Semantic segmentation

Matteo Moro, matteo.moro@unige.it

Segmentation

Let R be the spatial region occupied by an image

Image segmentation may be defined as the process of partitioning *R* into n subregions

$$\cup_{i=1}^{n} R_i = R$$

$$R_i \text{ is a connected set } i = 1, \ldots, n$$

$$R_i \cap R_j = \emptyset \text{ for all } i \text{ and } j, i \neq j$$

$$Q(R_i) = \text{ TRUE } i = 1, \ldots, n$$

$$Q(R_i \cup R_j) = \text{ FALSE } \text{ for any adjacent region } R_i \text{ and } R_j$$

Q is a logical predicate defined over the points in a set, for instance: Q(A)=TRUE if all pixels in A have the same intensity level

Segmentation – No semantic meaning

Let R be the spatial region occupied by an image

Image segmentation may be defined as the process of partitioning *R* into n subregions

$$\cup_{i=1}^{n} R_i = R$$

$$R_i \text{ is a connected set } i = 1, \ldots, n$$

$$R_i \cap R_j = \emptyset \text{ for all } i \text{ and } j, i \neq j$$

$$Q(R_i) = \text{ TRUE } i = 1, \ldots, n$$

$$Q(R_i \cup R_j) = \text{ FALSE } \text{ for any adjacent region } R_i \text{ and } R_j$$

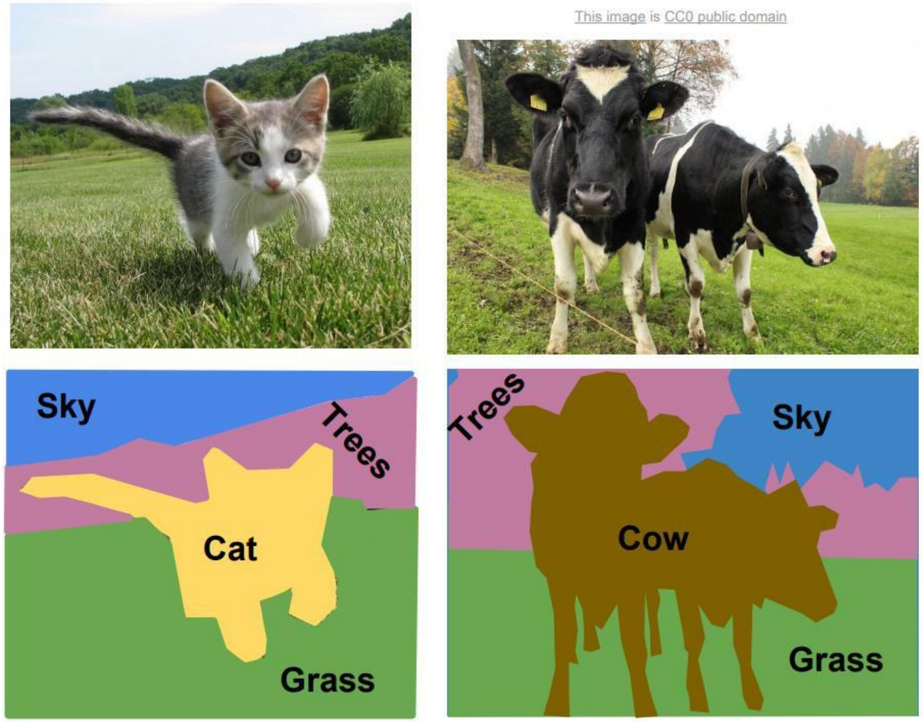
Q is a logical predicate defined over the points in a set, for instance: Q(A)=TRUE if all pixels in A have the same intensity level

