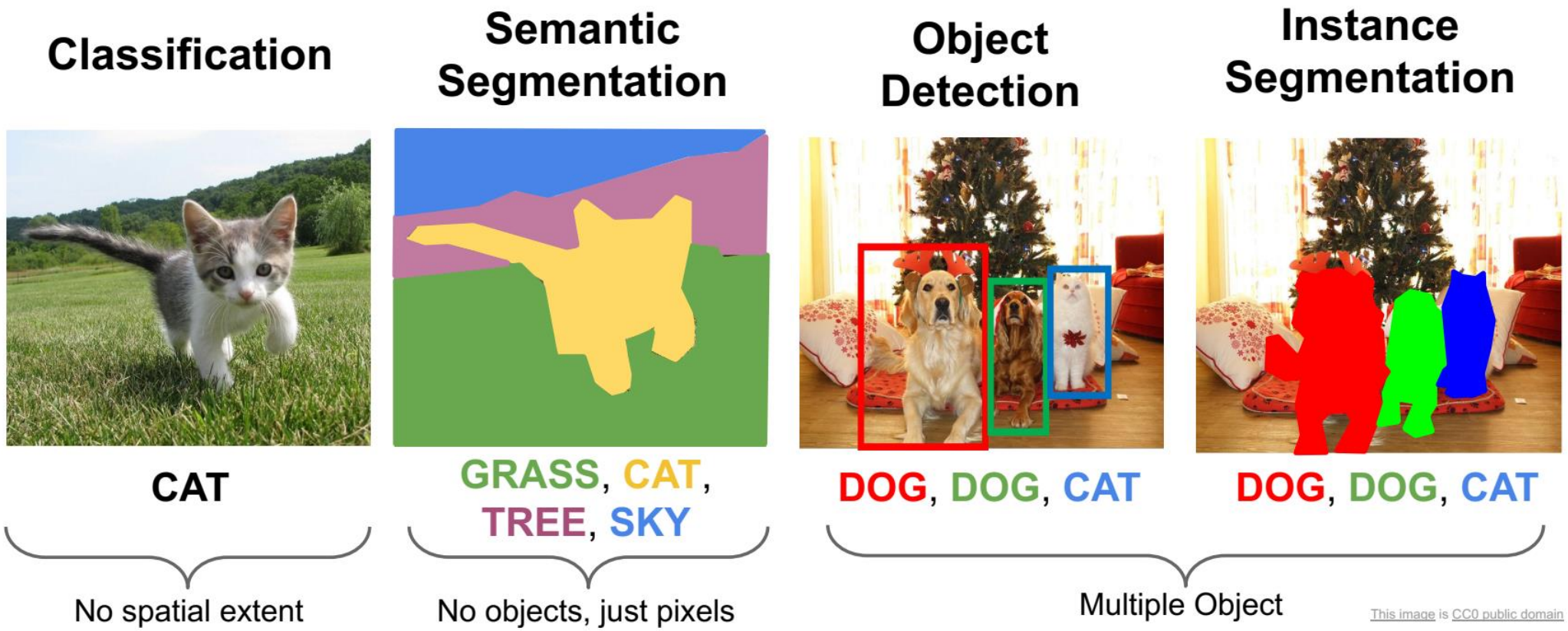
trees

motorbike

road

Cada pixel debe ser categorizado

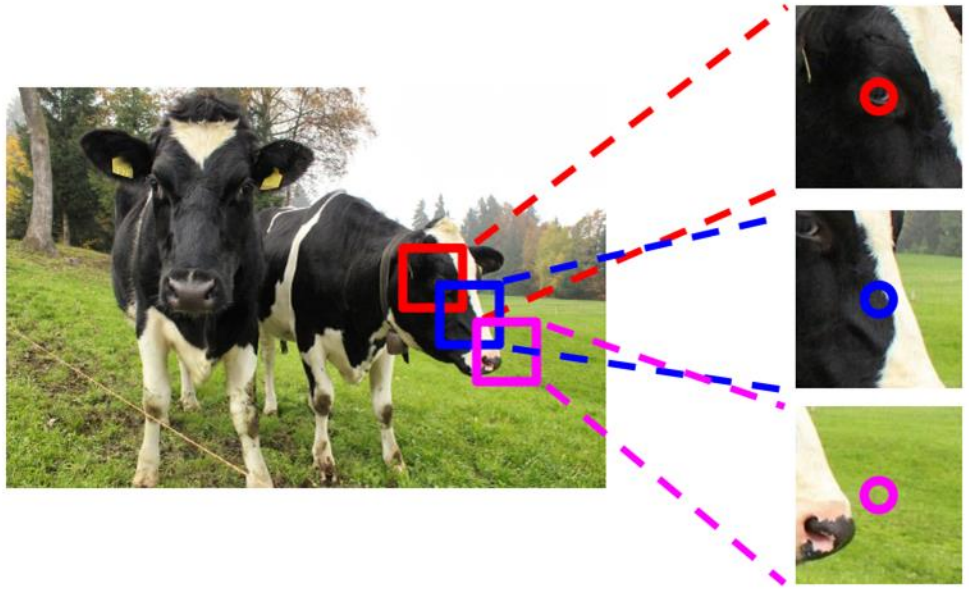
La segmentación semántica es una tarea clásica en visión por computador

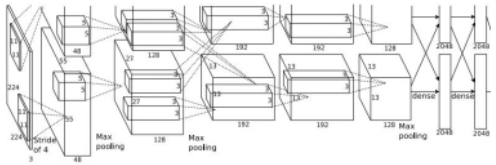


¿Por qué usar CNN para segmentación semántica?

