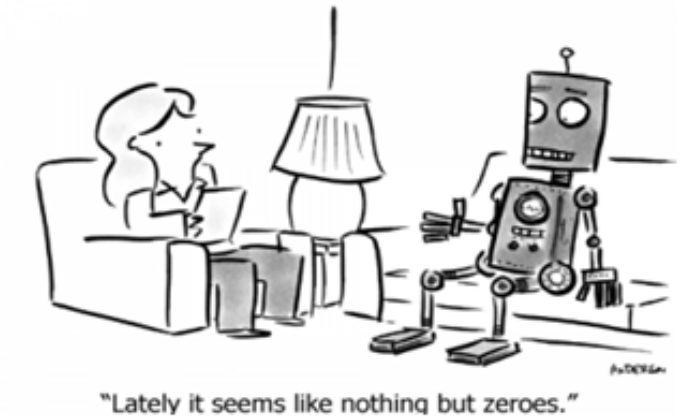
Semantic Segmentation Deep learning

Gianni FRANCHI
ENSTA-Paris



"Lately it seems like nothing but zeroes."

Semantic segmentation

Most images come from **[web1] [web2]**.
For a better understanding of the field please read : http://d2l.ai/

What is Semantic segmentation ?

Input

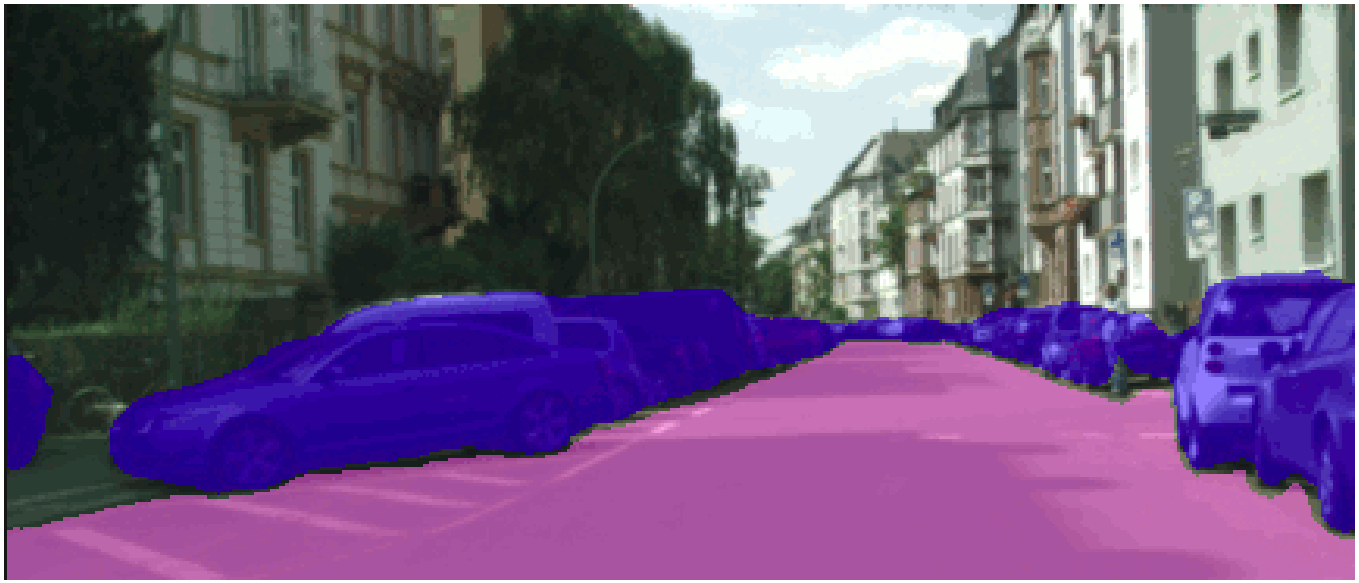
segmented

1: Person
2: Purse
3: Plants/Grass
4: Sidewalk
5: Building/Structures

3	3	3	3	3	3	3	3	3	3	3	3	5	5	5	5	5	5
3	3	3	3	3	3	3	3	3	3	3	3	5	5	5	5	5	5
3	3	3	3	3	3	1	1	3	3	3	3	5	5	5	5	5	5
3	3	3	3	3	1	1	1	1	3	3	3	5	5	5	5	5	5
3	3	3	3	3	3	1	1	3	3	3	5	5	5	5	5	5	5
5	5	3	3	3	3	1	1	3	3	5	5	5	5	5	5	5	5
4	4	3	4	1	1	1	1	1	1	4	4	4	5	5	5	5	5
4	4	3	4	1	1	1	1	1	1	4	4	4	4	4	5	5	5
4	4	4	1	1	1	1	1	1	1	1	4	4	4	4	4	4	4
3	3	3	1	1	1	1	1	1	1	1	4	4	4	4	4	4	4
3	3	3	1	2	2	1	1	1	1	1	4	4	4	4	4	4	4
3	3	3	1	2	2	1	1	1	1	1	4	4	4	4	4	4	4

Semantic Labels

Why do we do semantic segmentation ?



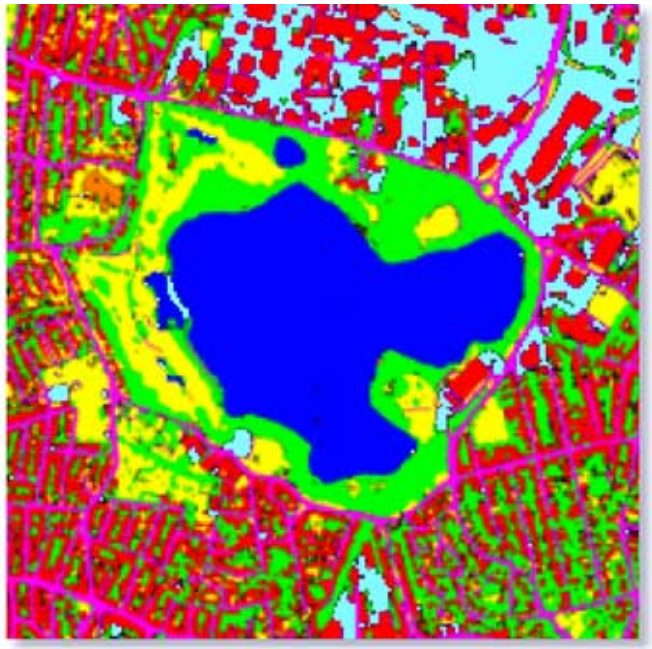
To have a better understanding of the scene.

Where do we apply semantic segmentation?