Semantic segmentation

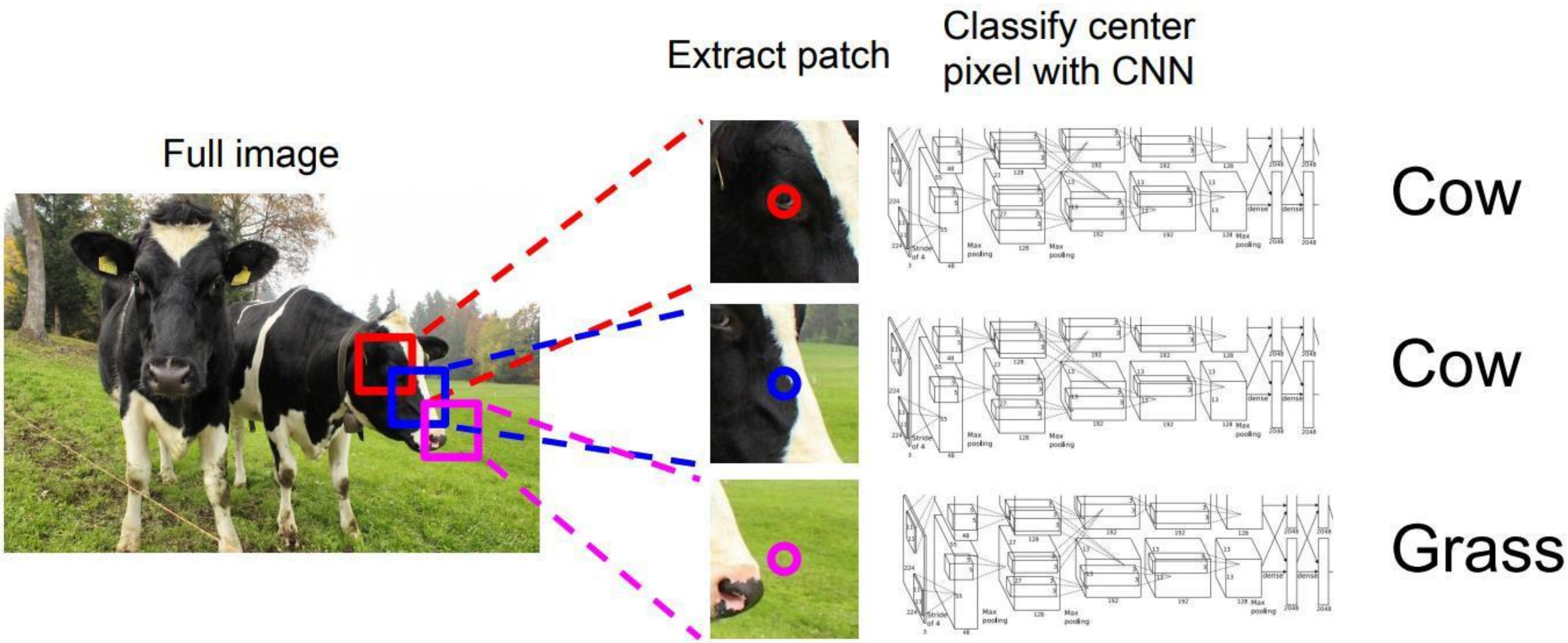
Semantic segmentation: general idea

- Label each pixel in the image with a category label

- Don't differentiate instances, only care about pixels



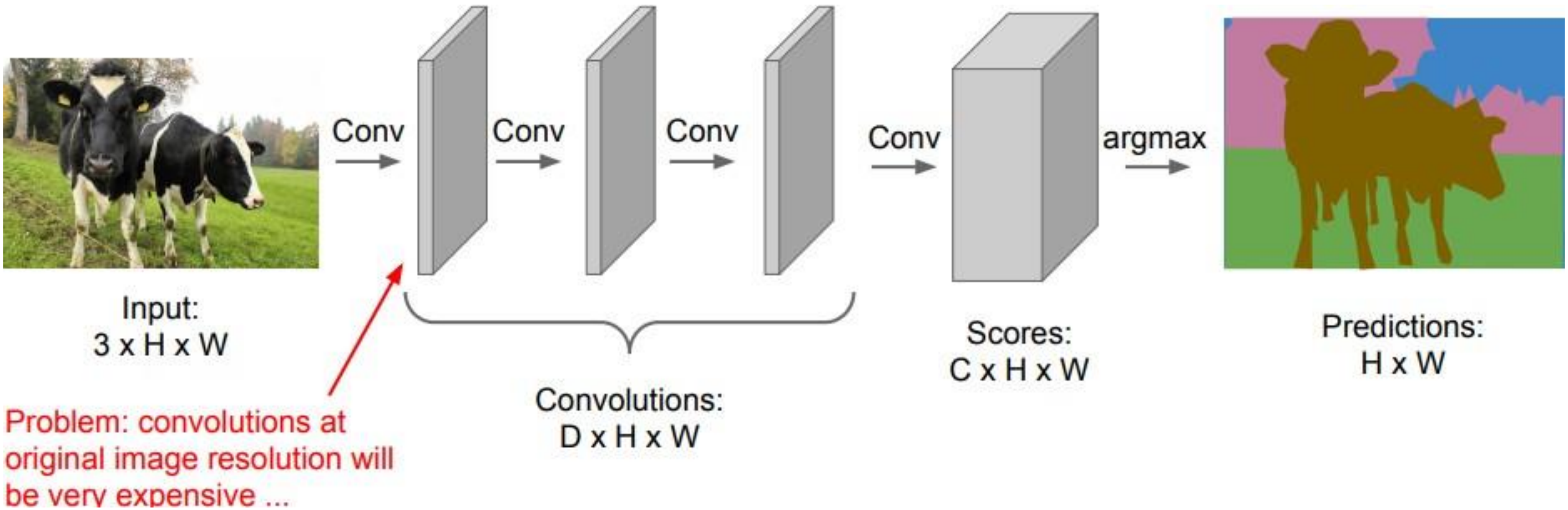
How can we do it? Sliding window



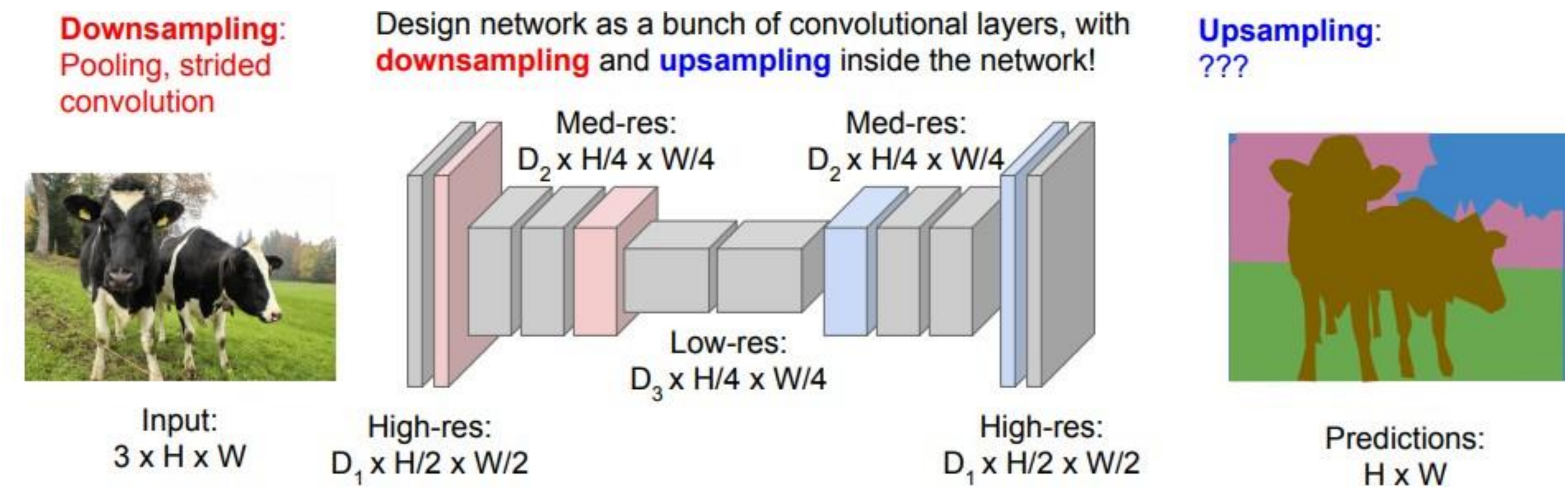
Problem: Very inefficient! Not reusing shared features between overlapping patches

Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013
Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