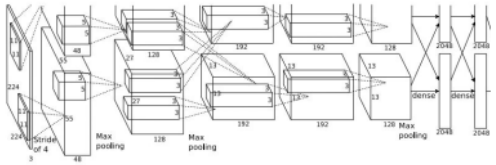
Imposible de clasificar sin algún tipo de contexto

¿Cómo usar CNN para segmentación semántica?

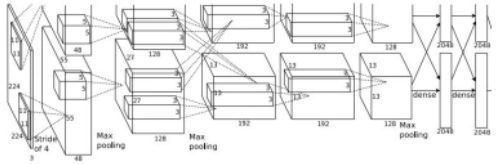




Vaca



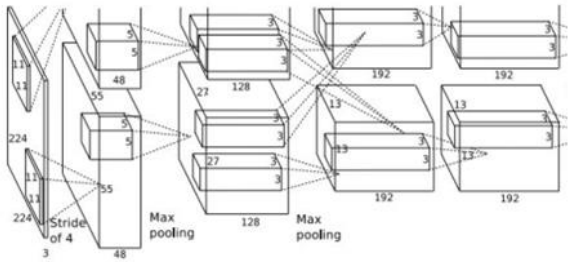
Vaca



Pasto

Altamente ineficiente, repite trabajo ya hecho en detecciones anteriores

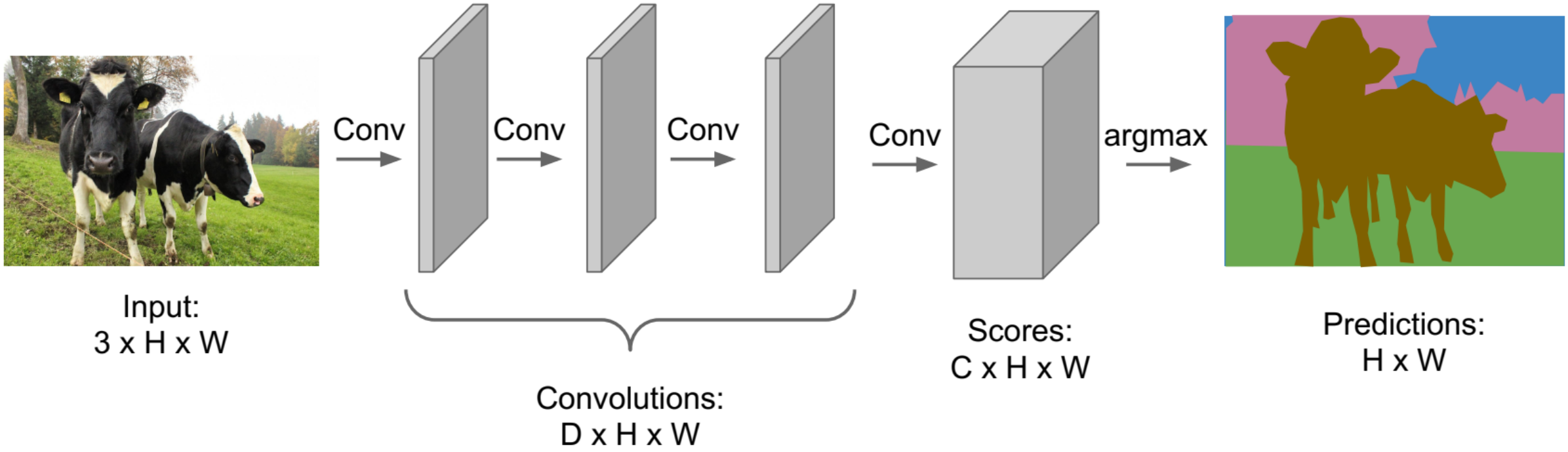
¿Cómo usar CNN para segmentación semántica?



Resolución es fuertemente reducida debido al proceso de la red

¿Cómo usar CNN para segmentación semántica?

Usamos CNN con filtros pequeños, sin pooling, con padding y manteniendo stride pequeño

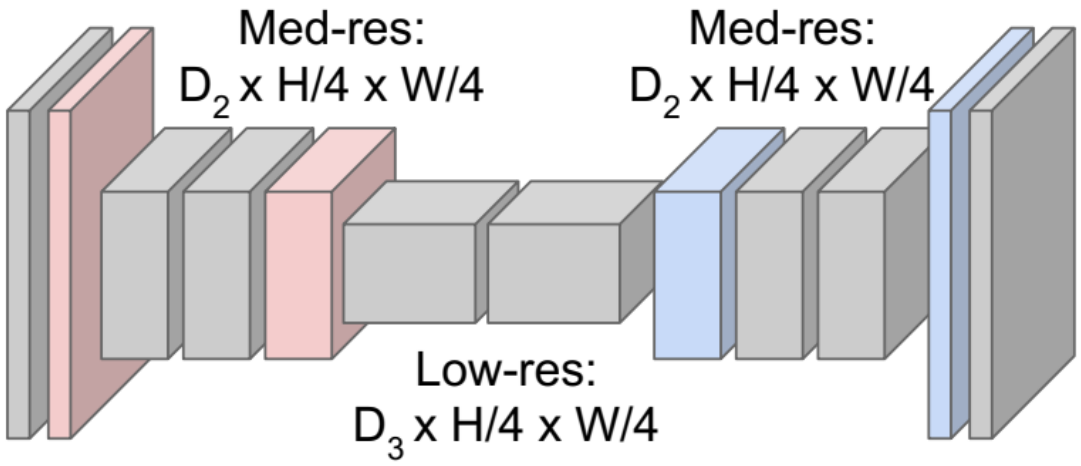


Excesivo costo a nivel de cómputo, no hay “cuello de botella” para capturar features *robustas*

¿Cómo usar CNN para segmentación semántica?

