Segmentare Semantica

Curs 5

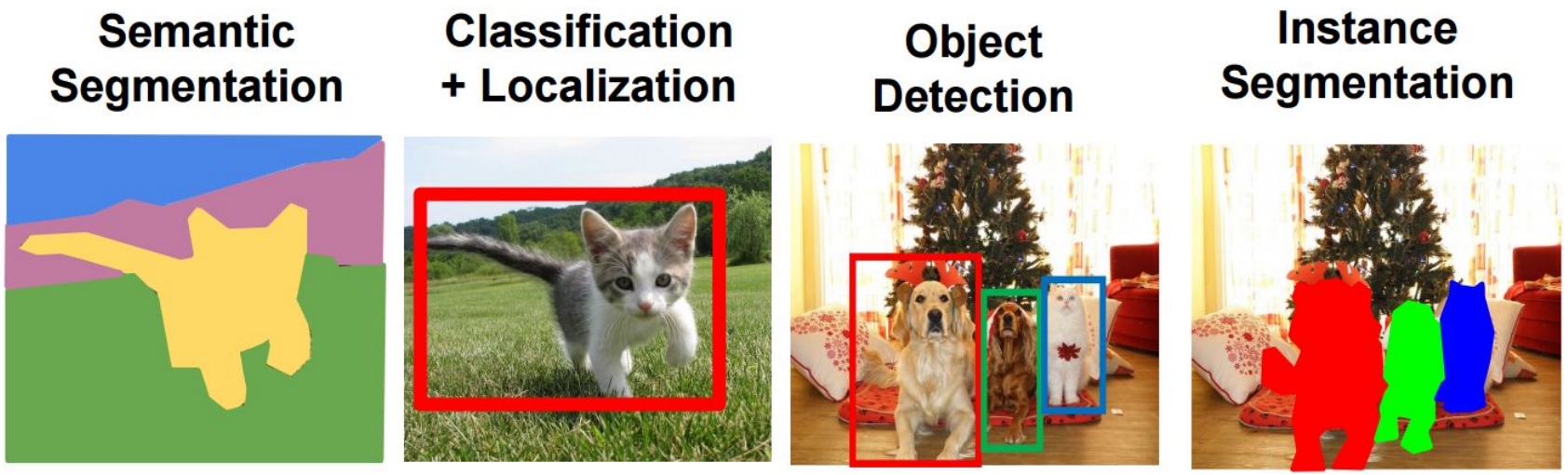
Recapitulare Task Clasificare

- Clasificare un obiect/categorie per imagine.
- Obiectul poate fi oriunde in poza, in orice pozitie. dar preferabil in foreground, de dimensiune mare
- CrossEntropy loss
- Etichete reprezentate ca One-Hot-Encodings.
- 1 sau mai multe straturi FC la sfarsitul retelei.
- Posibil sa folosim Global Average Pooling in loc de Flatten.

Limitari:

- Ce se intampla daca sunt mai multe obiecte in poza?
- Ce alta informatie ne intereseaza, legat de instanta fiecarui obiect?
- Ce se intampla daca obiectele se suprapun?

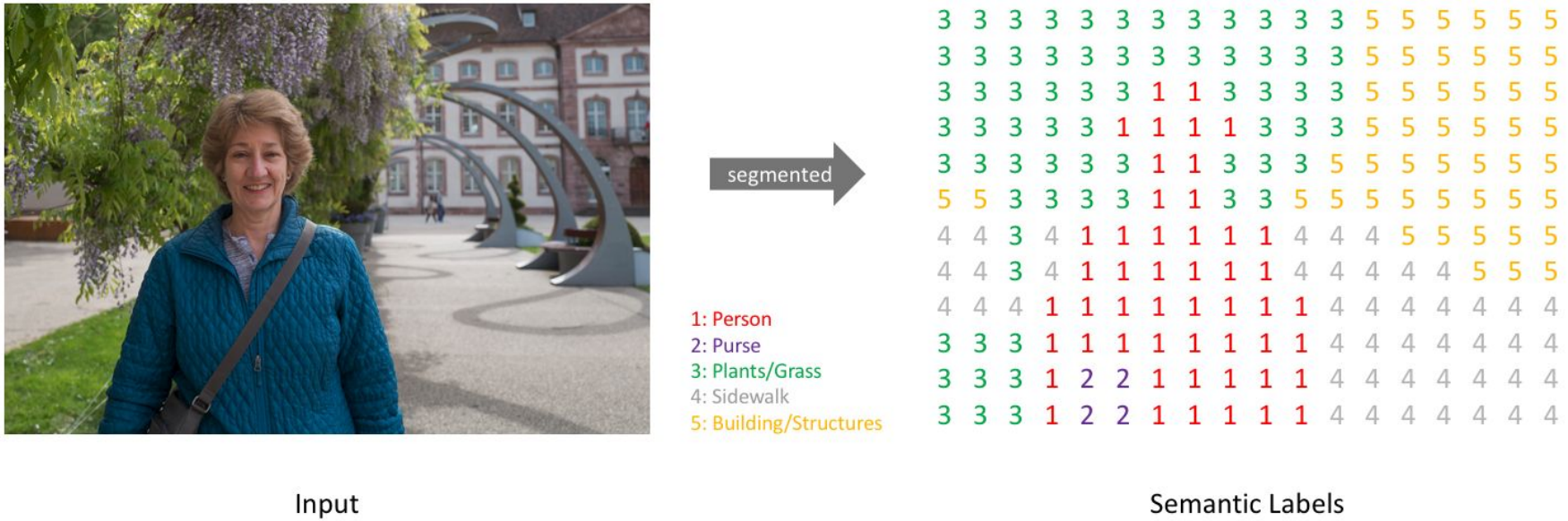
Urmatoarele Taskuri in CV



http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Segmentare Semantica

- Fiecare pixel este etichetat
- Nu diferentiaza intre instantele claselor
- Segmentare: pixelwise classification



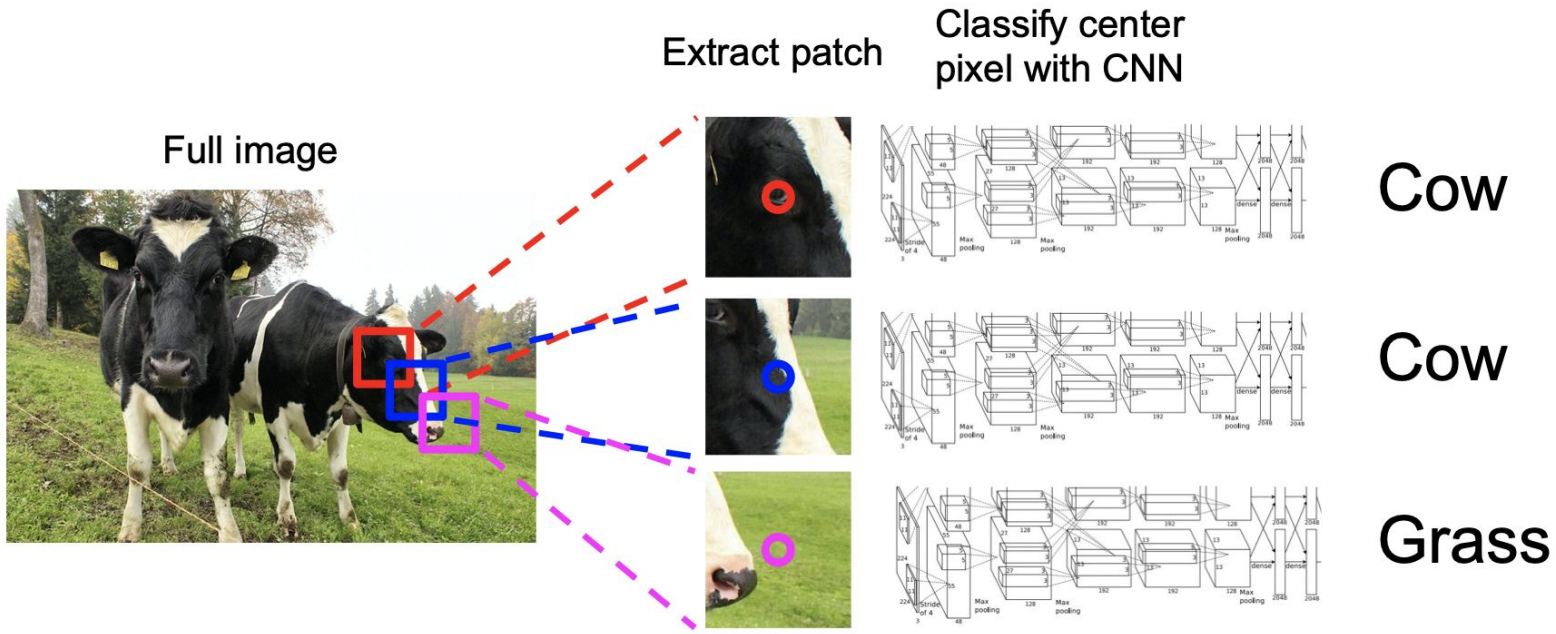
Input

Semantic Labels

https://www.jeremyjordan.me/semantic-segmentation/

Segmentare Semantica

Segmentare Semantica - Idee

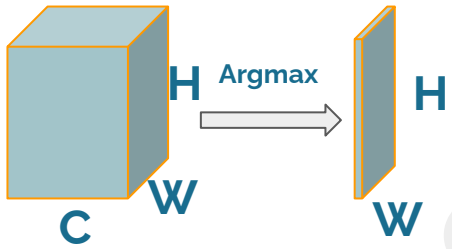


Farabet et al, “Learning Hierarchical Features for Scene Labeling,” TPAMI 2013
Pinheiro and Collobert, “Recurrent Convolutional Neural Networks for Scene Labeling”, ICML 2014