Semantic segmentation: Fully Convolutional



Semantic segmentation: Downsample and upsample



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

Downsampling : Pooling

2	1	7	1	2	5
5	0	3	4	1	2
1	7	8	3	3	0
0	3	2	0	1	1
3	6	5	3	0	3
3	6	0	2	1	0

Max pooling

8	5
6	3

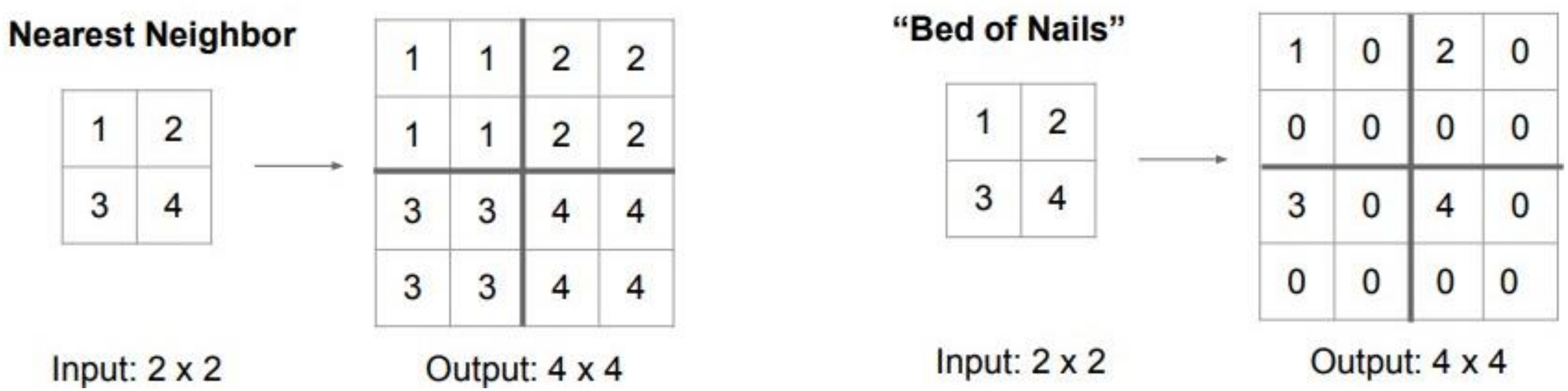
Average pooling

3.8	2.3
3	1.2

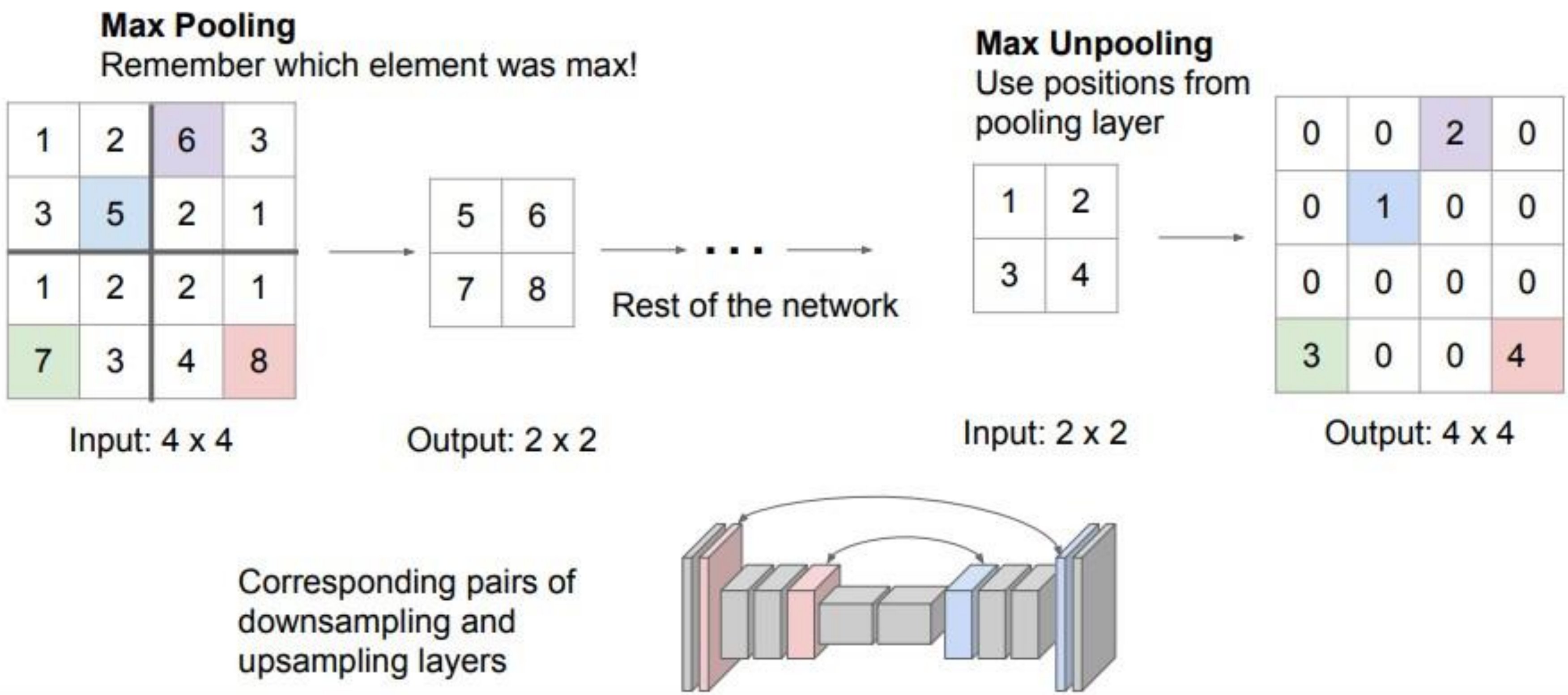
Pooling can help with local invariance although some information is lost

No parameter to be estimated here!