Luego de reducir la dimensionalidad de las features, expandimos el tamaño hasta llegar al original

Input:
3 x H x W



High-res:
D1 x H/2 x W/2

High-res:
D1 x H/2 x W/2

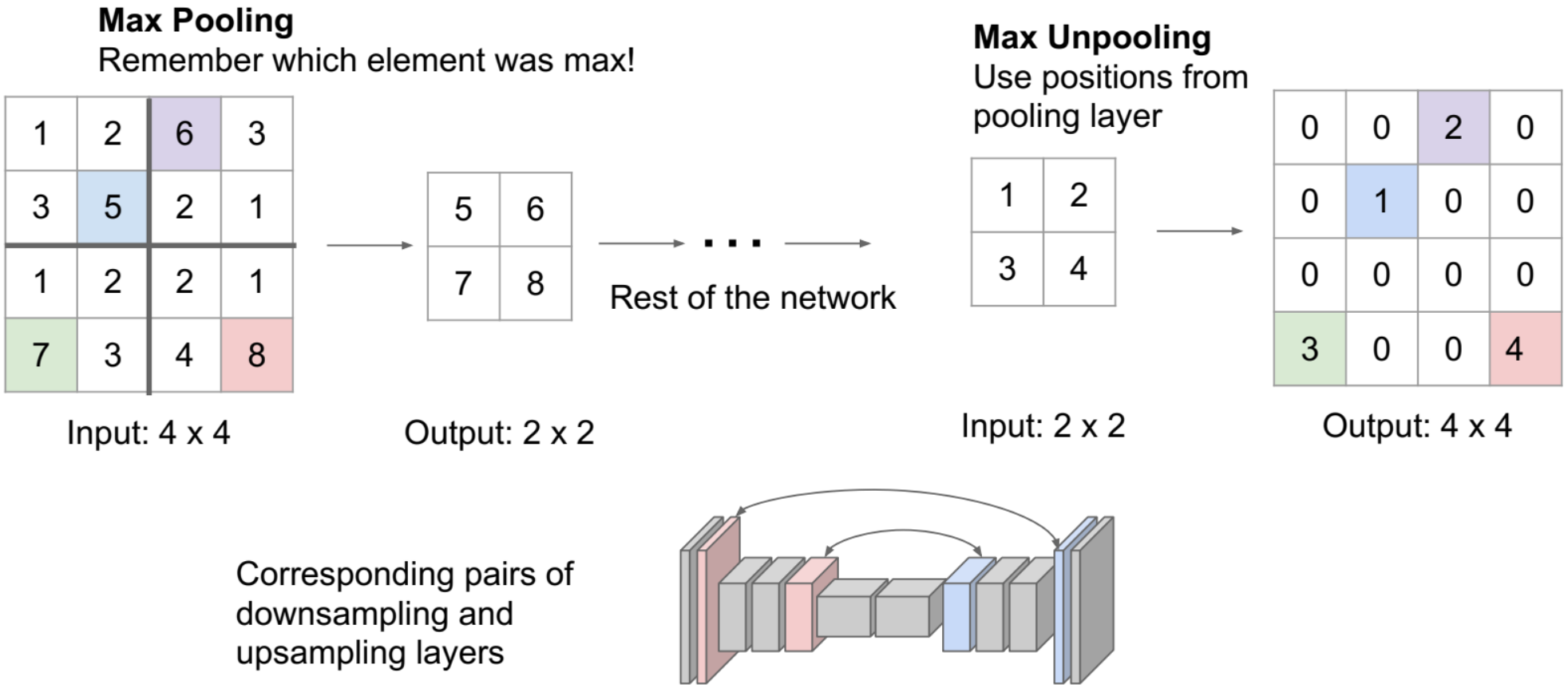
Predictions:
H x W

Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

¿Cómo podemos expandir el tamaño sin perder calidad?

Generalmente se usan dos métodos para expandir:

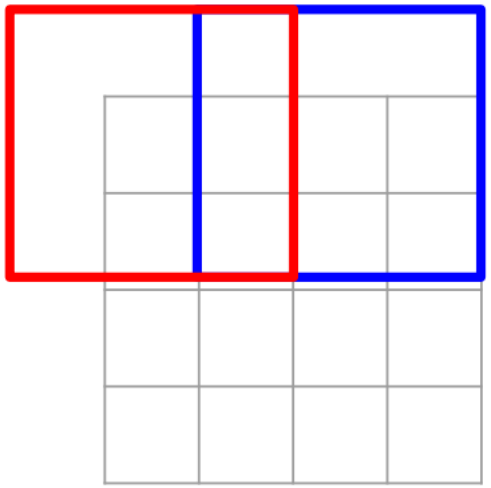
i) *max unpooling*



Corresponding pairs of downsampling and upsampling layers

Generalmente se usan dos métodos para expandir:
ii) filtros aprendibles (convolución transpuesta)