Autonomous driving

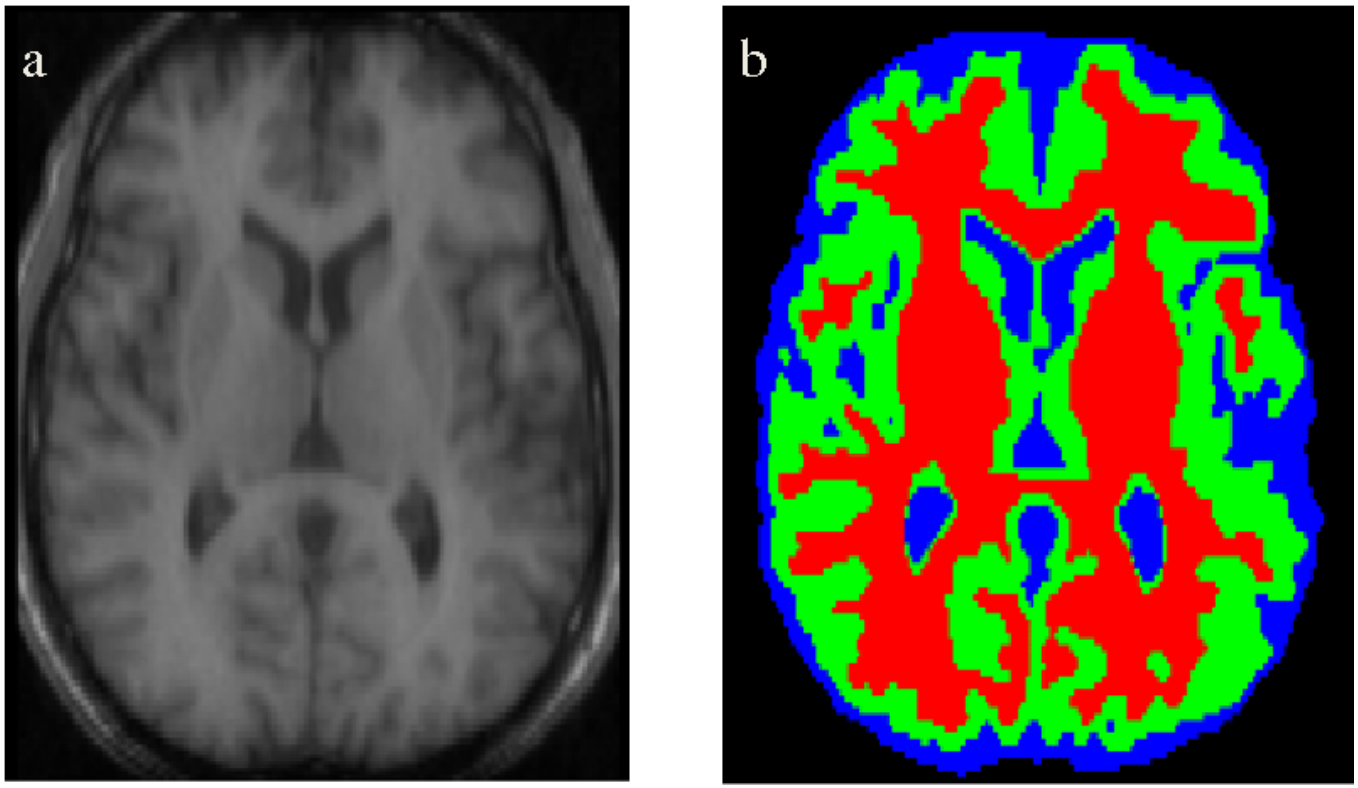
Where do we apply semantic segmentation?

Remote sensing images



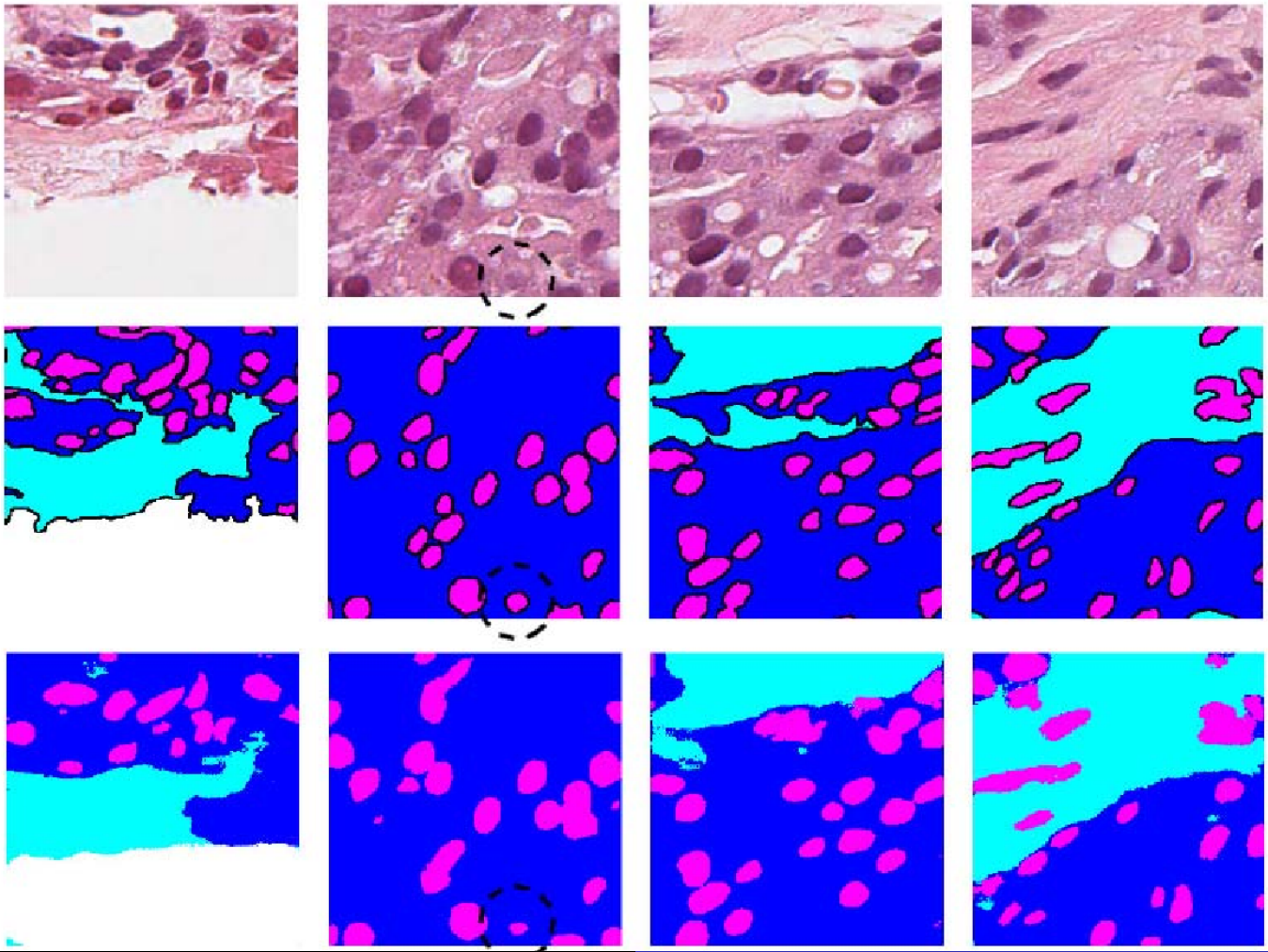
Where do we apply semantic segmentation?