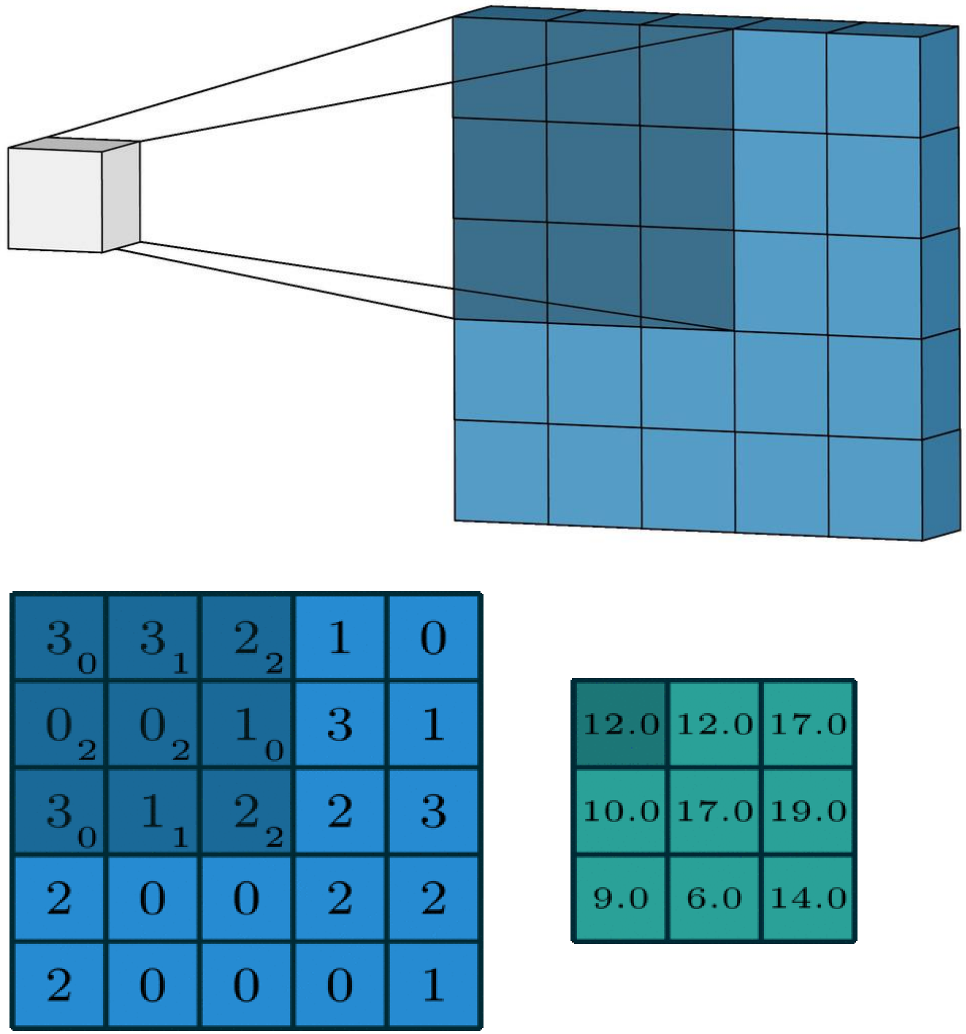
Upsampling: Unpooling



Upsampling: Max Unpooling

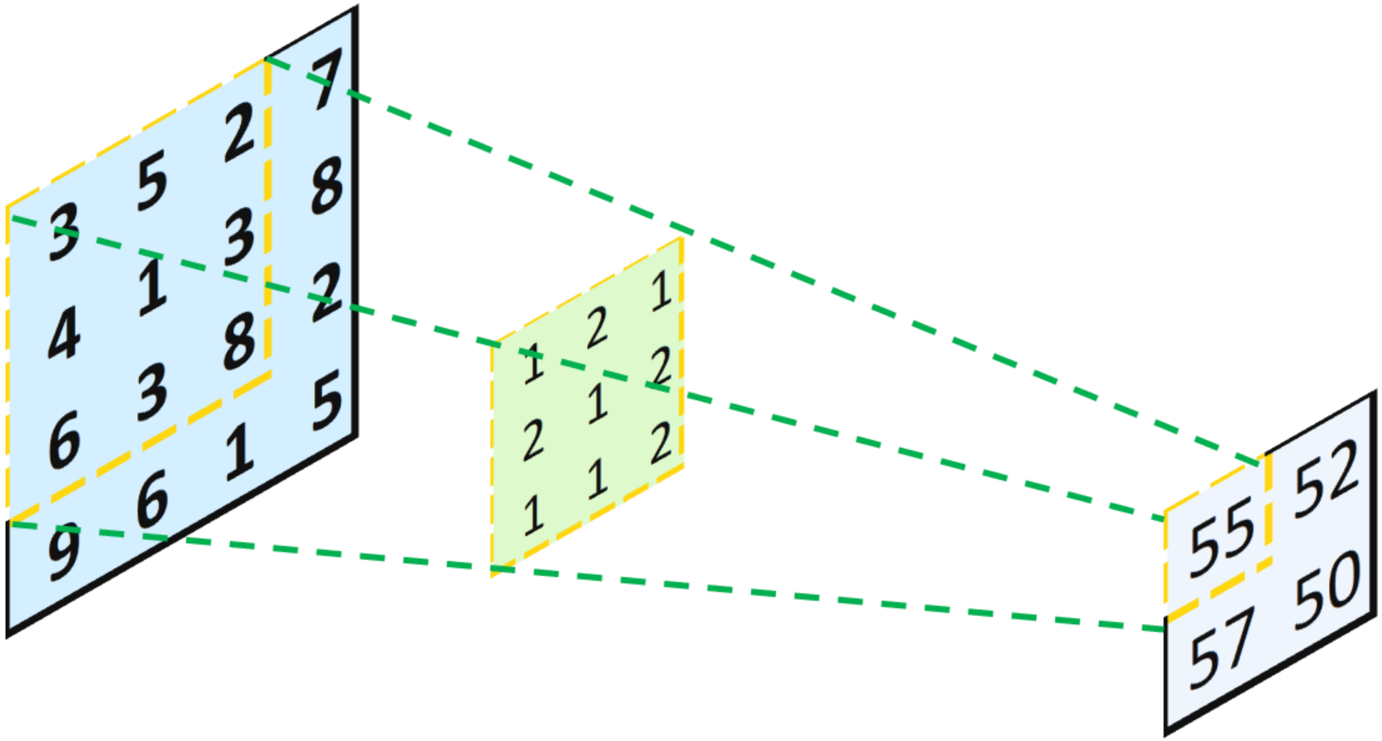


Downsampling: Convolution



Understanding upsampling convolution

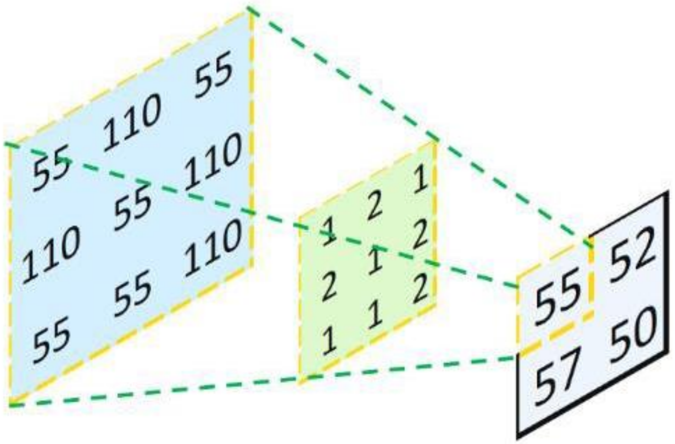
Let us consider this example where a 4x4 image is filtered with 3x3 convolution filters resulting in 2x2 matrix



Now we want to go back to the original image. We do the opposite.

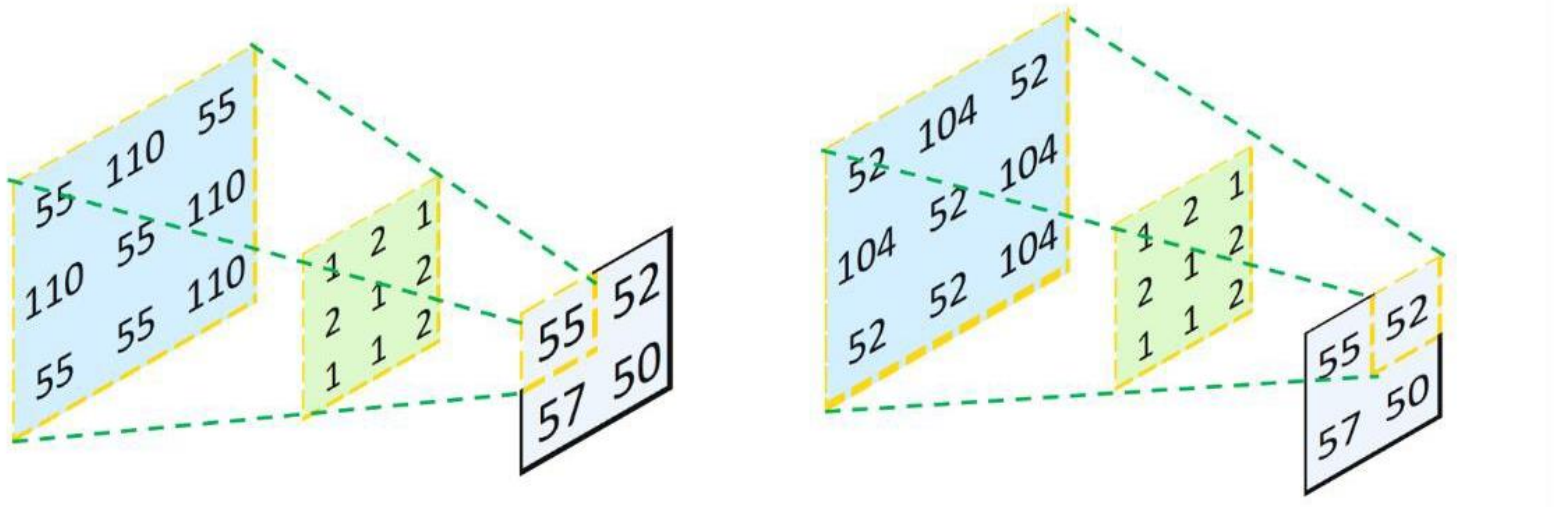
Upsampling: Transposed Convolution

For each pixel value we use the conv filter to transpose the convolution. So, we have one output map for each pixel