Recall: Normal 3 x 3 convolution, stride 2 pad 1



Input: 4 x 4

Dot product between filter and input

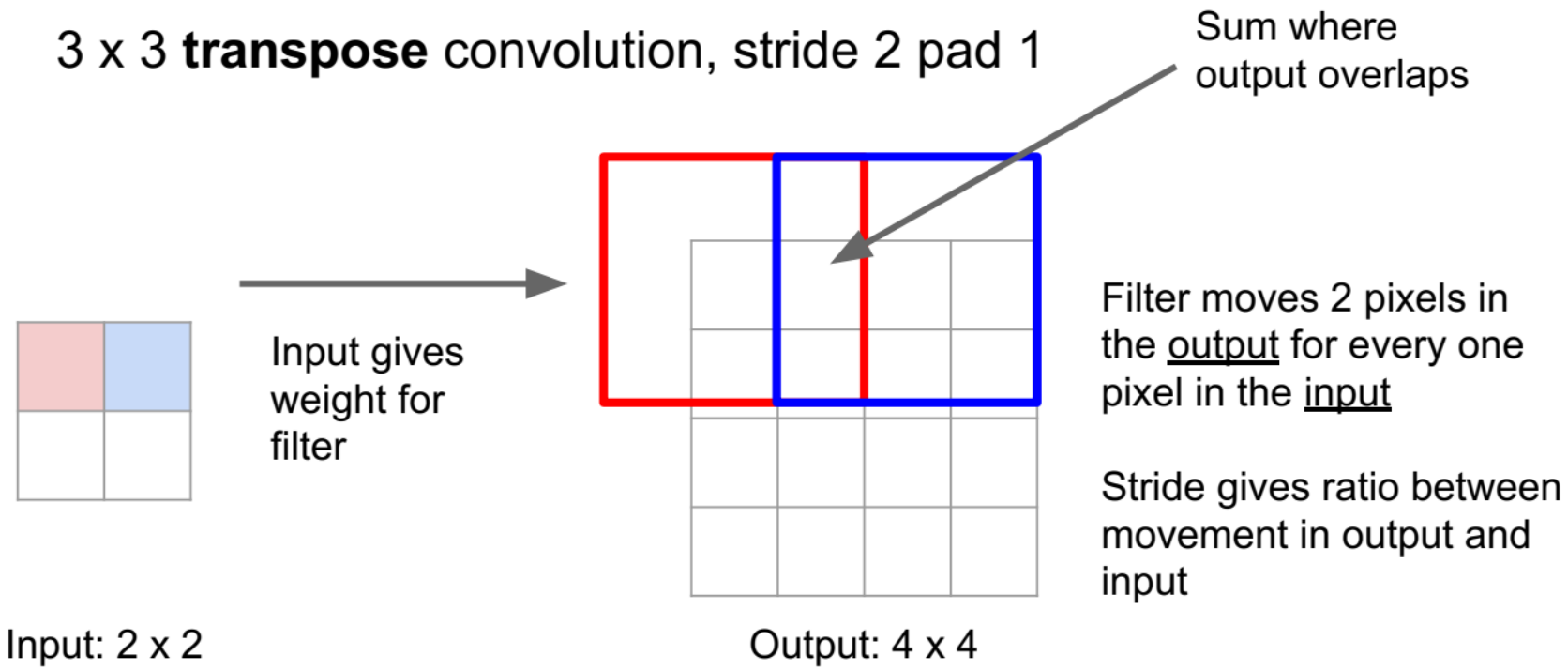
Output: 2 x 2

Filter moves 2 pixels in the input for every one pixel in the output

Stride gives ratio between movement in input and output

We can interpret strided convolution as “learnable downsampling”.

Generalmente se usan dos métodos para expandir:
ii) filtros aprendibles (convolución transpuesta)



Generalmente se usan dos métodos para expandir:
ii) filtros aprendibles (convolución transpuesta)

We can express convolution in terms of a matrix multiplication

$$\vec{x} * \vec{a} = X\vec{a}$$

$$\begin{bmatrix} x & y & x & 0 & 0 & 0 \\ 0 & 0 & x & y & x & 0 \end{bmatrix} \begin{bmatrix} 0 \\ a \\ b \\ c \\ d \\ 0 \end{bmatrix} = \begin{bmatrix} ay + bz \\ bx + cy + dz \end{bmatrix}$$

Example: 1D conv, kernel size=3, stride=2, padding=1

Convolution transpose multiplies by the transpose of the same matrix:

$$\vec{x} *^T \vec{a} = X^T \vec{a}$$

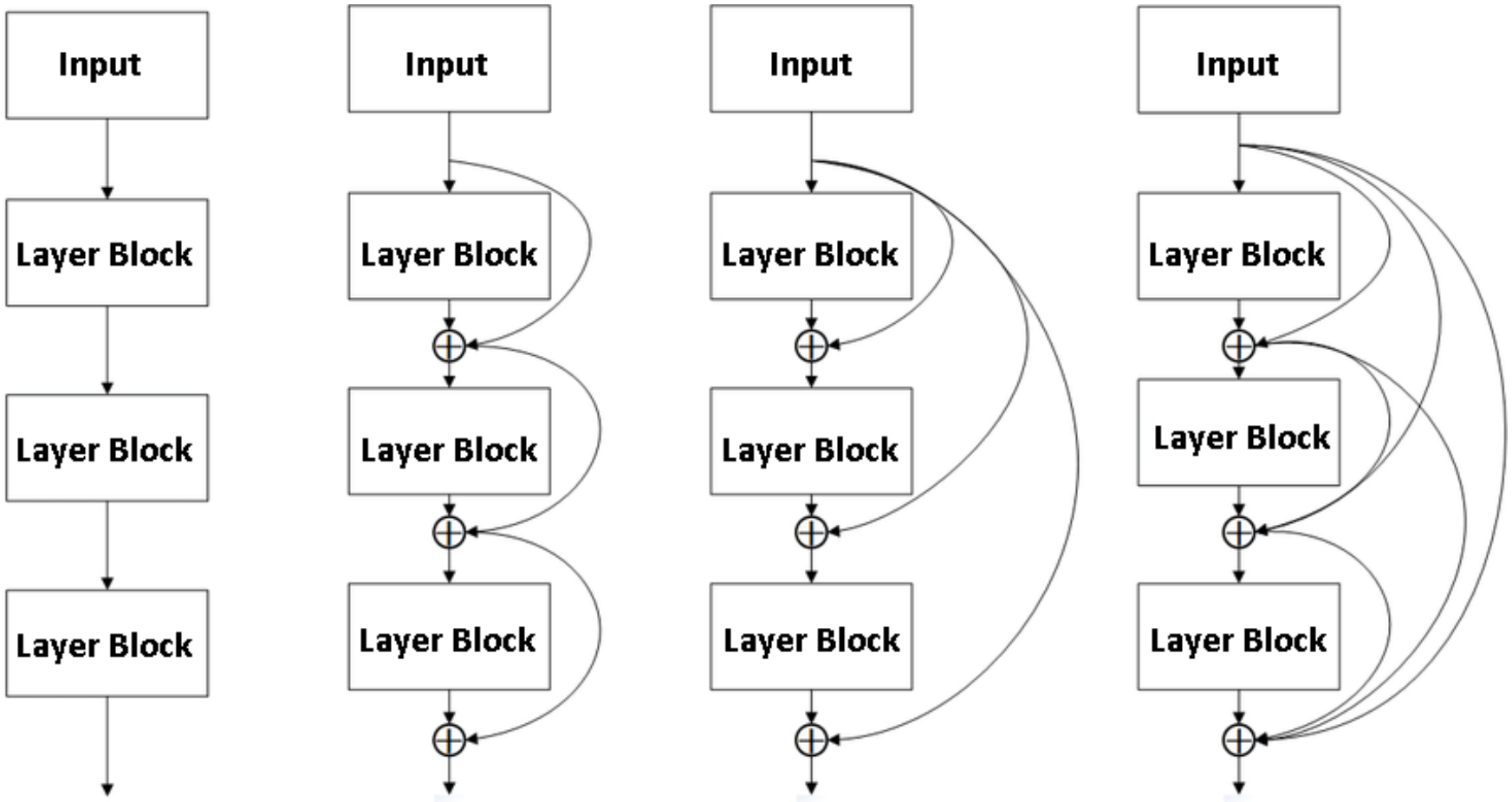
$$\begin{bmatrix} x & 0 \\ y & 0 \\ z & x \\ 0 & y \\ 0 & z \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} ax \\ ay \\ az + bx \\ by \\ bz \\ 0 \end{bmatrix}$$

Example: 1D transpose conv, kernel size=3, stride=2, padding=0

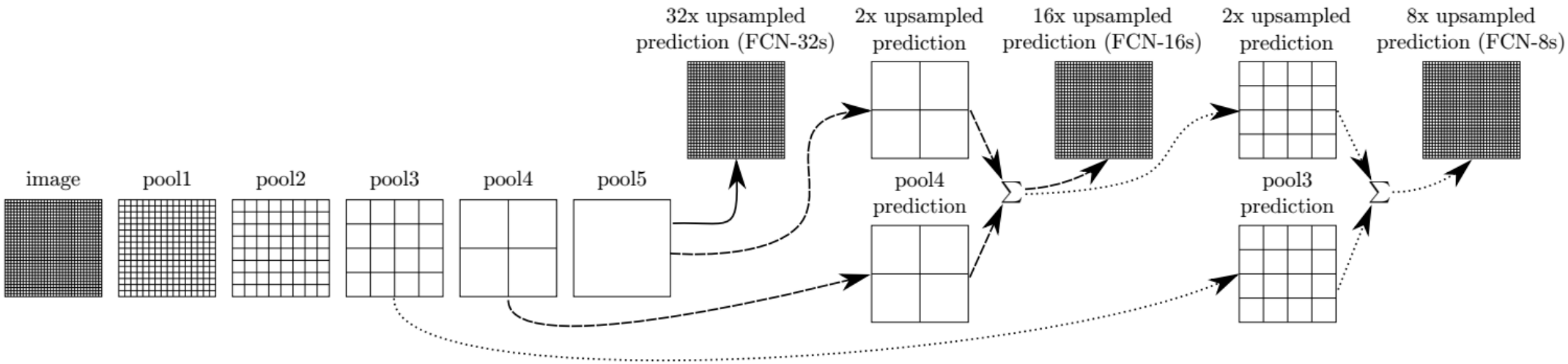
¿Y qué tal los resultados?