Medical images **[Withey2008]**



Where do we apply semantic segmentation?

microscopic images **[IsakssonIJCNN2017]**

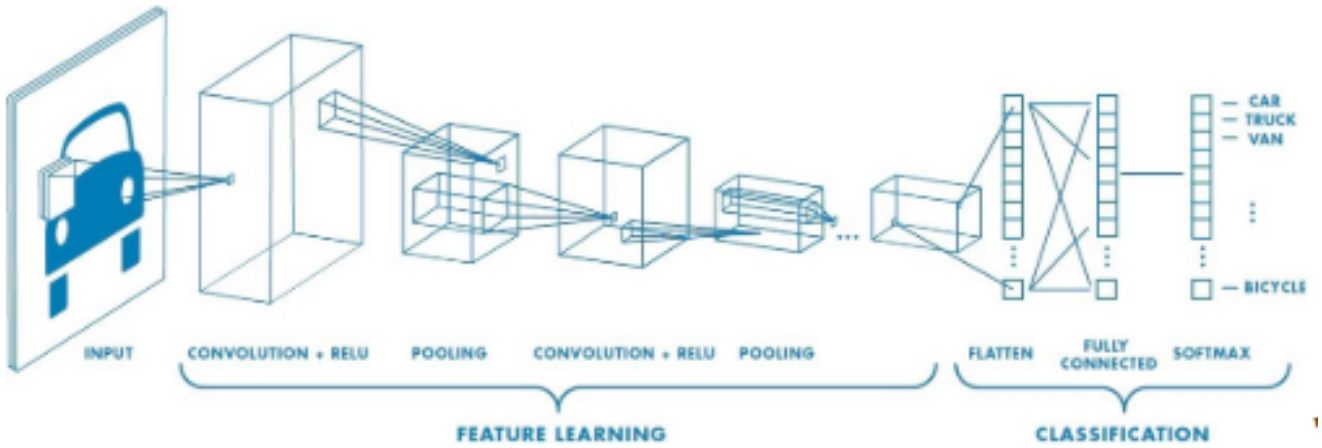


Traditional semantic segmentation

- Pixel based
- Superpixel based

How to do Semantic segmentation with DL?

Remember how a Deep Neural network works?

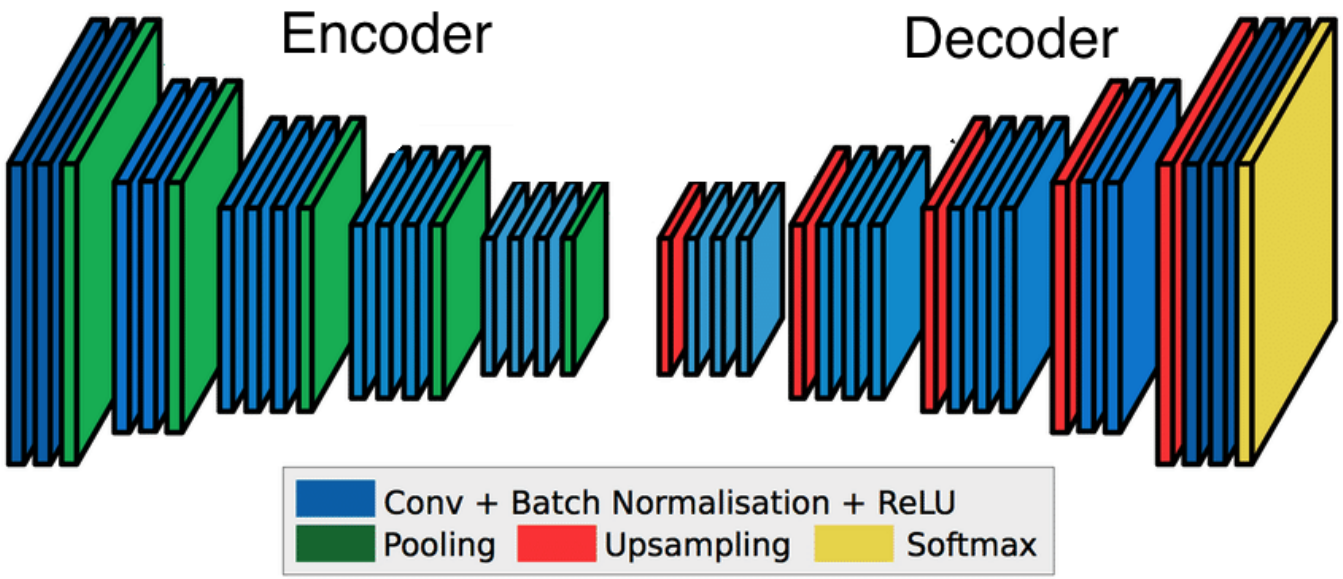


What are the problems?

- the classical DNN just encode the image
- we loose all the spatial information
- we must be able to handle change on the spatial size of images

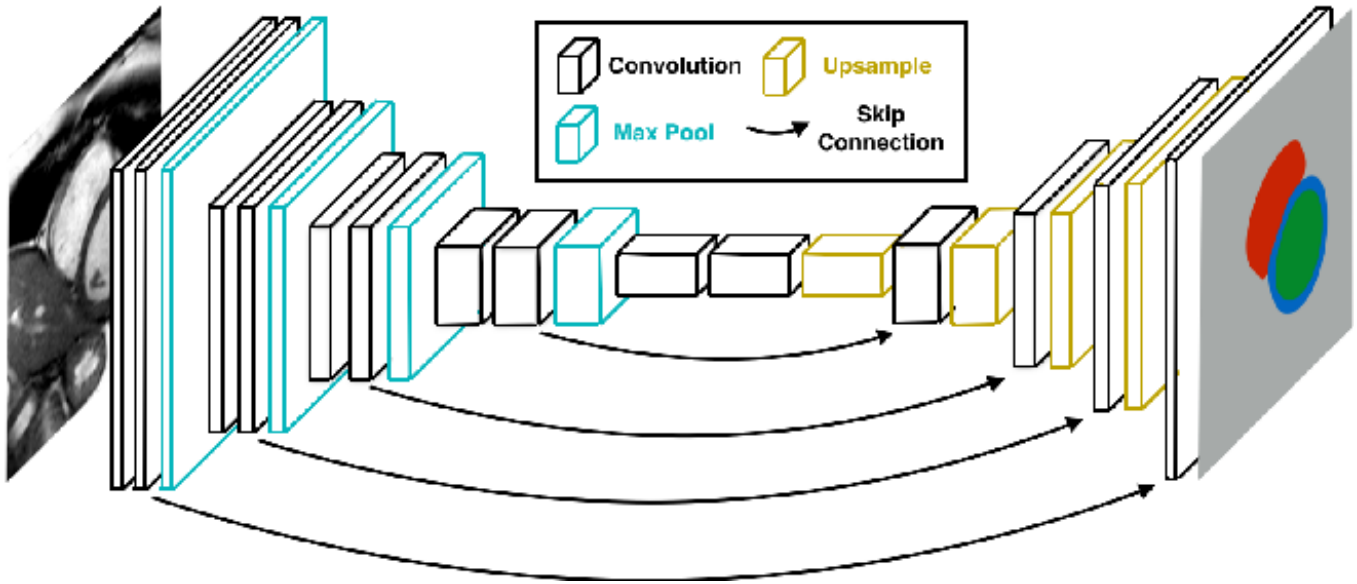
How to do Semantic segmentation with DL?