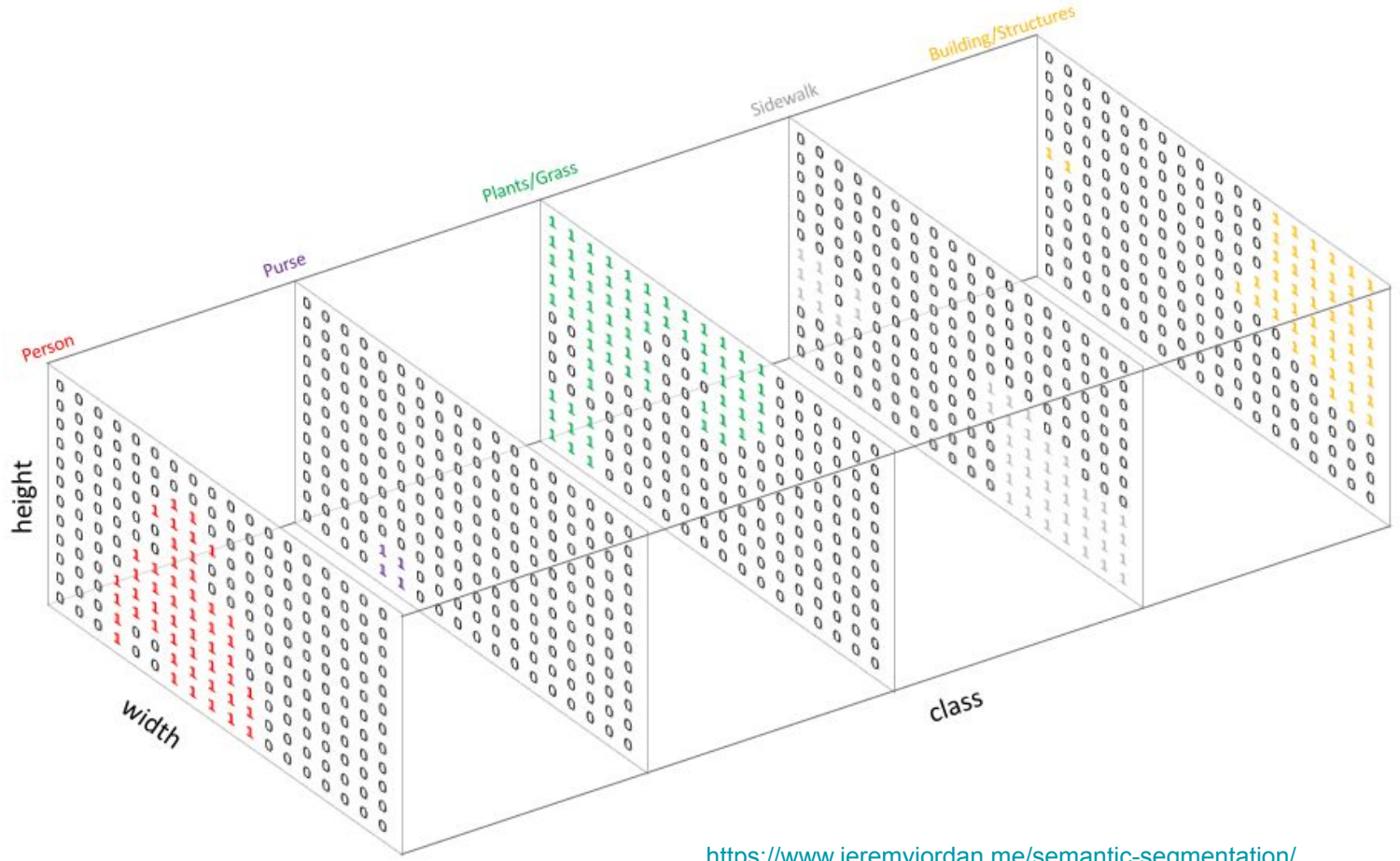
http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Segmentare Semantica - Intuitie

- Avem o retea de clasificare binara: 0/1 per imagine
- **Extindem la 0/1 per pixel (segmentare binara)**
- Convoluțiile păstrează ordinea spatiala, este intuitiv sa folosim doar convolutii pentru **dense classification** (segmentation)
- C clase: output-ul un volum CxHxW
- Softmax este aplicat pixelwise in dimensiunea C



Segmentare Semantica - Intuitie



https://www.jeremyjordan.me/semantic-segmentation/

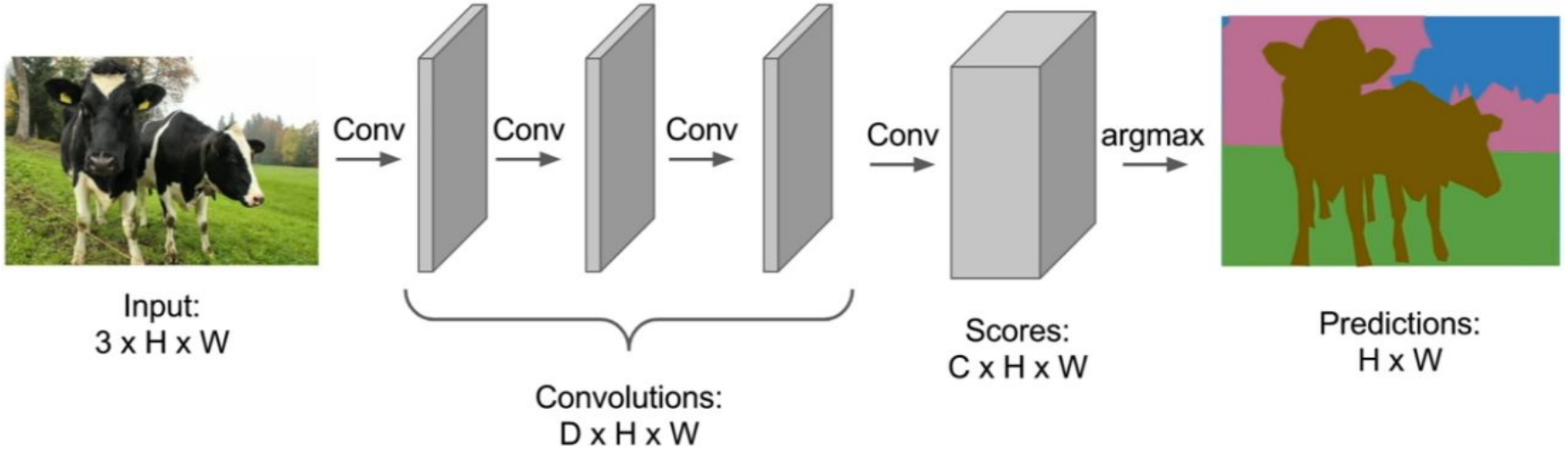
Segmentare Semantica - Intuitie

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

https://www.jeremyjordan.me/semantic-segmentation/

Segmentare Semantica - Idee

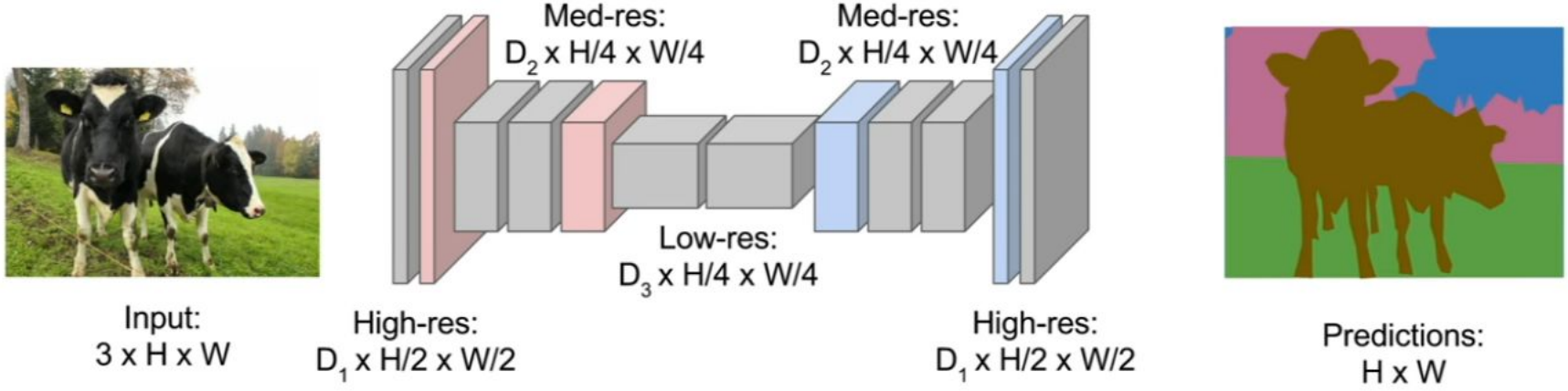
- Doar straturi convolutionale: Fully Convolutional Neural Network (FCN)
- Care este problema in cazul acesta**?**