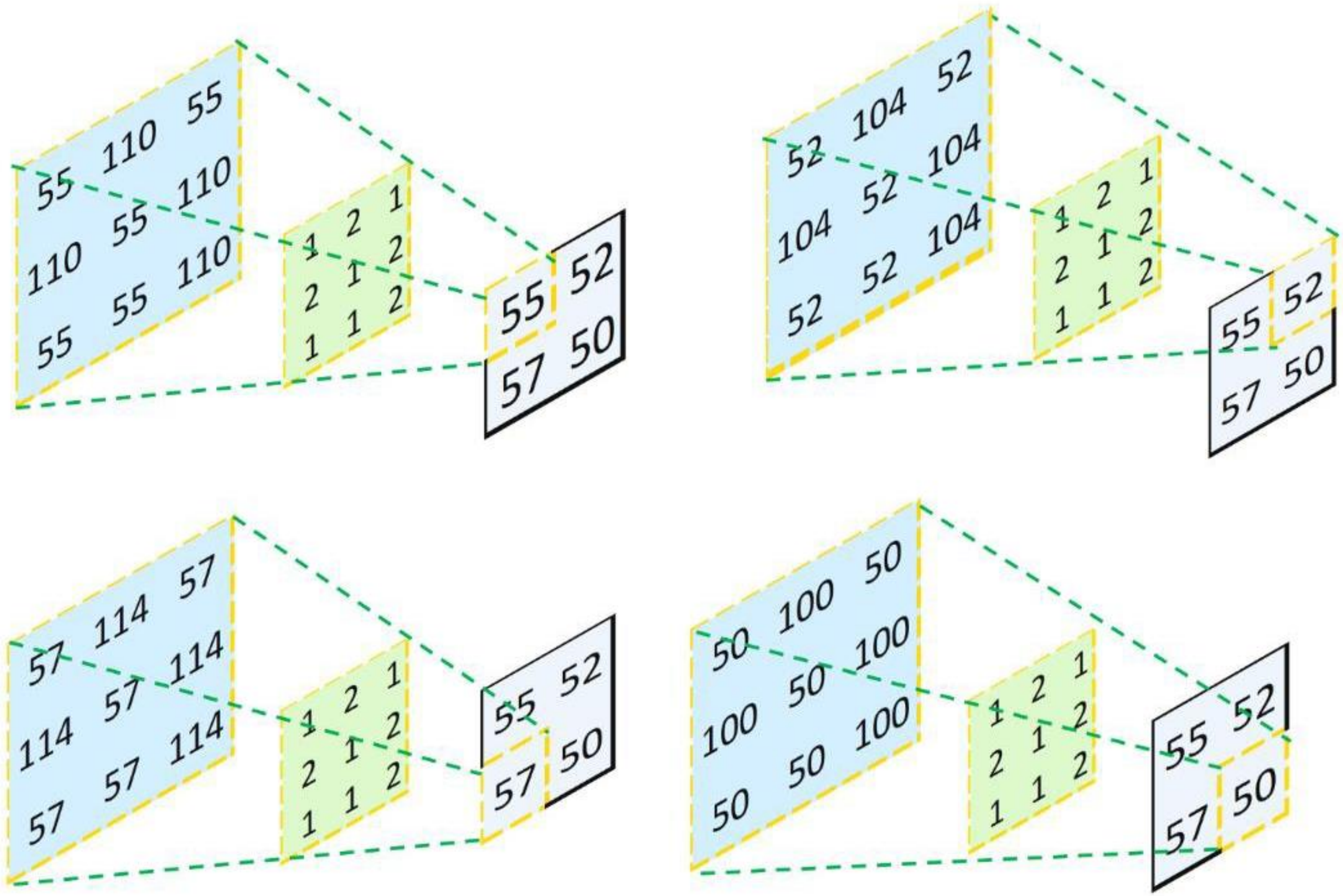
Upsampling: Transposed Convolution

For each pixel value we use the conv filter to transpose the convolution. So, we have one output map for each pixel



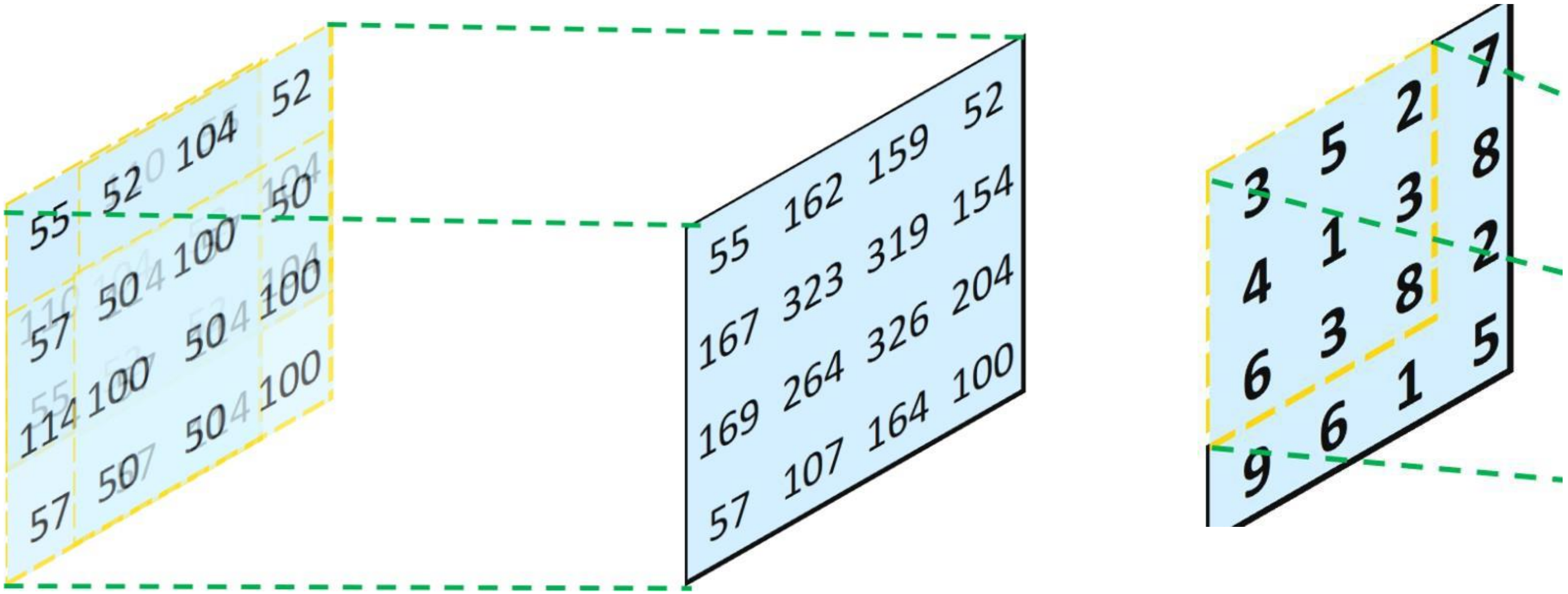
Upsampling: Transposed Convolution

For each pixel value we use the conv filter to transpose the convolution. So, we have one output map for each pixel



Upsampling: Transposed Convolution

We overlap the transposed maps (according to position, and we sum overlapping values obtaining the final output)



Values are different! Transpose convolution kernels are learnable (training)

UniGe | MaLGa

Basic neural networks architectures for segmentation

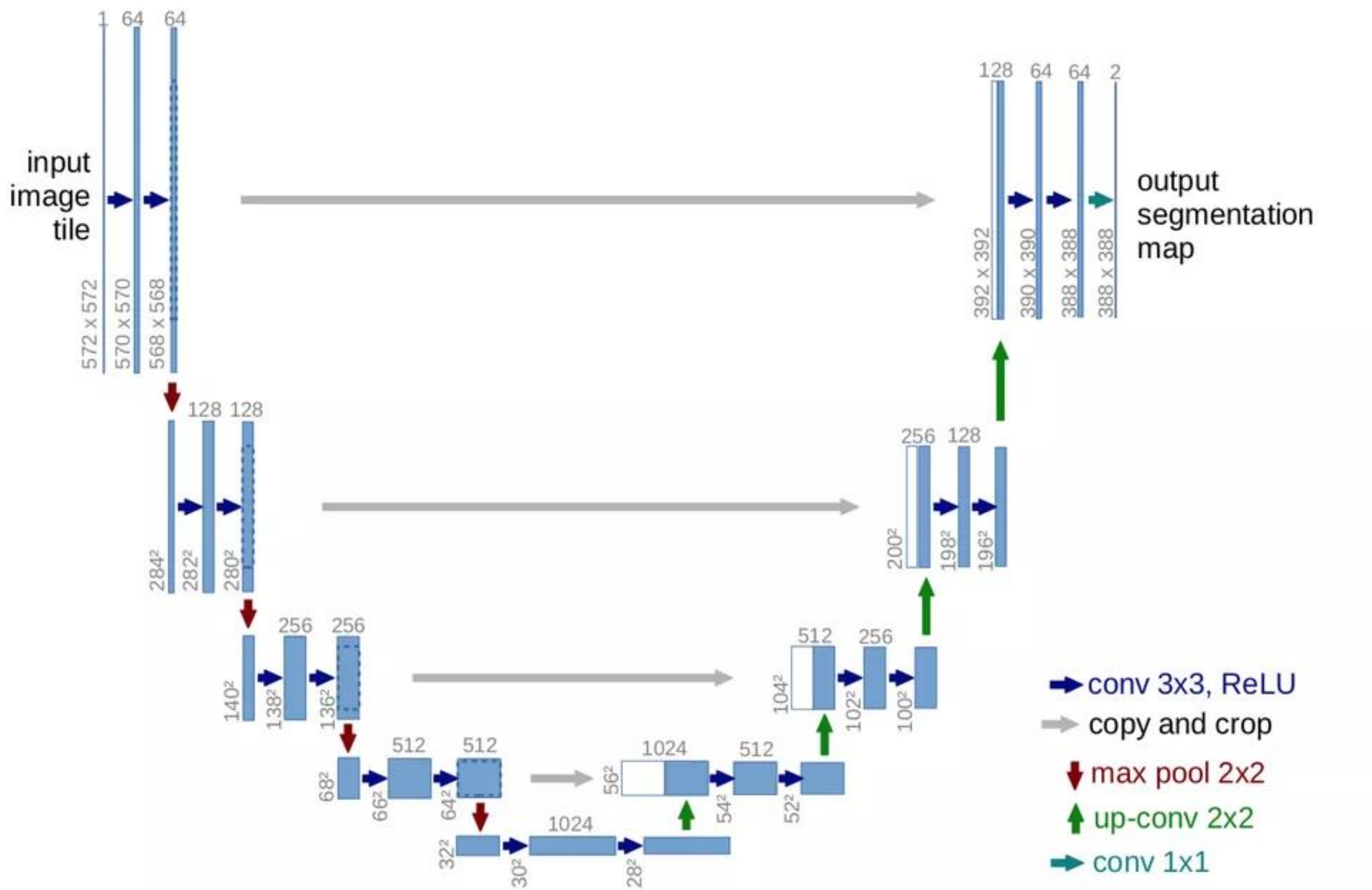
A basic network for semantic segmentation: The Unet