No tan buenos en realidad

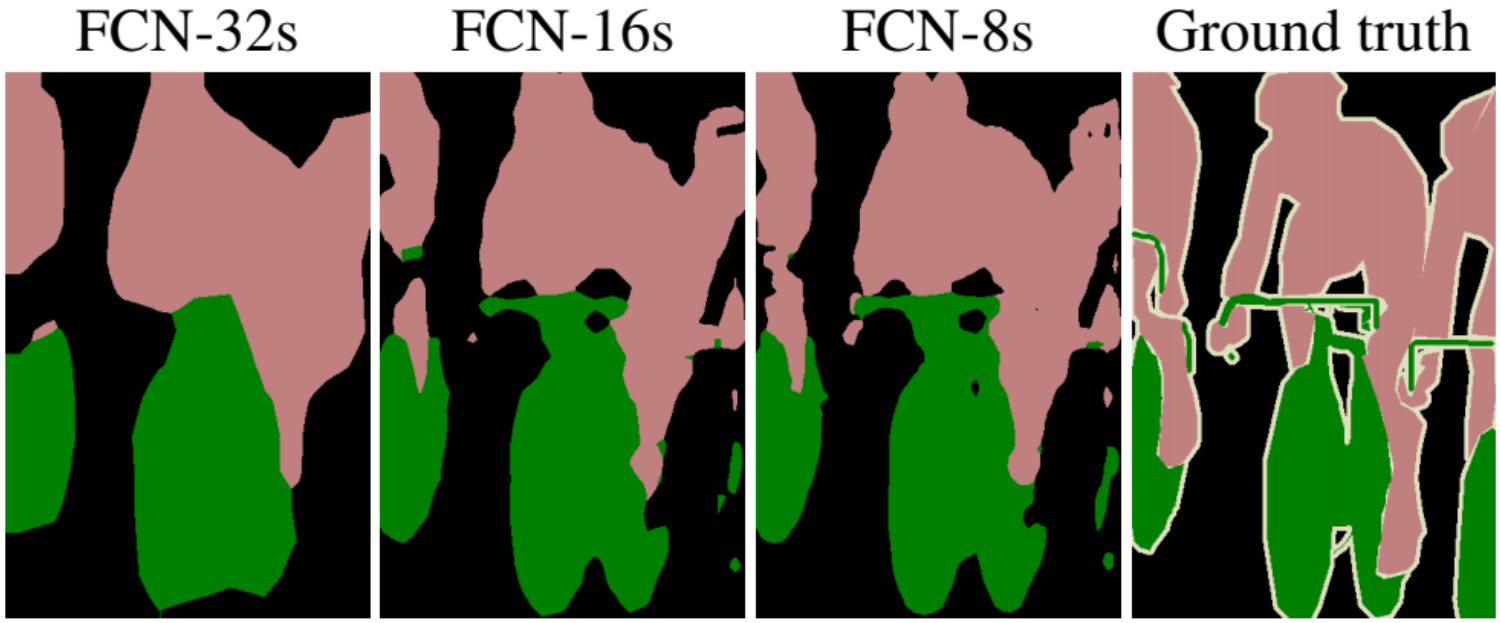
Interludio: *skip connections*



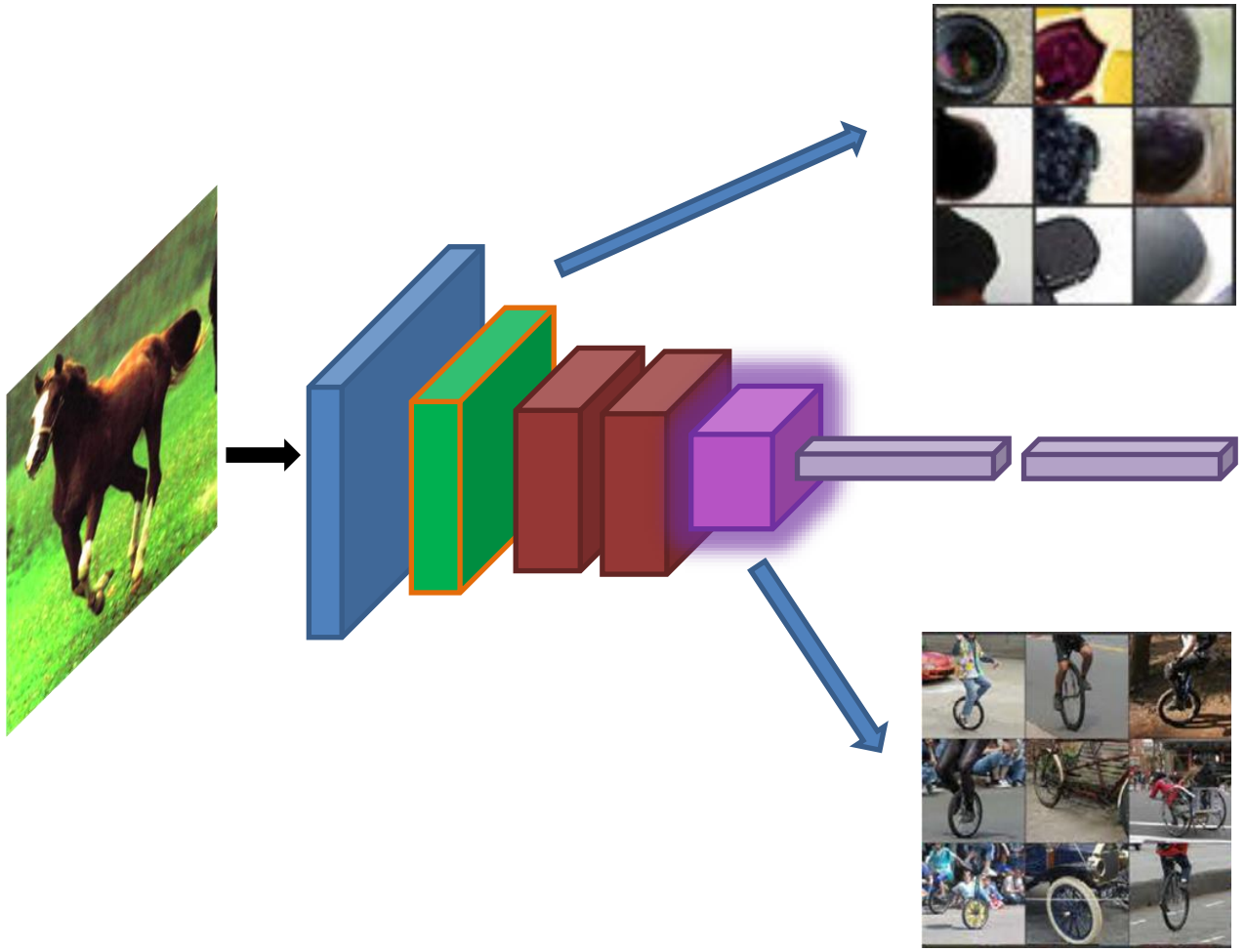
Podemos mejorar la resolución utilizando *skip connections*