http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

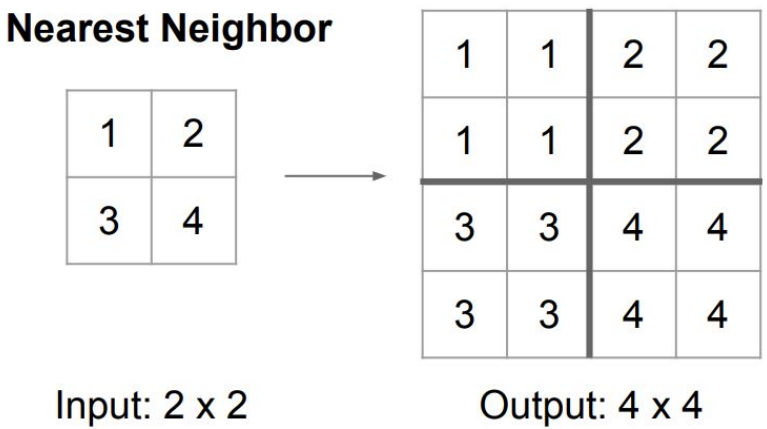
Segmentare Semantica

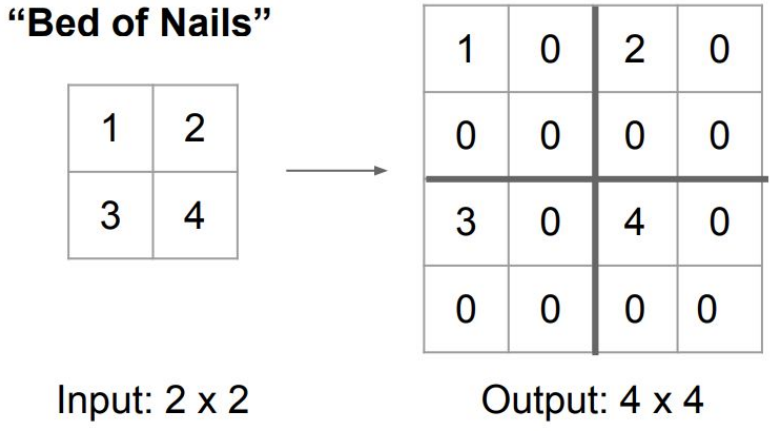
- In practica, este ineficient sa marim numarul de canale pastrand rezolutia spatiala a input-ului.
- Arhitectura Clepsidra (Hourglass) / Encoder-Decoder
  - Input Image -> Downsampling -> Upsampling -> Output Image Mask



http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

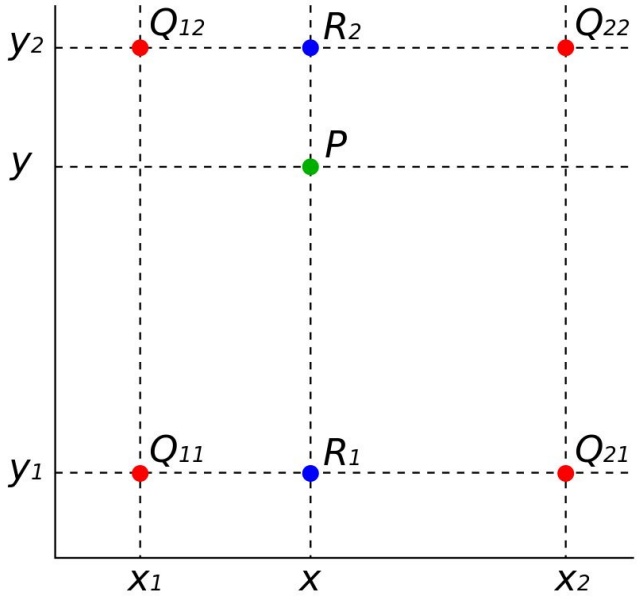
Upsampling





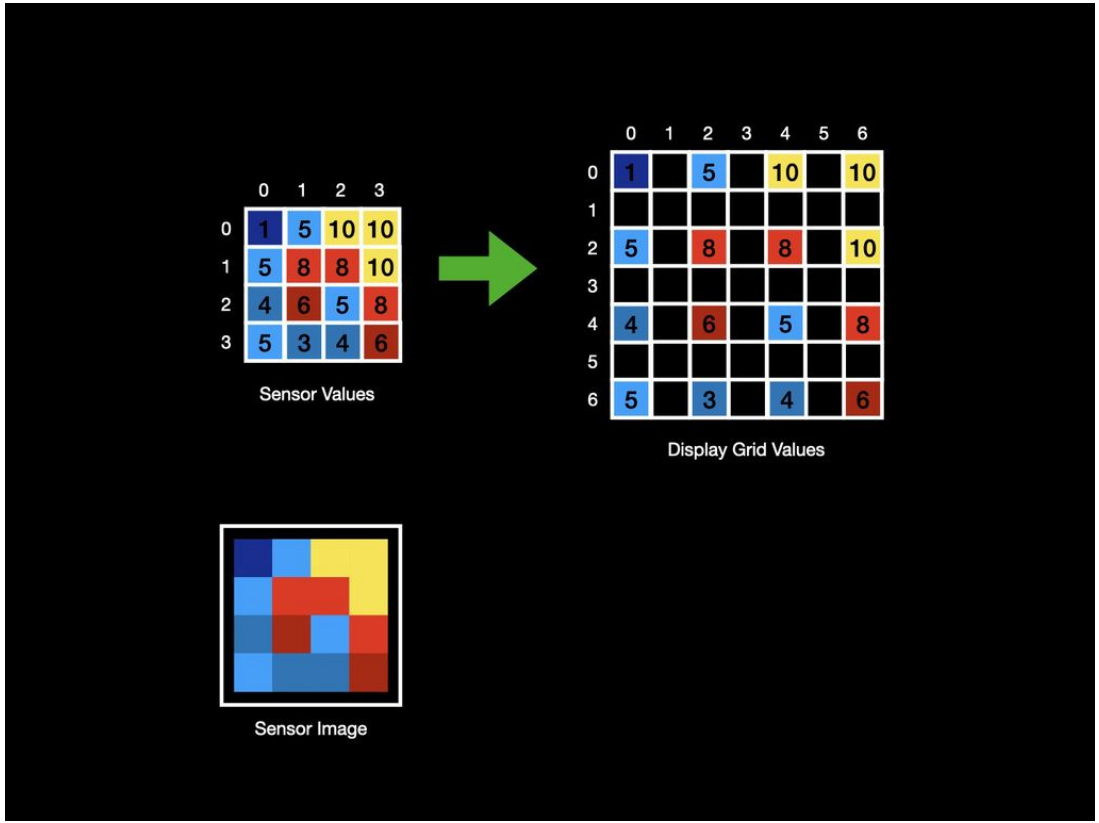
http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Upsampling: Bilinear Interpolation