Solution : An auto encoder composed of a **Fully Convolutional Network** (FCN)



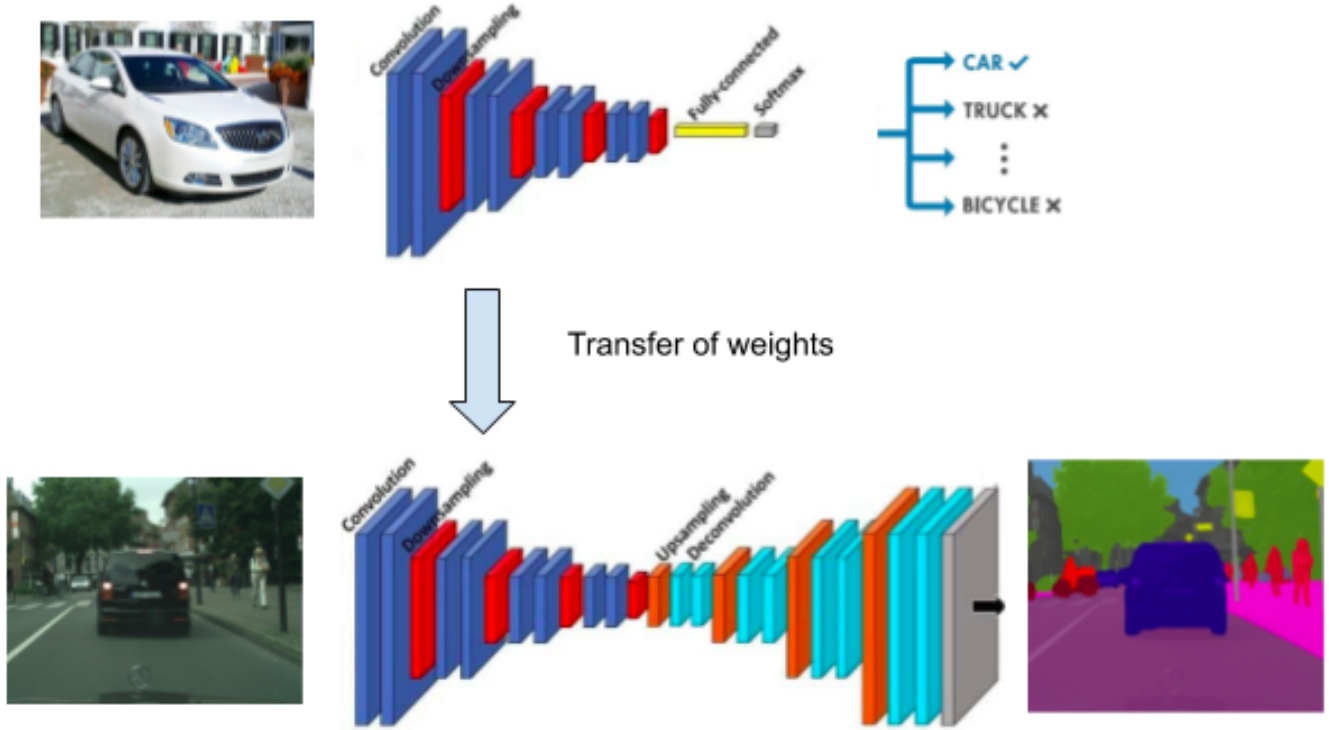
How to do Semantic segmentation with DL?

We might loose spatial and low-level information. A solution use Skip connection



How to do Semantic segmentation with DL?

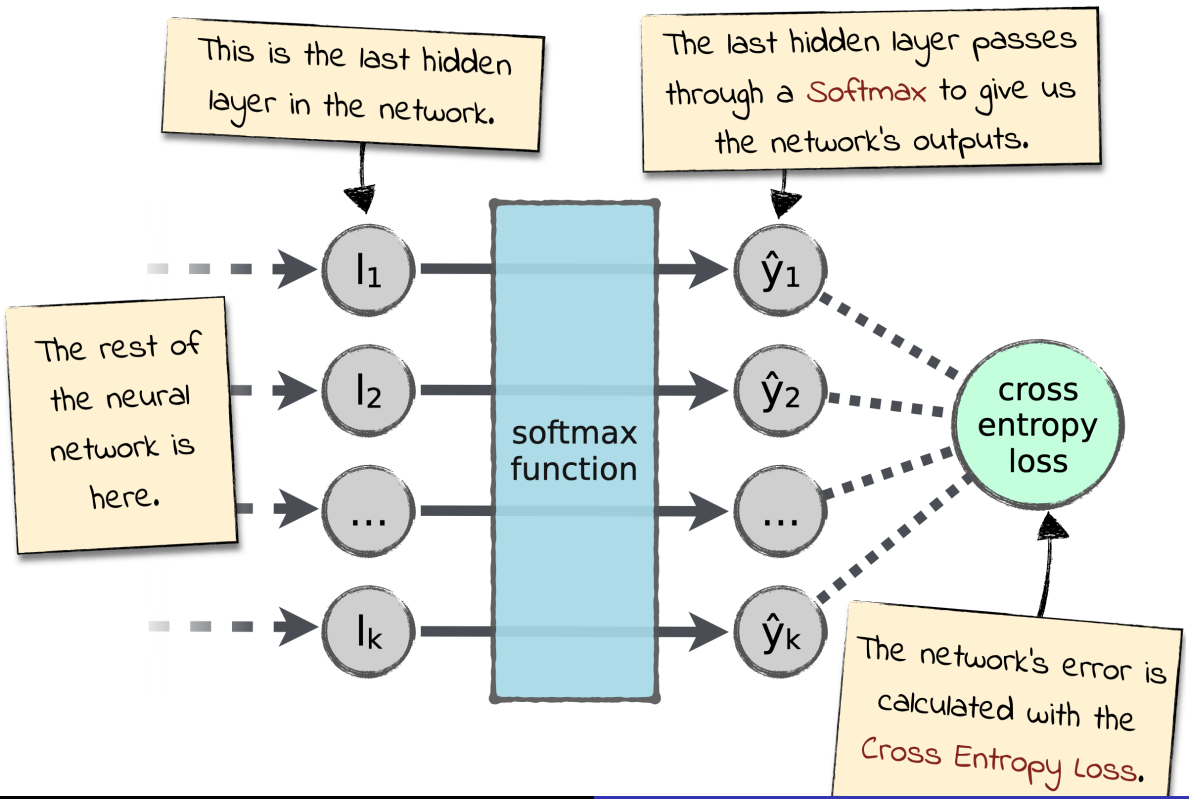
Training a DNN from scratch might be problematic
A solution use **pre-trained** DNNs



Which loss do we choose?