Podemos mejorar la resolución utilizando *skip connections*



Un problema de este enfoque es que las primeras capas de una red no son “muy semánticas”



Problema radica en el “*receptive field*” de los filtros

Convoluciones dilatadas entregan una posible solución

- Subsampling (pooling) permite que filtros pequeños capturen mayor información contextual, pero perdemos resolución.

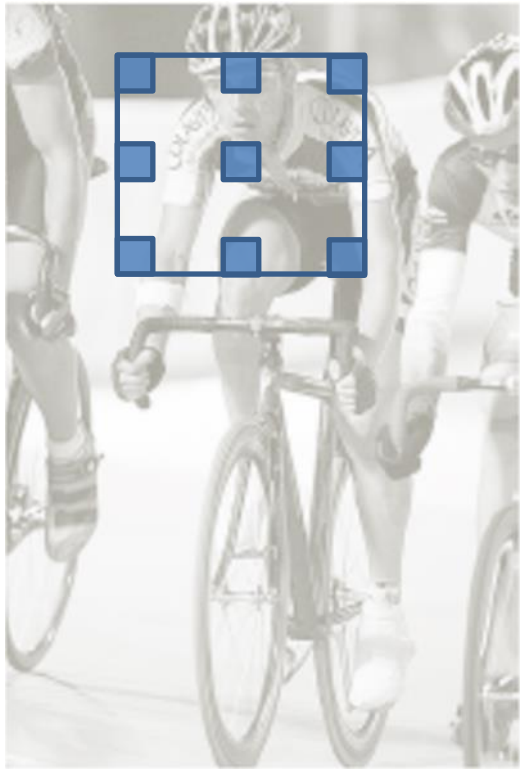
- Filtros grandes capturan mayor información de alta resolución, pero son ruidosos y consumen mucha memoria.

- Es posible aprovechar lo mejor de ambos esquemas, ampliando el tamaño de los filtros, pero no su cantidad de coeficientes.

Convoluciones dilatadas entregan una posible solución

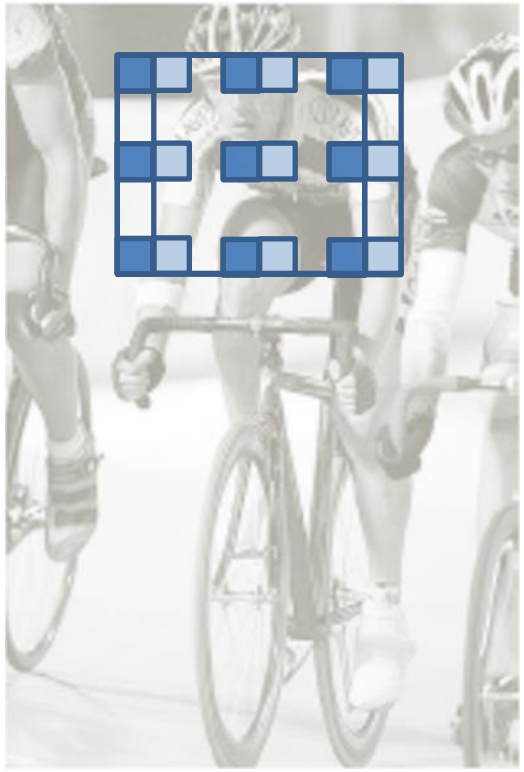
Yu, F., Koltun, V. (2016). Multi-Scale Context Aggregation by Dilated Convolutions. International Conference on Learning Representations (ICLR).

Convoluciones dilatadas entregan una posible solución



Yu, F., Koltun, V. (2016). Multi-Scale Context Aggregation by Dilated Convolutions. International Conference on Learning Representations (ICLR).

Convoluciones dilatadas entregan una posible solución



Yu, F., Koltun, V. (2016). Multi-Scale Context Aggregation by Dilated Convolutions. International Conference on Learning Representations (ICLR).