- The *UNET* was developed by Olaf Ronneberger et al. for Biomedical Image Segmentation

- Encoder-decoder network

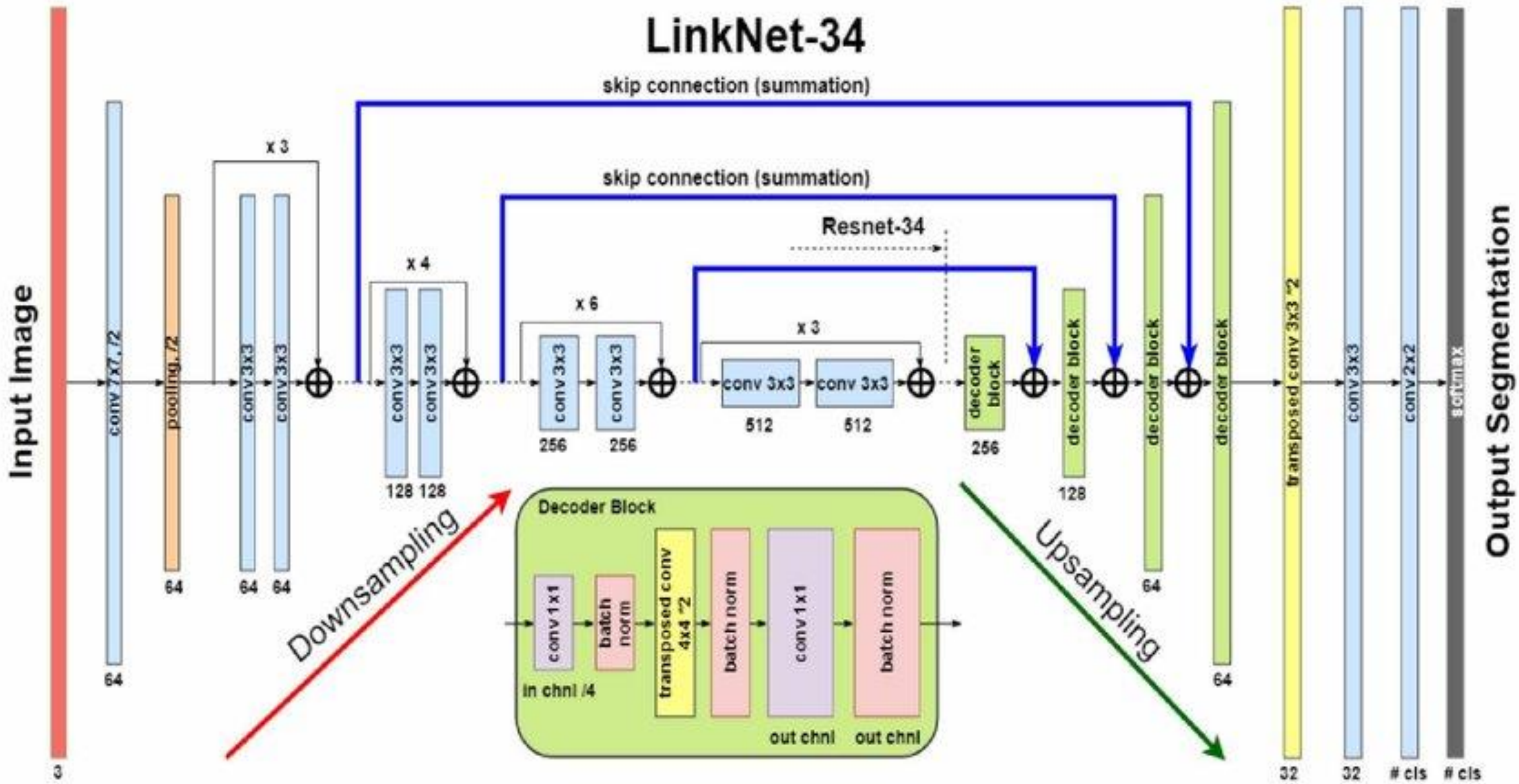
- It won the **Grand Challenge for Computer-Automated Detection of Caries in Bitewing Radiography at ISBI 2015**, and the **Cell Tracking Challenge at ISBI 2015**

Segmentation with Unet: Architecture



Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.