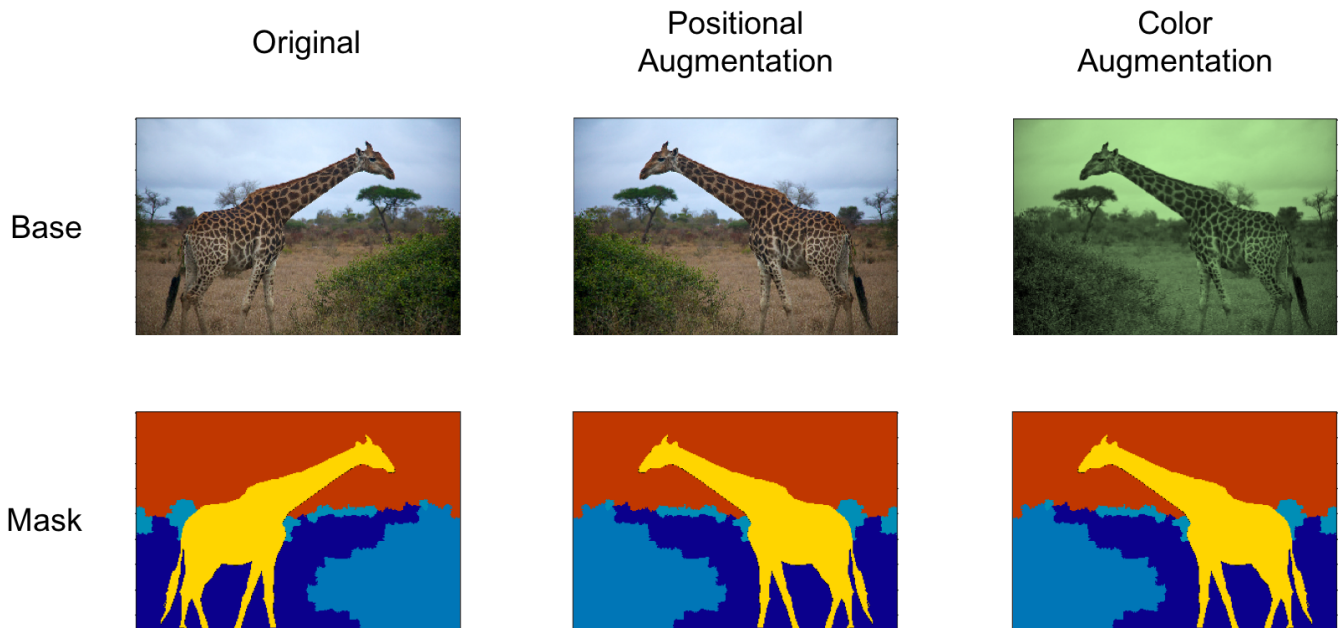
A solution use the **cross entropy loss**.

$$\mathcal{L}(\hat{y}, y) = -\sum_{i}^{C} y_i \log(\hat{y}_i) \tag{1}$$



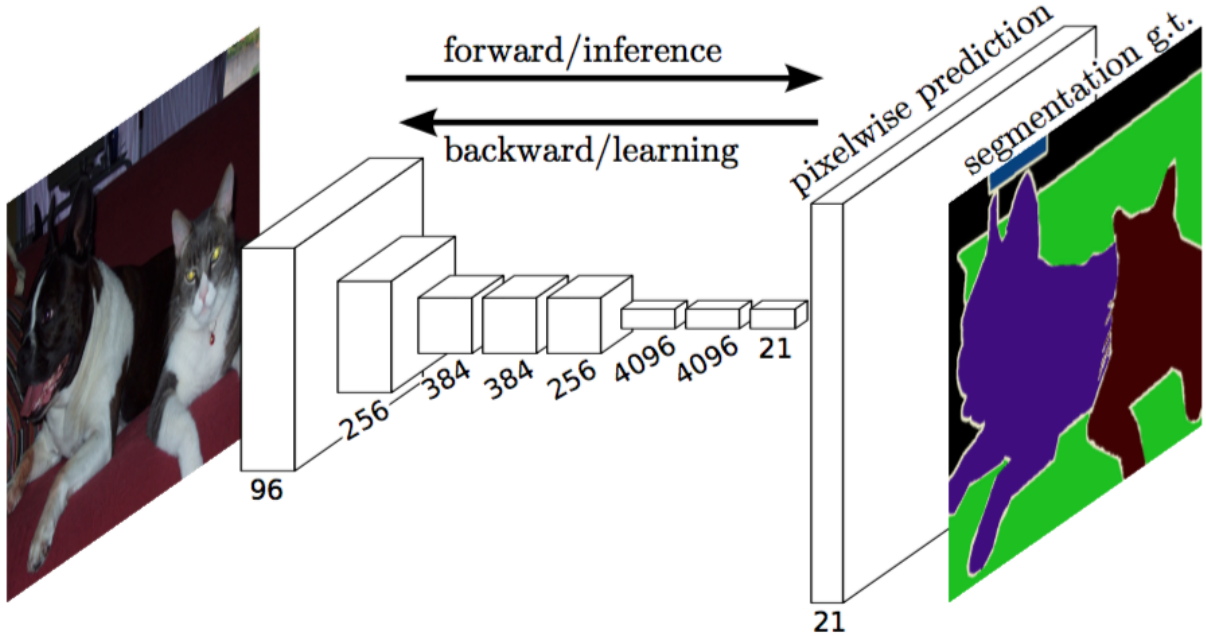
How to train Semantic segmentation with DL?

Training a DNN from scratch might be problematic
A solution use **data augmentations**.



FCN

FCN is the first proposed models for end-to-end semantic segmentation, at the end a big transposed convolution was used.



FCN + CRF

FCN is followed by a CRF to refined the labels.

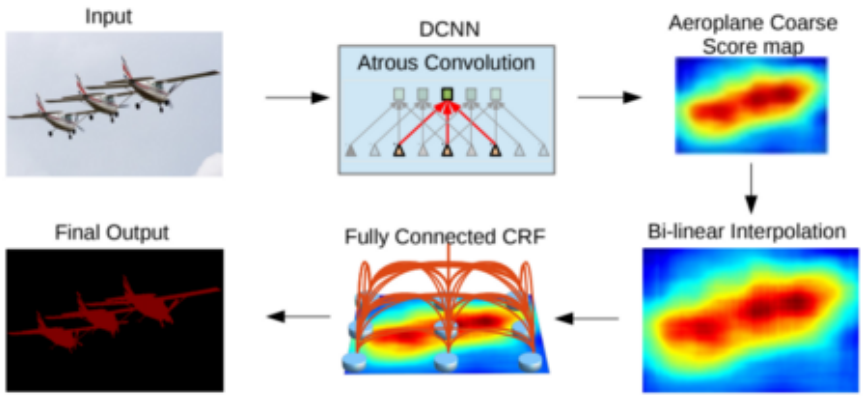
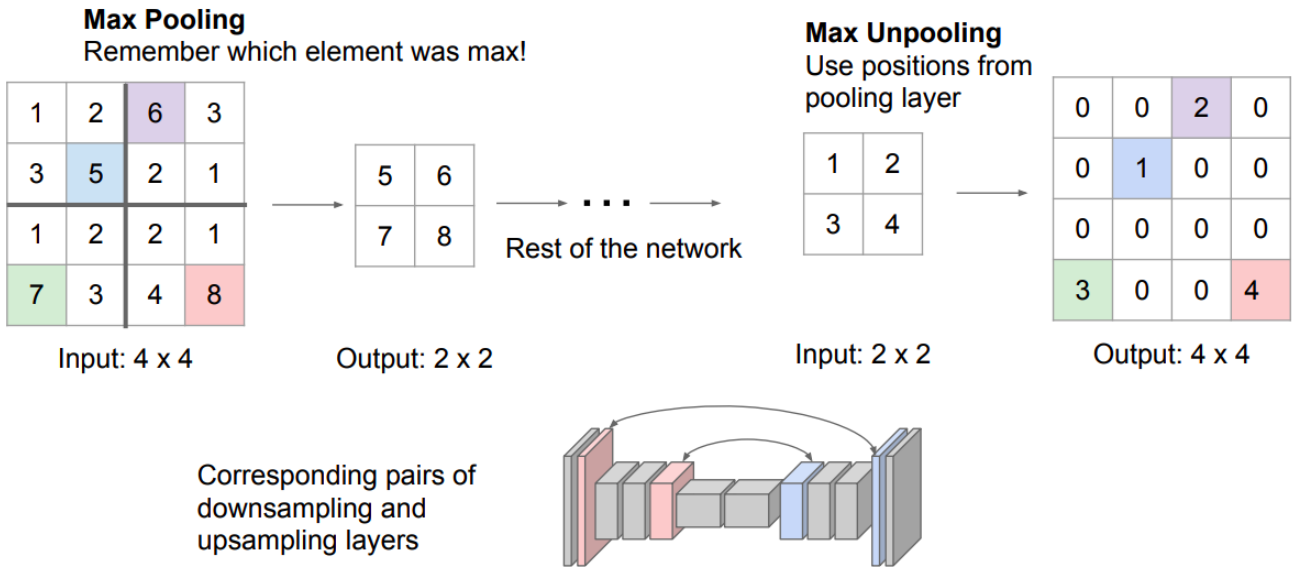