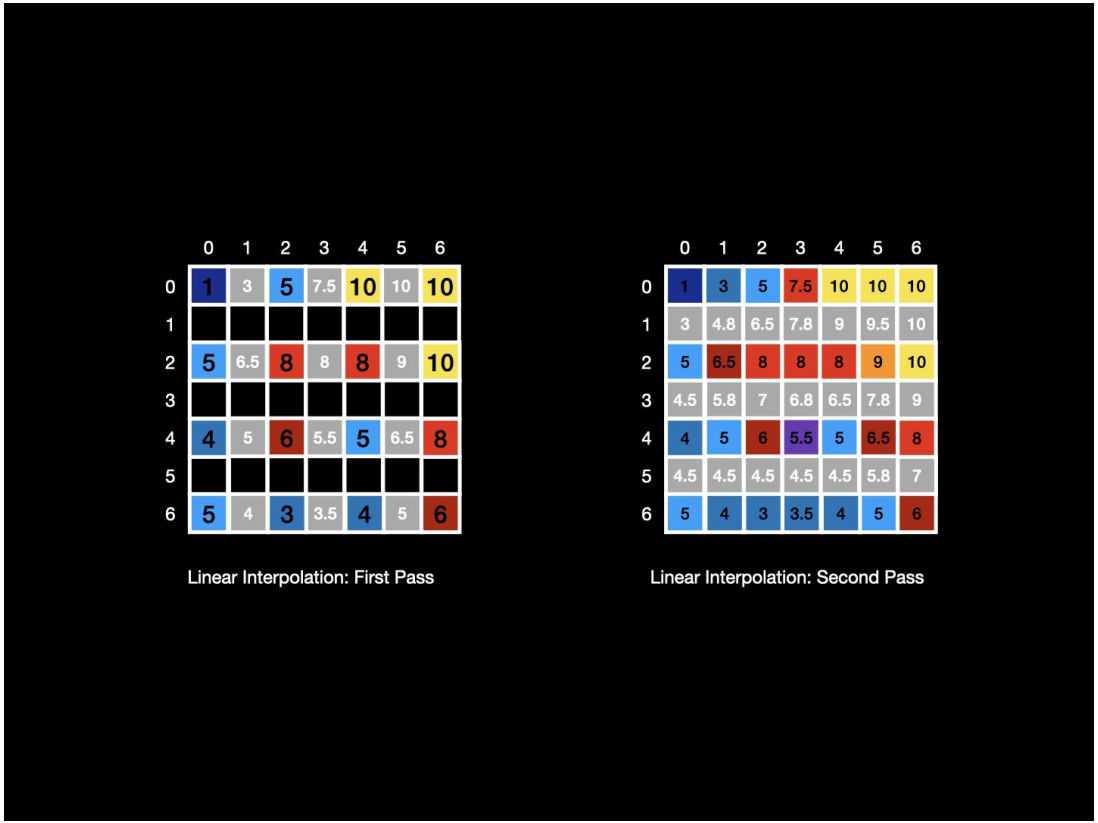
https://en.wikipedia.org/wiki/Bilinear_interpolation

Upsampling: Bilinear Interpolation



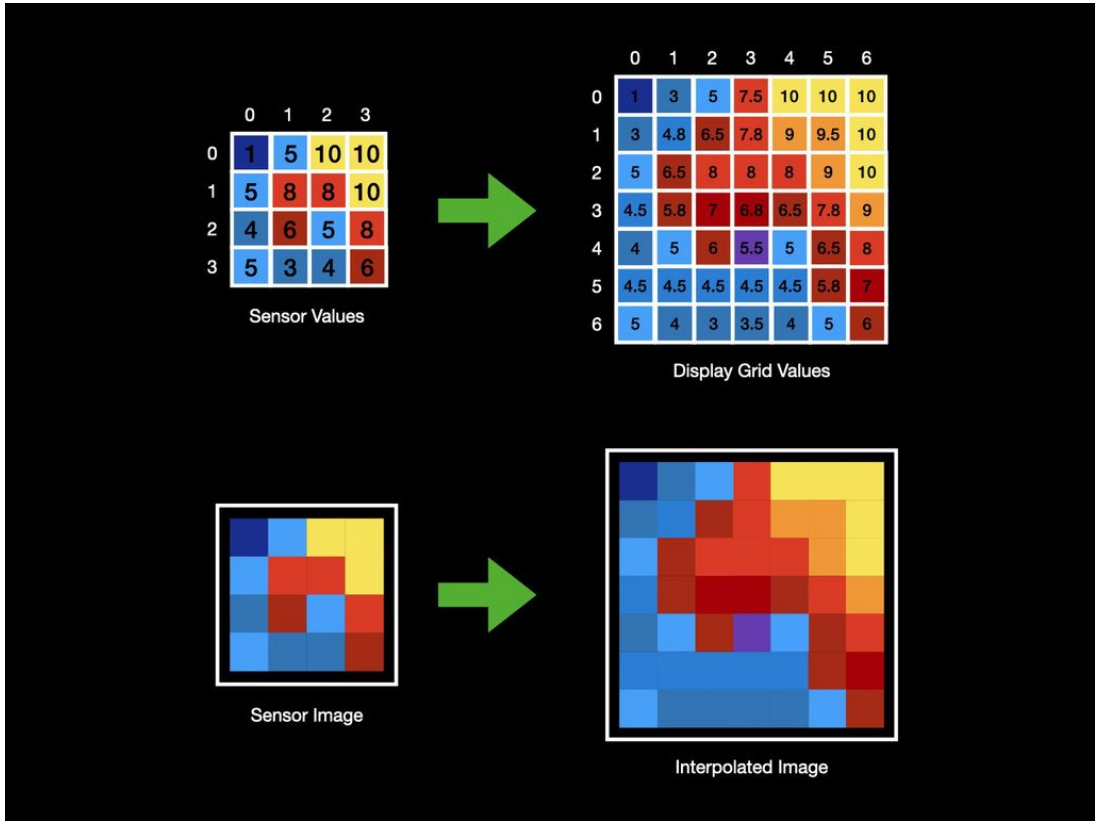
https://learn.adafruit.com/assets/102695

Upsampling: Bilinear Interpolation



https://learn.adafruit.com/assets/102696

Upsampling: Bilinear Interpolation



https://learn.adafruit.com/assets/102697

Upsampling: Max Unpooling

Max Pooling
Remember which element was max!

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

Input: 4 x 4

5	6
7	8

Output: 2 x 2

Rest of the network

Max Unpooling
Use positions from pooling layer

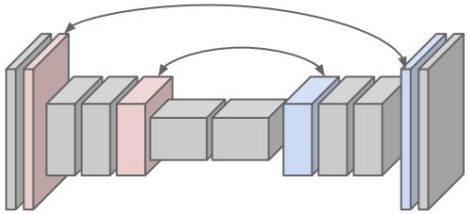
1	2
3	4

Input: 2 x 2

0	0	2	0
0	1	0	0
0	0	0	0
3	0	0	4

Output: 4 x 4

Corresponding pairs of downsampling and upsampling layers



http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture11.pdf

Segmentare Semantica: SegNet

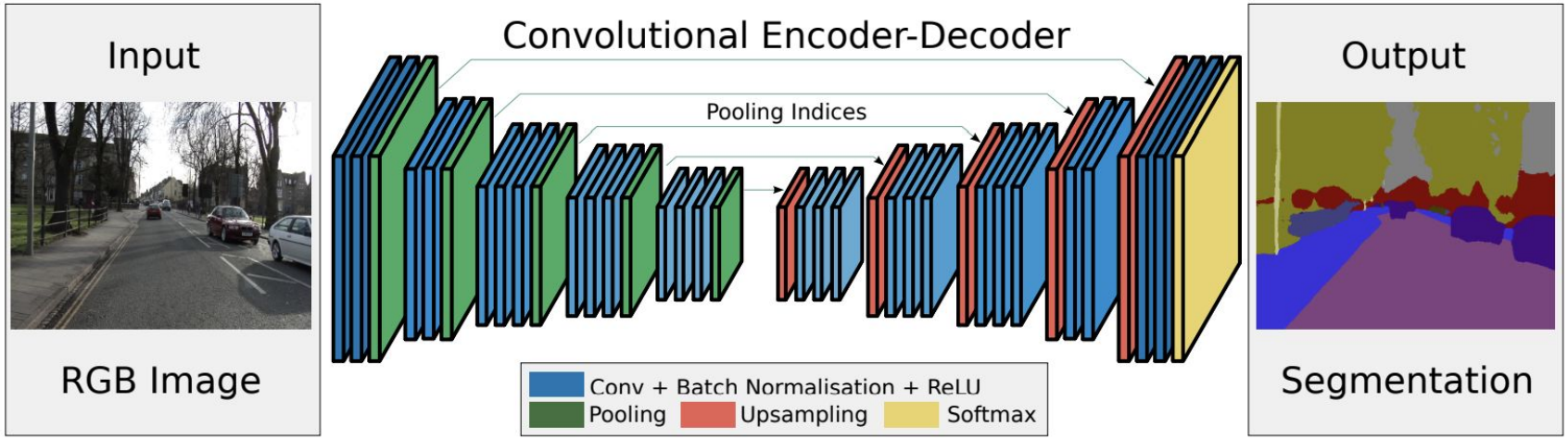


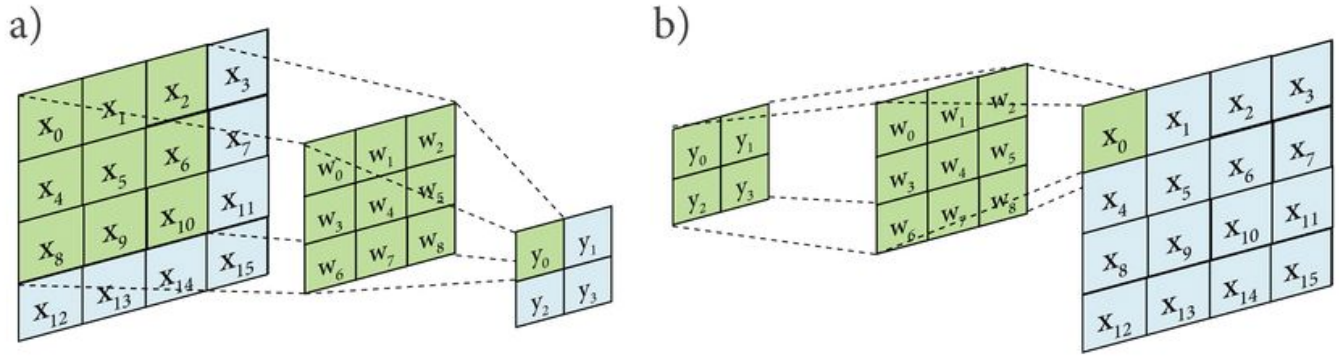
Fig. 2. An illustration of the SegNet architecture. There are no fully connected layers and hence it is only convolutional. A decoder upsamples its input using the transferred pool indices from its encoder to produce a sparse feature map(s). It then performs convolution with a trainable filter bank to densify the feature map. The final decoder output feature maps are fed to a soft-max classifier for pixel-wise classification.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *39*(12), 2481-2495.