Enfoques más recientes refinan idea de *skip-connections* para hacer predicción *coarse-to-fine*

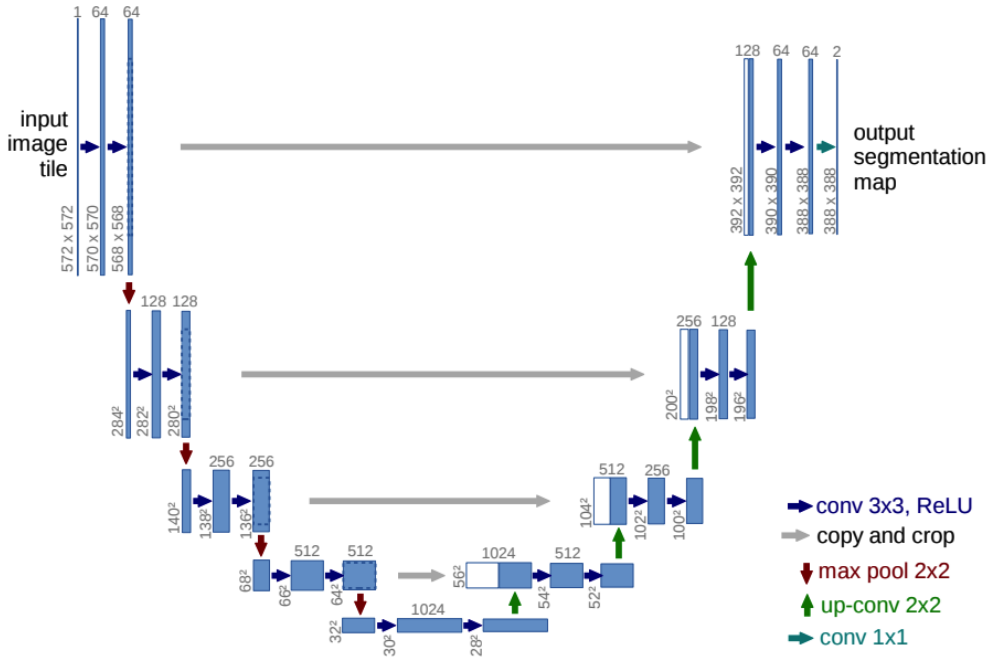
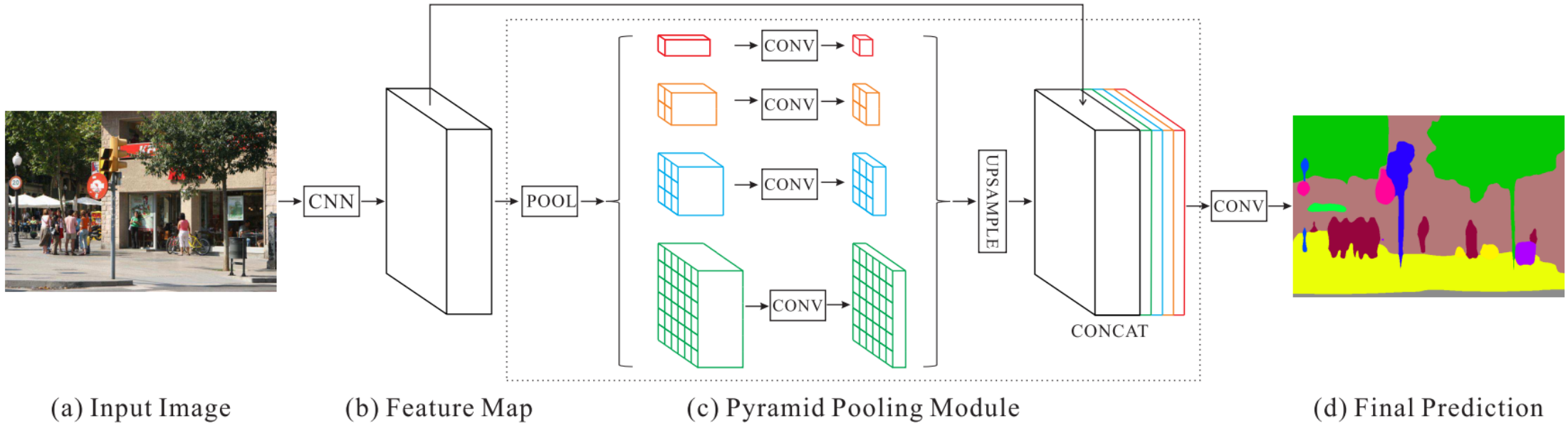


Fig. 1. U-net architecture (example for 32x32 pixels in the lowest resolution). Each blue box corresponds to a multi-channel feature map. The number of channels is denoted on top of the box. The x-y-size is provided at the lower left edge of the box. White boxes represent copied feature maps. The arrows denote the different operations.

Ronneberger, O., Fischer, P., and Brox, T. (2015) "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention.

Otra opción es realizar una subdivisión estructurada de las imágenes para dar mejor contexto



Zha, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017). Multi-Scale Context Aggregation by Dilated Convolutions. Conference on Computer Vision and Pattern Recognition (CVPR)

Otra opción es realizar una subdivisión estructurada de las imágenes para dar mejor contexto

https://youtu.be/HYghTzmbv6Q

Zha, H., Shi, J., Qi, X., Wang, X., Jia, J. (2017). Multi-Scale Context Aggregation by Dilated Convolutions. Conference on Computer Vision and Pattern Recognition (CVPR)

Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

Sistemas Urbanos Inteligentes

Segmentación Semántica

Hans Löbel
Dpto. Ingeniería de Transporte y Logística
Dpto. Ciencia de la Computación