Fig. 1: Model Illustration. A Deep Convolutional Neural Network such as VGG-16 or ResNet-101 is employed in a fully convolutional fashion, using atrous convolution to reduce the degree of signal downsampling (from 32x down 8x). A bilinear interpolation stage enlarges the feature maps to the original image resolution. A fully connected CRF is then applied to refine the segmentation result and better capture the object boundaries.

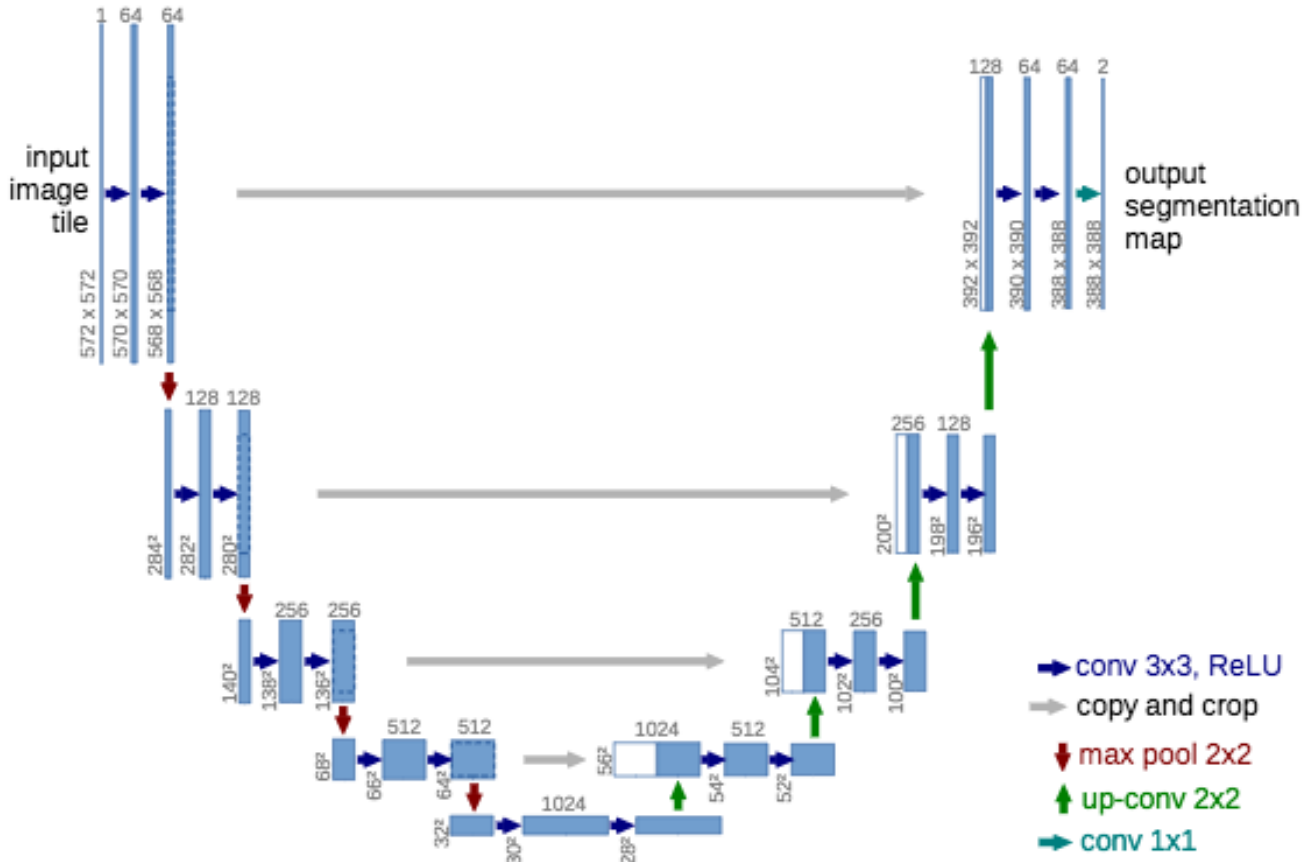
SegNet

For SegNet, the encoder and decoder layers are symmetrical to each other without any skip connections. The upsampling operation of the decoder layers use the max-pooling indices of the corresponding encoder layers. Unlike FCN, no learnable parameters are used for upsampling.



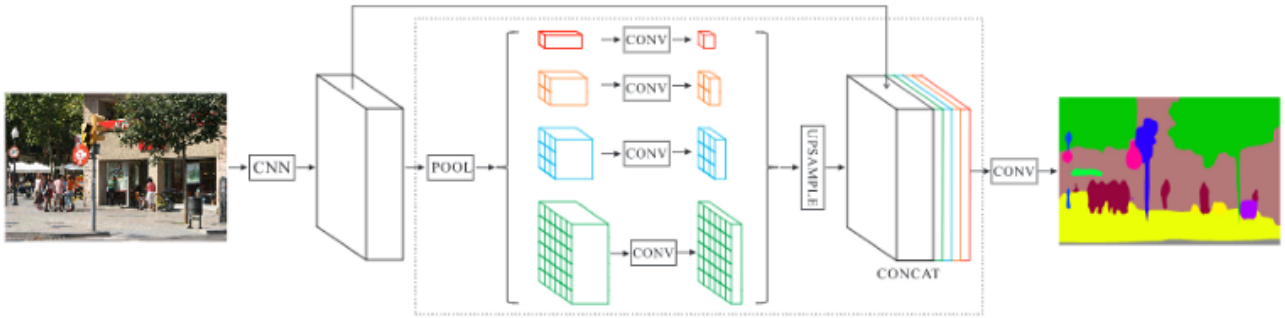
U-Net

For U-Net, the encoder and decoder layers are symmetrical to each other **with skip connections**. The U-Net uses interpolation followed by convolution operation on the decoder.