Upsampling: Convolutia Transpusa

Upsampling cu parametri invatabili : Convolutie Transpusa

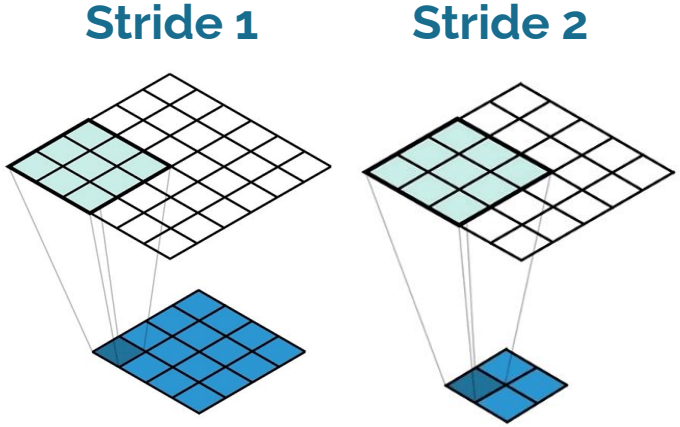
- In forward pass avem convolutie normala. In backward pass avem convolutie transpusa .
- Face upsampling la feature-map invatand prin parametrii unei convolutii felul in care trebuie sa distribuie, ponderat, valorile pentru urmatorul feature-map.
- Valorile filtrului W sunt inmultite cu valoarea de input din feature map => In output vor exista copii ponderate ale W.
- Acolo unde aceste copii se suprapun, valorile W*pondere se aduna.
- Se mai numeste si “Deconvolution”/Fractionally Strided Conv



http://deeplearning.net/software/theano/tutorial/conv_arithmetic.html
https://arxiv.org/pdf/1603.07285.pdf
https://github.com/vdumoulin/conv_arithmetic

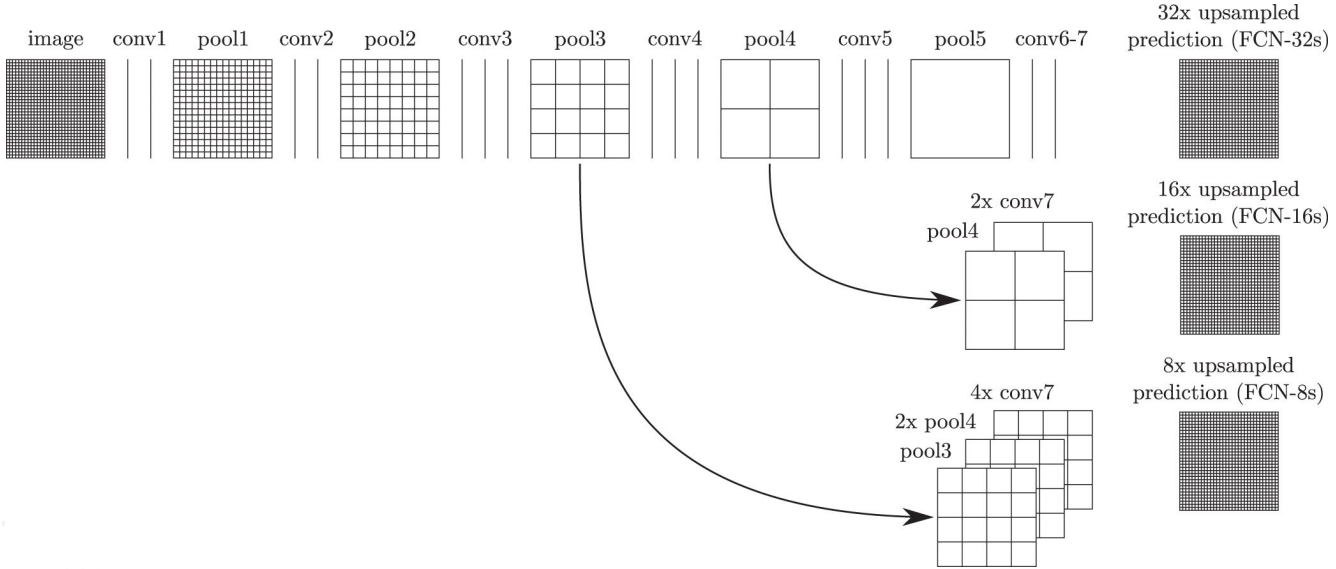
Convolutia Transpusa

- Intuitiv, cu cat o valoare (pixel) e mai aproape de centru in output, cu atat acumuleaza **“fractiuni”** din mai multe valori (pixeli) de input
- Stride-ul se aplica acum in **output**: stride mai mare -> dimensiune mai mare
  - Stride-ul mare produce un output mai mic decat input/s in convolutia directa
- Inversam input/output: input-ul devine output si vice versa:
  - $\mathbf{Input}^T = \mathbf{Output}_C$
  - $\mathbf{Output}^T = \mathbf{Input}_C$
  - Se doreste reconstituirea input-ului care a fost folosit in convolutia directa pentru generara output-ului
- Dimensiuni output: (I - 1) * S + K

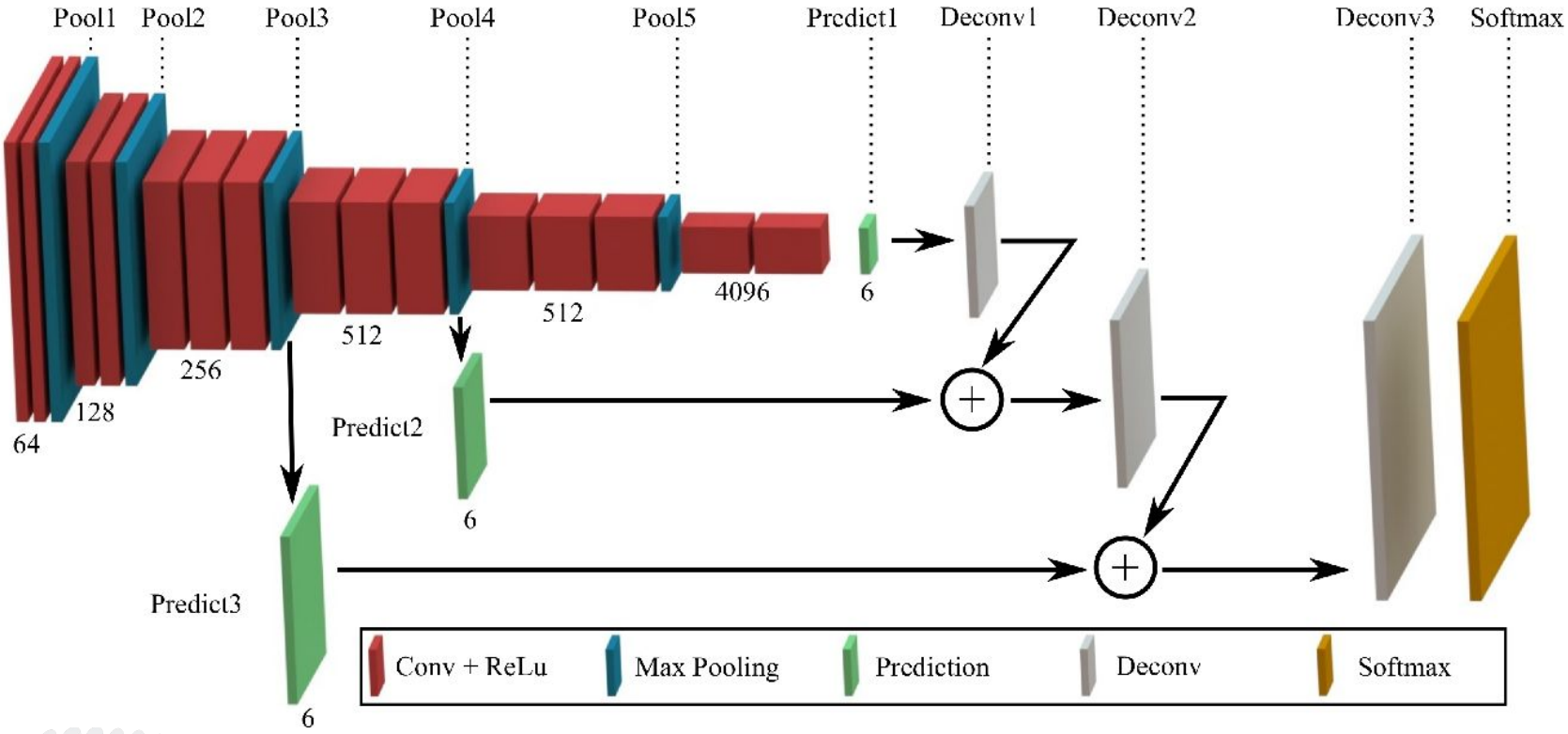


Fully Convolutional Networks - E Shelhamer

- Skip connections
- Upsample from stride 16 to stride 8:
- Use 1x1 bottlenecks to compress to number of classes
- Add with encoder counterpart from stride 8 (1x1 compressed to number of classes)