Unet variants: LinkNet



Unet variants: Pyramid Scene Parsing Network

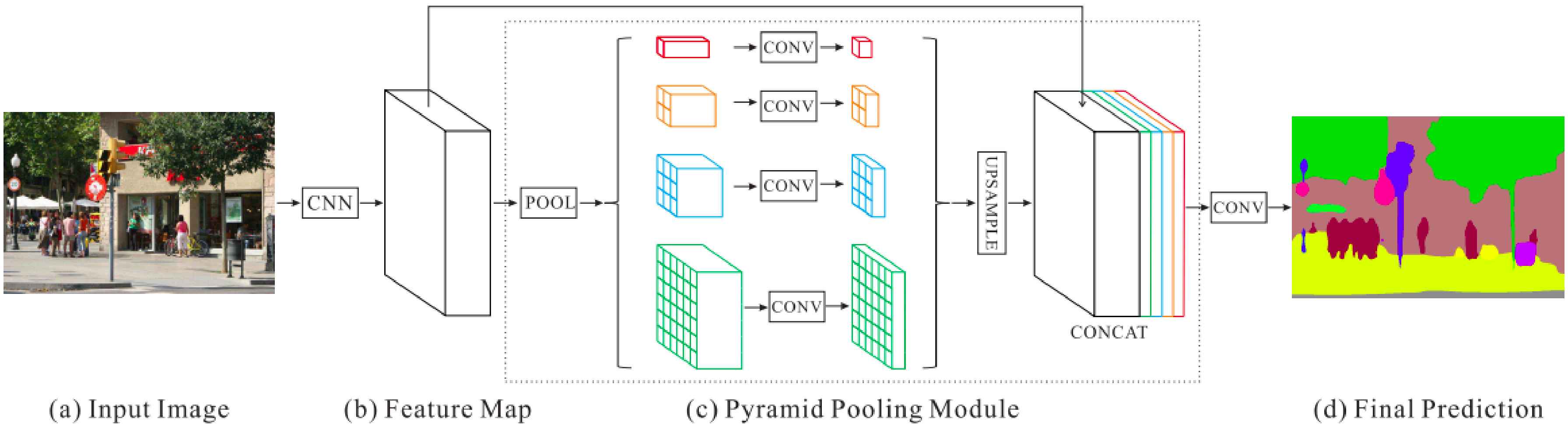
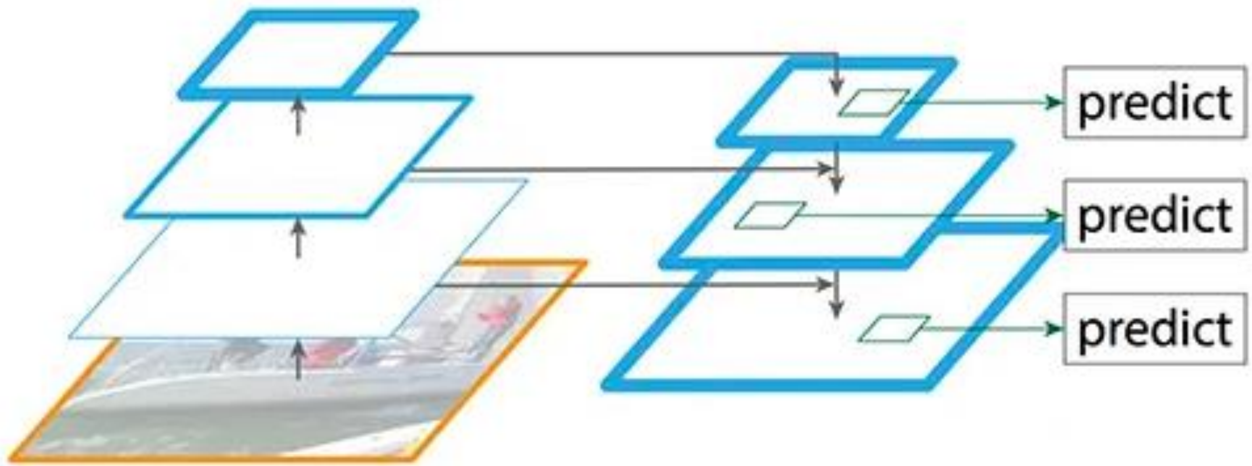


Figure 3. Overview of our proposed PSPNet. Given an input image (a), we first use CNN to get the feature map of the last convolutional layer (b), then a pyramid parsing module is applied to harvest different sub-region representations, followed by upsampling and concatenation layers to form the final feature representation, which carries both local and global context information in (c). Finally, the representation is fed into a convolution layer to get the final per-pixel prediction (d).

PSPNet outperformed state-of-the-art available neural networks in the task of segmentation in 2017.

Unet variants: Feature Pyramid Network



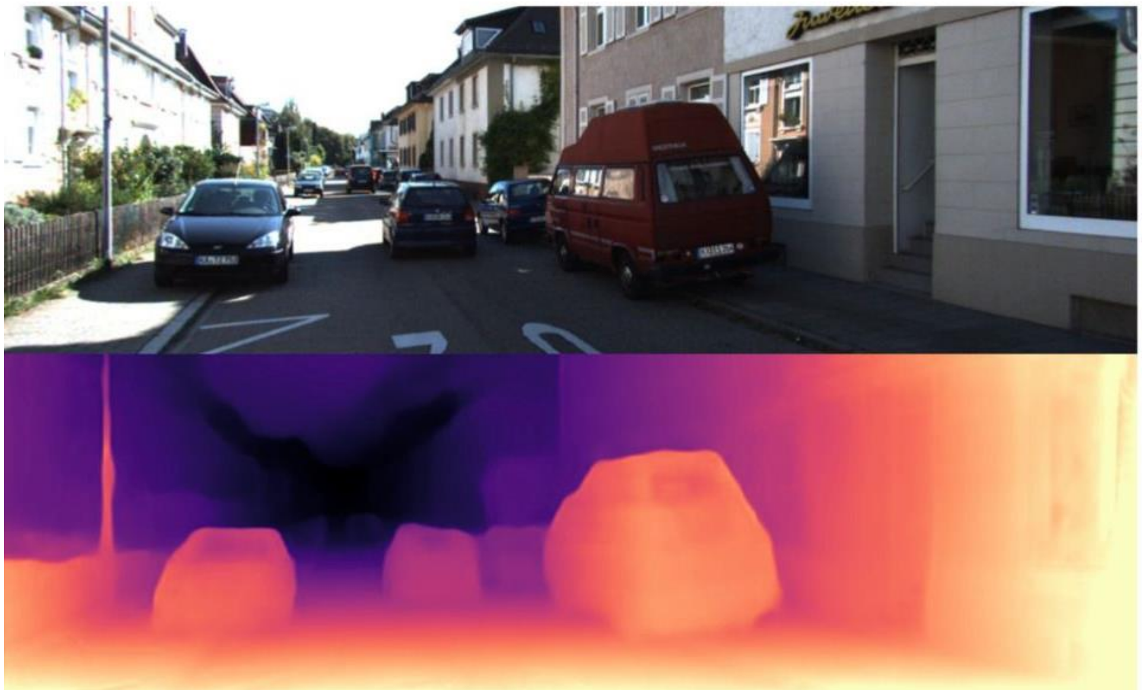
FPN (Source)

FPN composes of a **bottom-up** and a **top-down** pathway. The bottom-up pathway is the usual convolutional network for feature extraction. As we go up, the spatial resolution decreases. With more high-level structures detected, the **semantic value** for each layer increases.

Depth estimation

Monocular depth estimation

Monocular Depth Estimation is the task of estimating the **depth value (distance relative to the camera)** of each pixel given a single (monocular) RGB image.



Monocular depth estimation

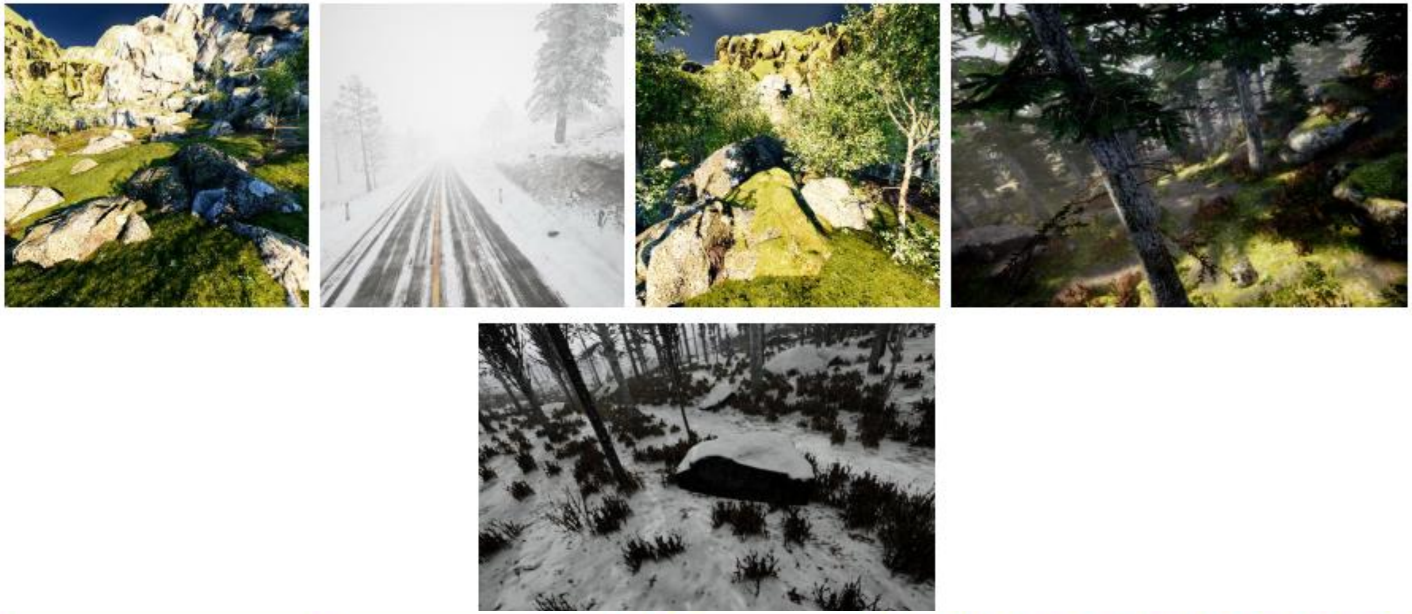
State-of-the-art methods usually fall into one of two categories:

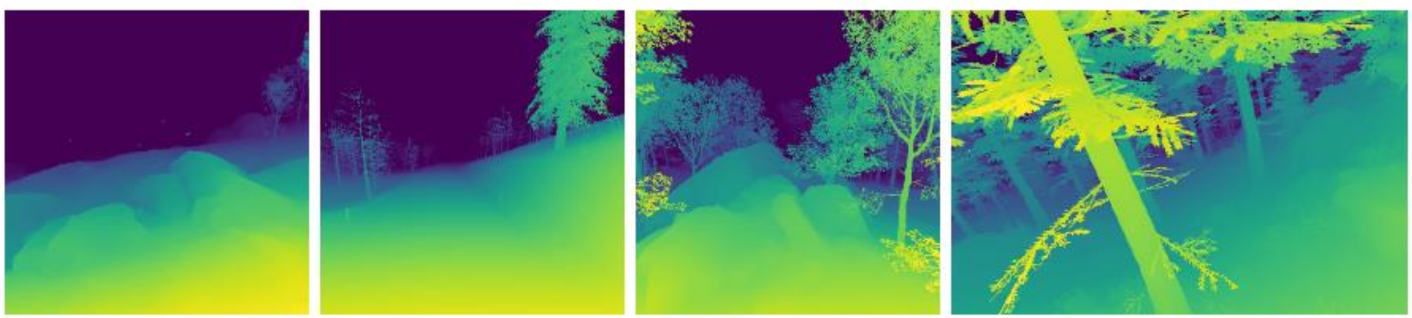
-) designing a complex network that is powerful enough to directly regress the depth map

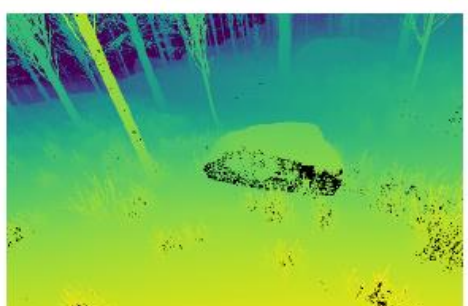
-) splitting the input into bins or windows to reduce computational complexity

Lack ok ground truth datasets for supervised training

M4Depth

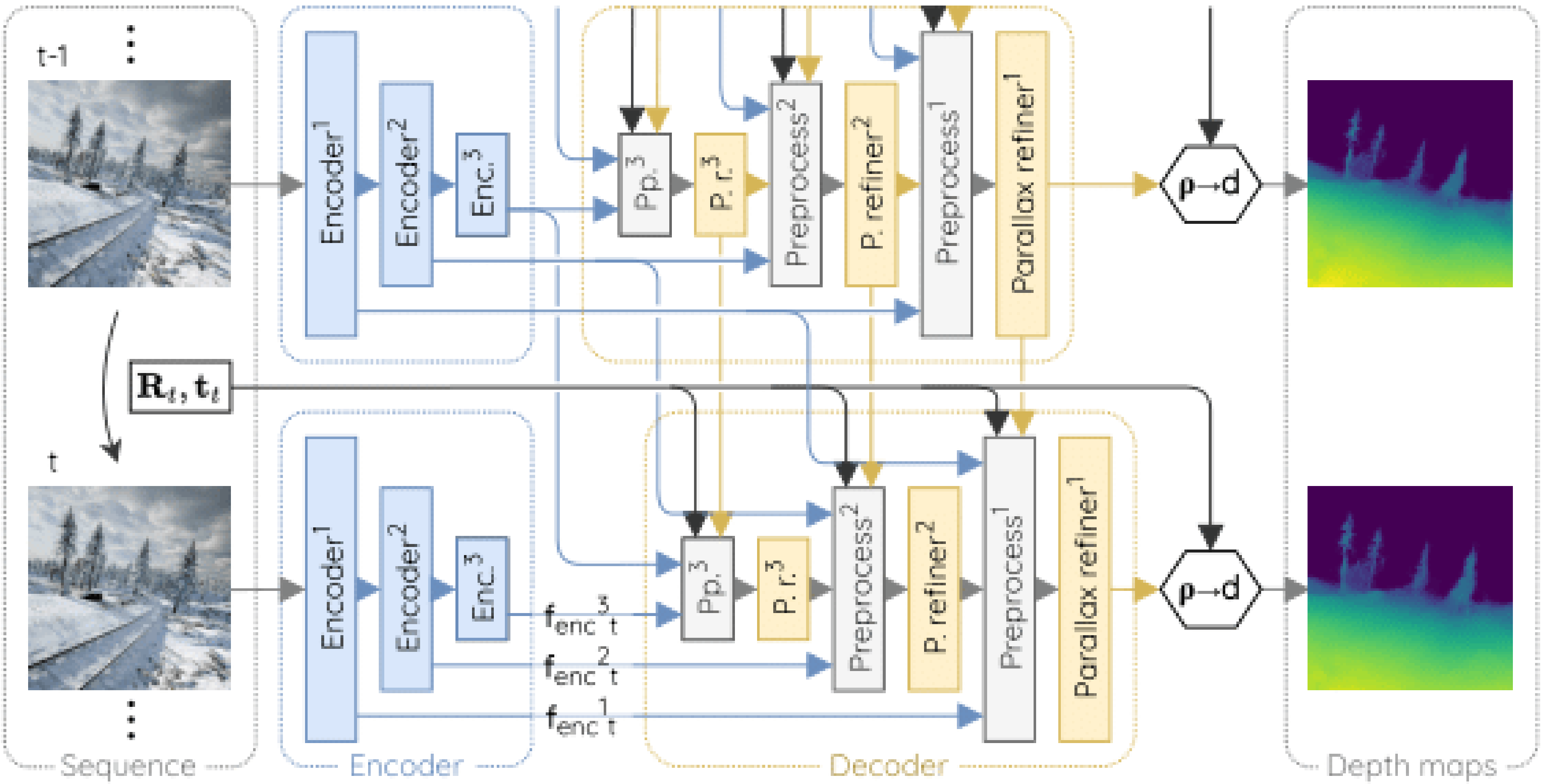






Fonder, Michaël, Damien Ernst, and Marc Van Droogenbroeck. "Parallax inference for robust temporal monocular depth estimation in unstructured environments." Sensors 22.23 (2022): 9374.

M4Depth



Fonder, Michaël, Damien Ernst, and Marc Van Droogenbroeck. "Parallax inference for robust temporal monocular depth estimation in unstructured environments." Sensors 22.23 (2022): 9374.

UniGe

MaLGa