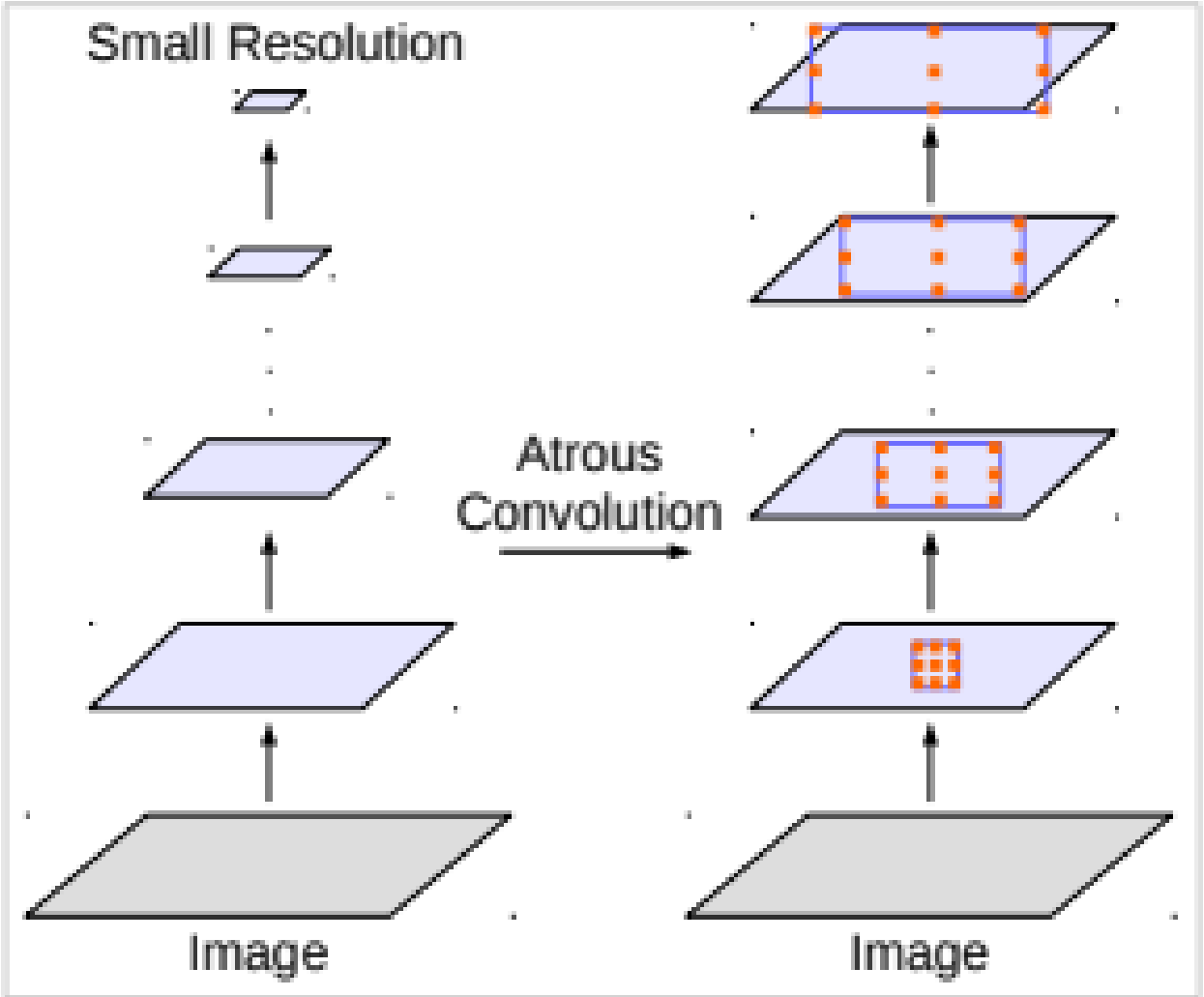
PSPNet

For PSPNet, first use encoder CNN to get the feature map of the last convolutional laye. Then, it uses a pyramid parsing module to improve the representation. This is followed by upsampling and concatenation layers to form the final feature representation.

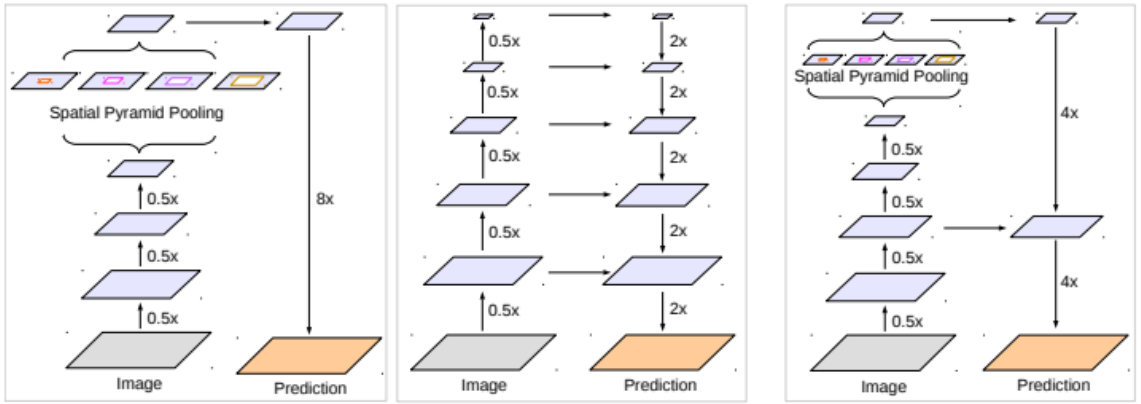


(a) Input Image (b) Feature Map (c) Pyramid Pooling Module (d) Final Prediction

Deeplab



Deeplab v3+



(a) Spatial Pyramid Pooling (b) Encoder-Decoder (c) Encoder-Decoder with Atrous Conv

Fig. 1. We improve DeepLabv3, which employs the spatial pyramid pooling module (a), with the encoder-decoder structure (b). The proposed model, DeepLabv3+, contains rich semantic information from the encoder module, while the detailed object boundaries are recovered by the simple yet effective decoder module. The encoder module allows us to extract features at an arbitrary resolution by applying atrous convolution.