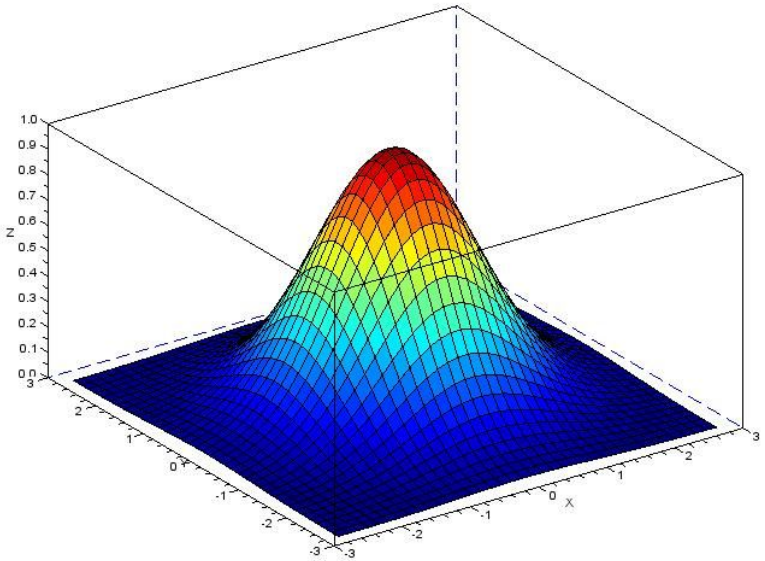
Fully Convolutional Networks



Studiu de caz: Pose Estimation

Reprezentarea unui punct in plan folosind distributie de probabilitate:

- Heatmap
- HRNet
- Task de predictie densa:
  - Regresie pentru fiecare pixel



HRNet

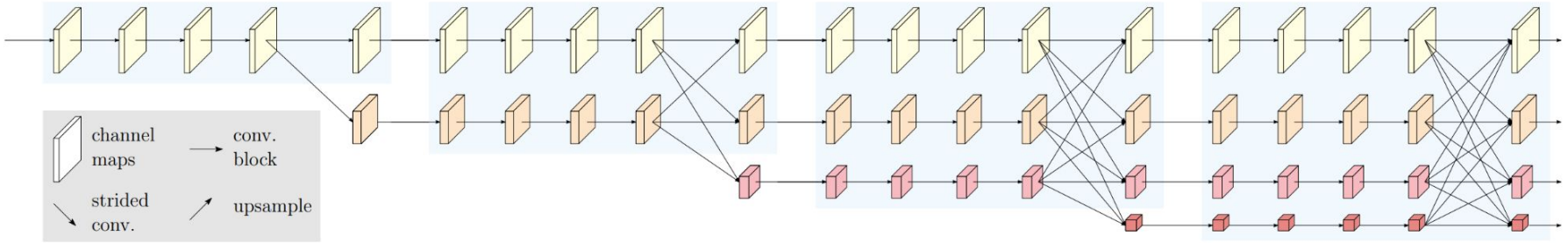
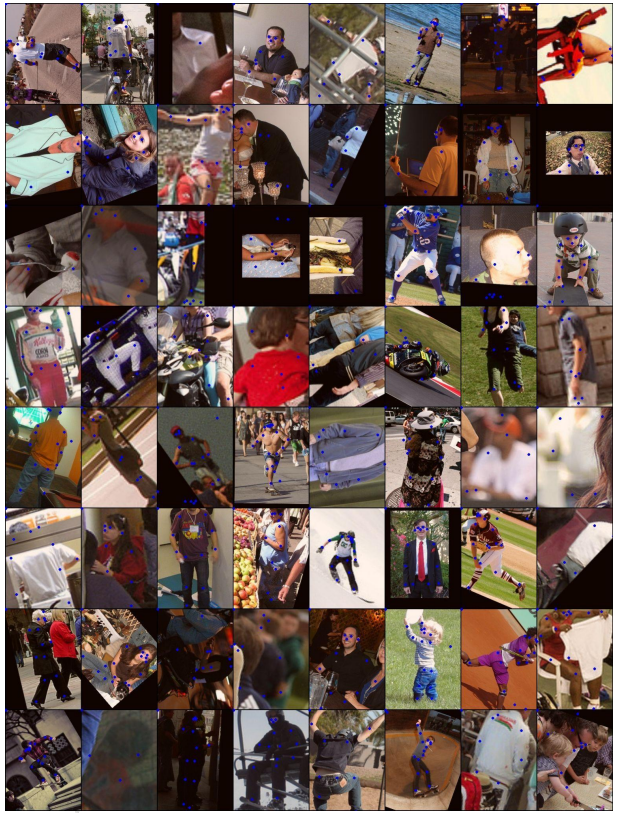


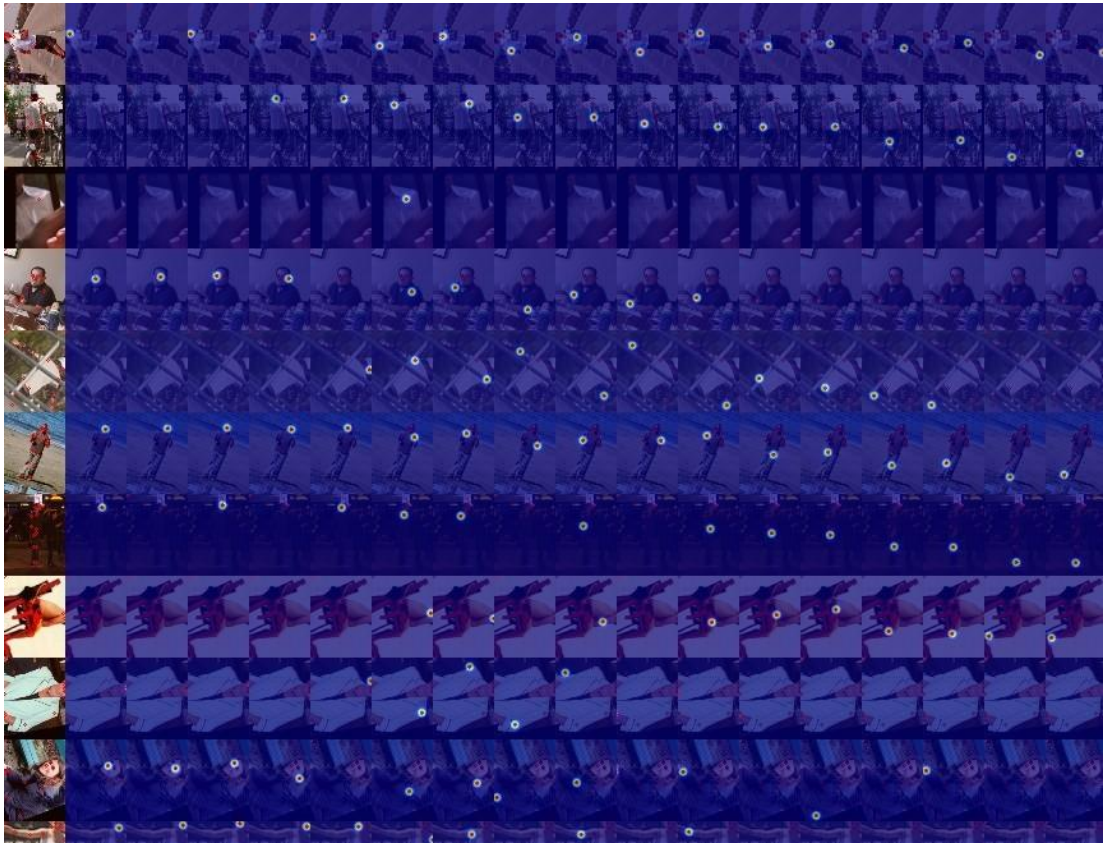
Figure 1. A simple example of a high-resolution network. There are four stages. The 1st stage consists of high-resolution convolutions. The 2nd (3rd, 4th) stage repeats two-resolution (three-resolution, four-resolution) blocks. The detail is given in Section 3.

- Metodele encoder-decoder reconstruiesc input-ul dintr-o rezolutie micsorata si skip connections din encoder
- HRNet (**Hi**gh **R**esolution **Net**work mentine reprezentarea de inalta rezolutie pe tot parcurusul
- Se adauga branch-uri succesive de stride /2, /4, /8, /16 pe parcurs
- Operatii de subsampling (strided convolutions not pooling)
- Operatii de upsampling (nearest neighbours upsampling)
- Se face in mod repeat multi-scale **fusion** de la mai mult rezolutii
- Pytorch code

Studiu de caz: Pose Estimation

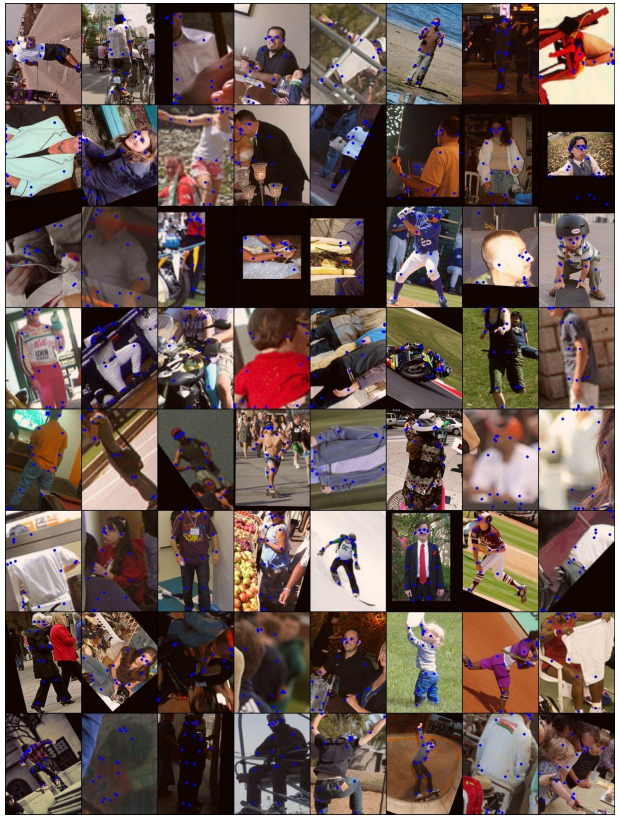


Keypoints Ground truth

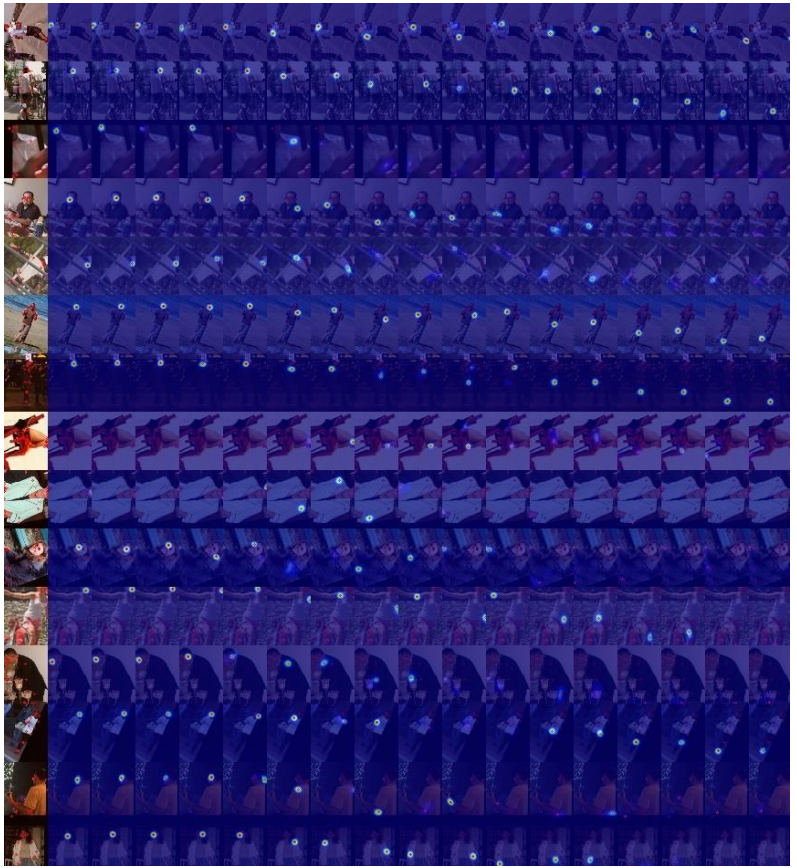


Keypoints Heatmap Representation

Studiu de caz: Pose Estimation

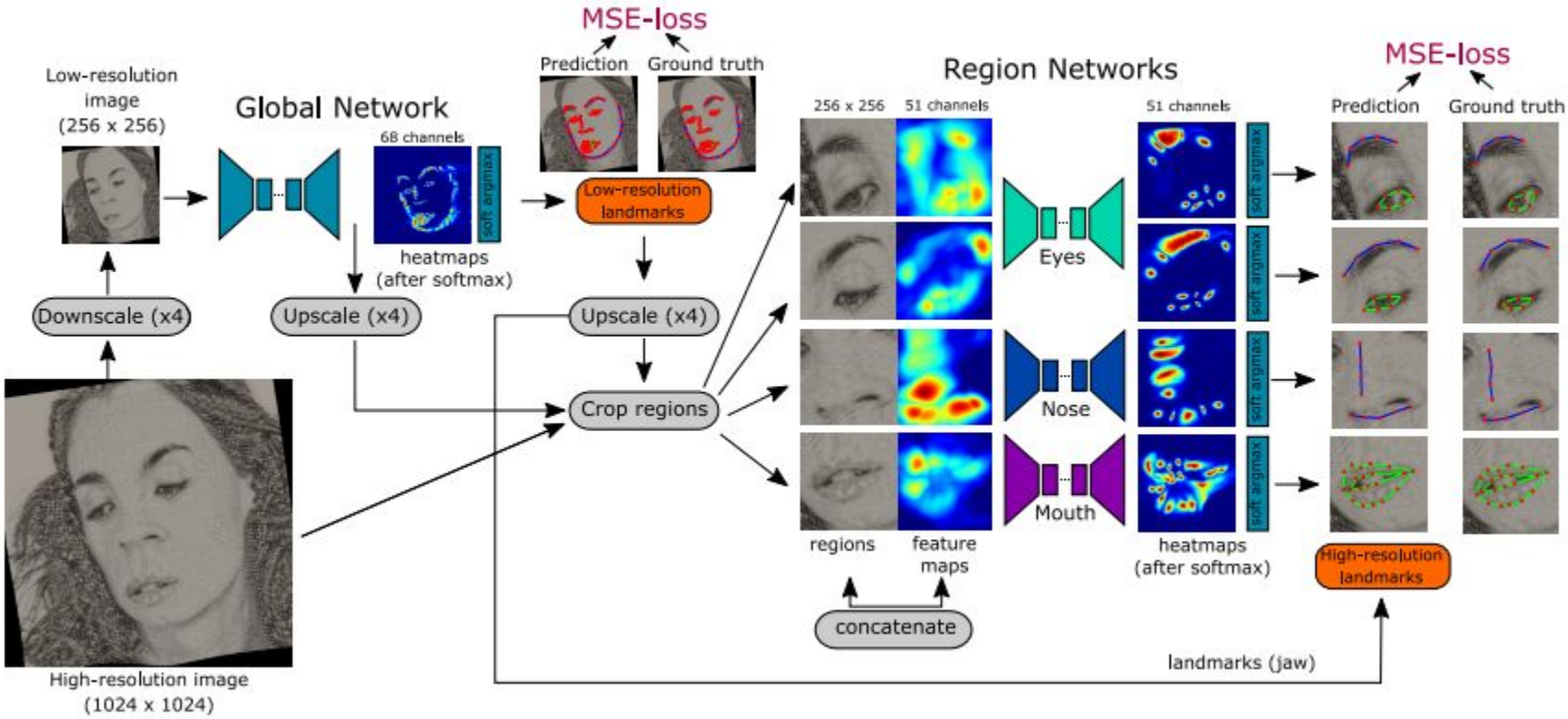


Keypoints Predictions



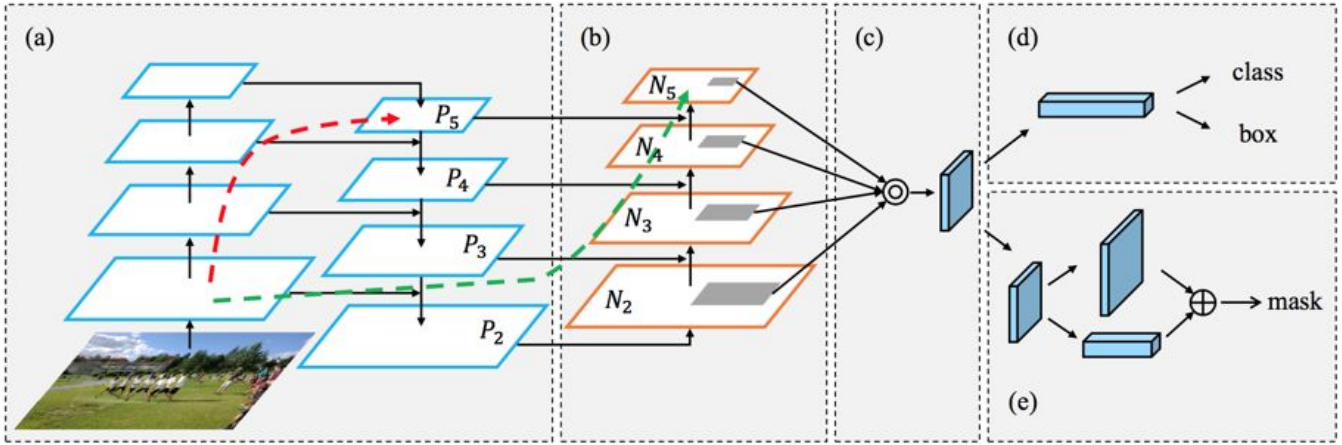
Keypoints Heatmap Prediction

Predicting kpts on High Resolution Images



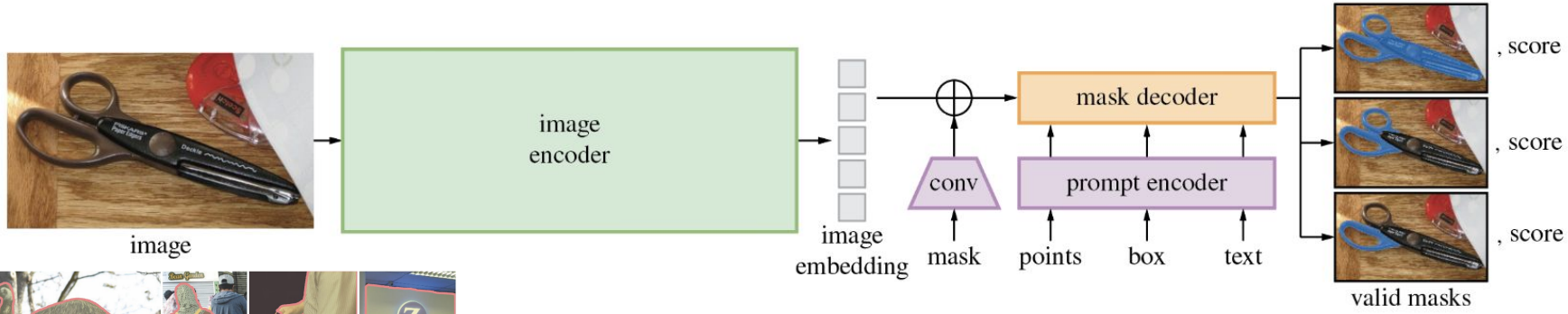
Studiu de Caz: FPN

- Feature Pyramid Network
- Utile in object detection - folosite in Faster R-CNN, Mask R-CNN, RetinaNet, YOLOV5
- A Feature Pyramid Network, or FPN, is **a feature extractor that takes a single-scale image of an arbitrary size as input, and outputs proportionally sized feature maps at multiple levels, in a fully convolutional fashion**