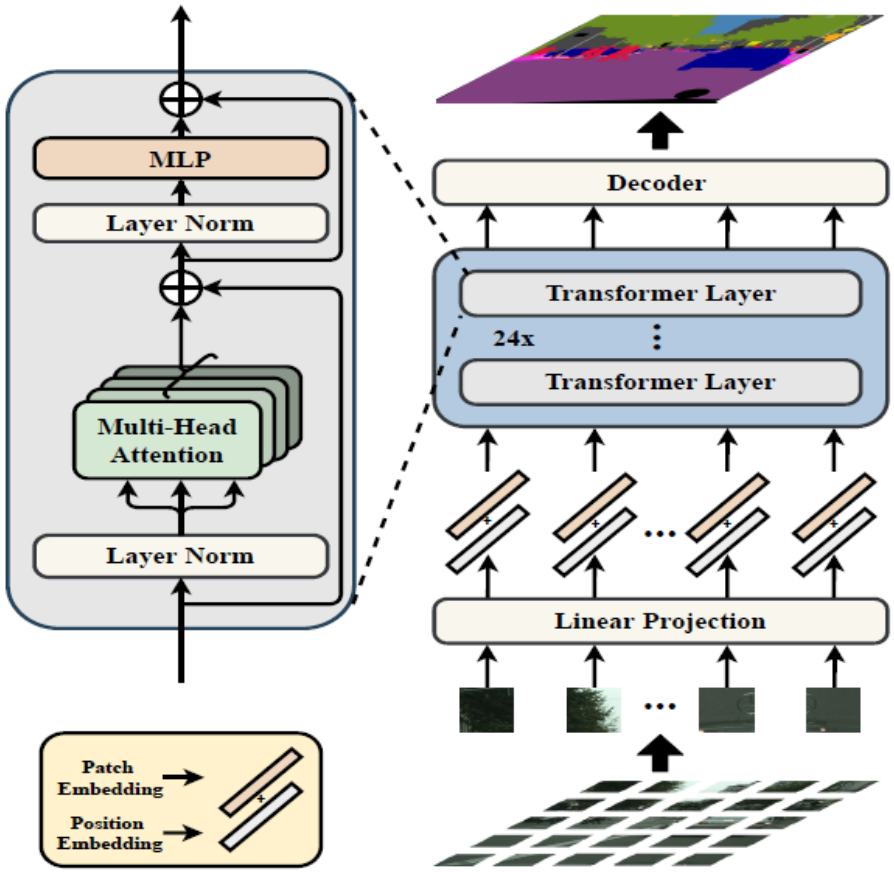
SEgmentation TRansformers (SETR) [ZhengCVPR2021]

A standard Semantic segmentation model has an **encoder-decoder** architecture:

- **encoder**: for feature representation learning
- **decoder**: for pixel-level classification of the feature representations yielded by the encoder

They propose to use for the encoder a DNN inspired by VIT. For the decoder they propose 3 techniques.

SEgmentation TRansformers (SETR) [ZhengCVPR2021]



(a)

SEgmentation TRansformers (SETR) **ENCODER** [ZhengCVPR2021]

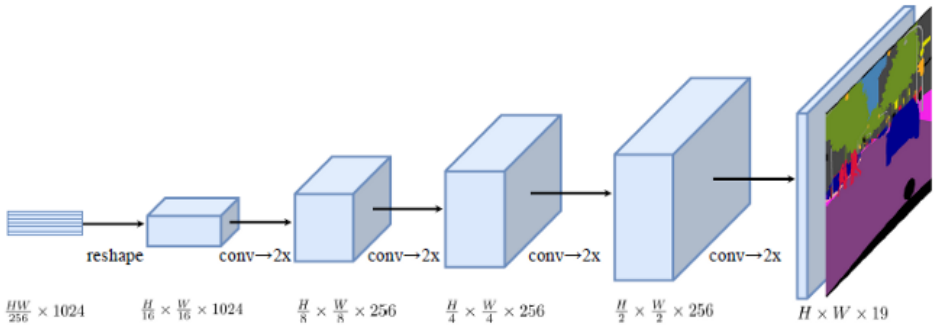
typicaly the encoder is designed for semantic segmentation would downsample a 2D image $x \in \mathrm{R}^{H \times W \times 3}$ into a feature map $x - f \in \mathrm{R}^{\frac{H}{16} \times \frac{W}{16} \times C}$ with C the number of channels. They decided to set the transformer input sequence length L to $\frac{H}{16} \times \frac{W}{16}$

SEgmentation TRansformers (SETR) **DECODER** [ZhengCVPR2021]

The decoder is a function $(\frac{H}{16} \times \frac{W}{16} \times C) \rightarrow (H \times W \times K)$, with K the number of classes. They propose 3 kind of decoder:

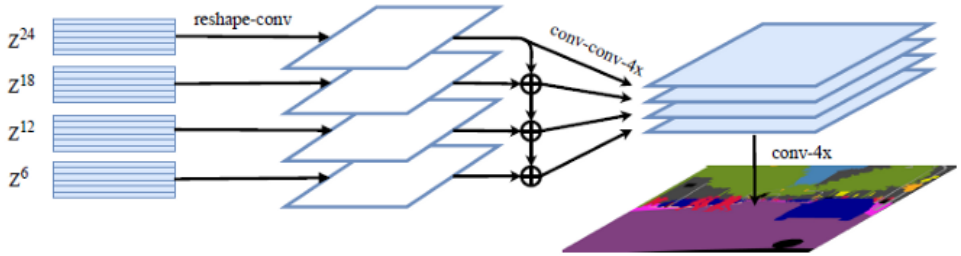
- Naive upsampling (Naive) : 1×1 conv + sync batch norm (w/ ReLU) + 1×1 conv + bilinear upsample
- Progressive UPsampling (PUP) : Not just on big upsample but a progressive upsampling strategy that alternates conv layers and upsampling operations.
- Multi-Level feature Aggregation (MLA) : a progressive upsampling strategy + multi-level feature aggregation.

SEgmentation TRansformers (SETR) PUP DECODER [ZhengCVPR2021]



(b)

SEgmentation TRansformers (SETR) MLA DECODER [ZhengCVPR2021]



(c)

SEgmentation TRansformers (SETR) [ZhengCVPR2021]

Method	Pre	Backbone	#Params	40k	80k
FCN [38]	1K	R-101	68.59	73.93	75.52
Semantic FPN [38]	1K	R-101	47.51	-	75.80
Hybrid-Base	R	T-Base	112.59	74.48	77.36
Hybrid-Base	21K	T-Base	112.59	76.76	76.57
Hybrid-DeiT	21K	T-Base	112.59	77.42	78.28
SETR-*Naïve*	21K	T-Large	305.67	77.37	77.90
SETR-*MLA*	21K	T-Large	310.57	76.65	77.24
SETR-*PUP*	21K	T-Large	318.31	78.39	79.34
SETR-*PUP*	R	T-Large	318.31	42.27	-
SETR-*Naïve-Base*	21K	T-Base	87.69	75.54	76.25
SETR-*MLA-Base*	21K	T-Base	92.59	75.60	76.87
SETR-*PUP-Base*	21K	T-Base	97.64	76.71	78.02
SETR-*Naïve-DeiT*	1K	T-Base	87.69	77.85	78.66
SETR-*MLA-DeiT*	1K	T-Base	92.59	78.04	78.98
SETR-*PUP-DeiT*	1K	T-Base	97.64	**78.79**	**79.45**

Table 2. **Comparing SETR variants** on different pre-training strategies and backbones. All experiments are trained on Cityscapes train fine set with batch size 8, and evaluated using the single scale test protocol on the Cityscapes validation set in mean IoU (%) rate. "Pre" denotes the pre-training of transformer part. "R" means the transformer part is randomly initialized.

Bibliography

- **[web1]** https://heartbeat.fritz.ai/a-2019-guide-to-semantic-segmentation-ca8242f5a7fc
- **[web2]** https://heartbeat.fritz.ai/a-2019-guide-to-semantic-segmentation-ca8242f5a7fc
- **[Withey2008]** Withey, Daniel J., and Zoltan J. Koles. "A review of medical image segmentation: methods and available software." International Journal of Bioelectromagnetism 10.3 (2008): 125-148.
- **[IsakssonIJCNN2017]** Isaksson, Johan, et al. "Semantic segmentation of microscopic images of H&E stained prostatic tissue using CNN." 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017.

Bibliography

 **[ZhengCVPR2021]** Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., ... & Zhang, L. (2021). Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6881-6890).