- Nu reconstruiesc input ci lucreaza la nivel de feature maps pentru obtinerea de reprezentari semantice si cu rezolutie mare
- Pot fi aplicate de 2 ori (bottom up si top-down) ca mai jos
- Sunt idendepndente de encoder-ul folosit



Mask R-CNN

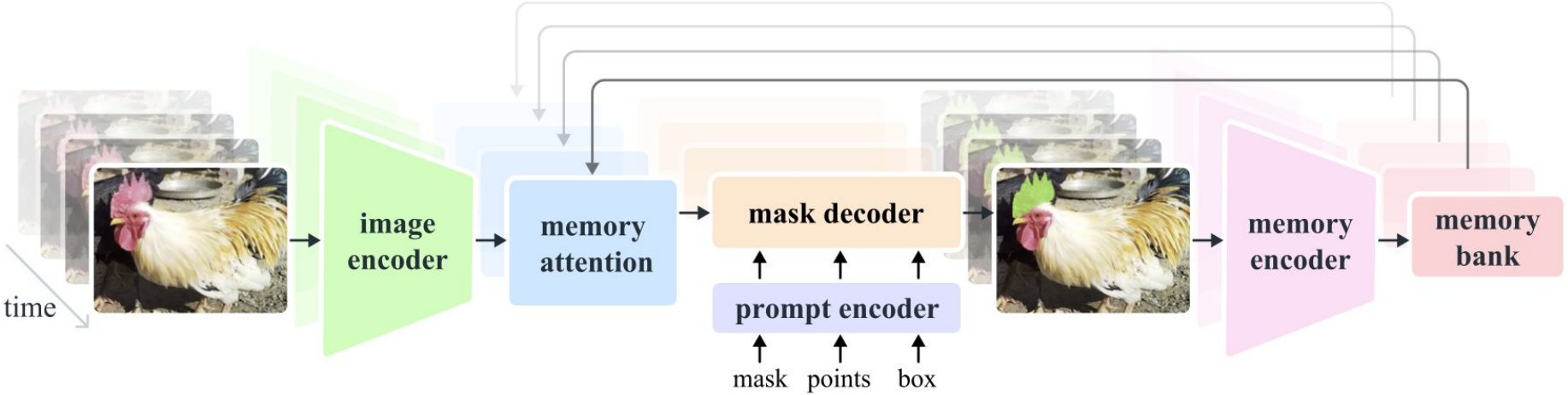
Segment Anything Model (SAM)



https://github.com/facebookresearch/segment-anything

https://arxiv.org/pdf/2304.02643.pdf

Segment Anything Model 2 (SAM2)

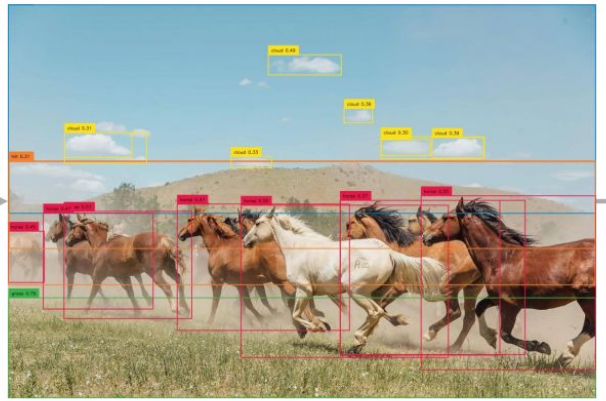


- Unified architecture (image & video)
- Memory Conditioning System
- Performance optimizations

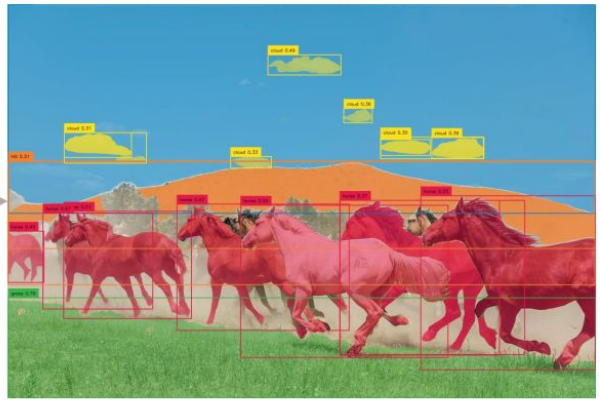
https://github.com/facebookresearch/sam2

Grounded Segment Anything

Text Prompt:
“Horse. Clouds. Grasses. Sky. Hill.”



Grounding DINO:
Detect Everything



Grounded-SAM:
Detect and Segment Everything

- Prompt text to SAM

https://github.com/IDEA-Research/Grounded-Segment-Anything

Resurse

- Aritmetica Convolutiilor
  - http://deeplearning.net/software/theano/tutorial/conv_arithmetic.html
  - https://arxiv.org/pdf/1603.07285.pdf
  - https://github.com/vdumoulin/conv_arithmetic
  - https://towardsdatascience.com/up-sampling-with-transposed-convolution-9ae4f2df52d0#d907
  - https://medium.com/apache-mxnet/transposed-convolutions-explained-with-ms-excel-52d13030c7e8
- Segmentare Semantica folosind FCN
  - https://github.com/shelhamer/fcn.berkeleyvision.org