Nizhny Novgorod State University

Institute of Information Technologies, Mathematics and Mechanics

Department of Computer software and supercomputer technologies

Educational course
«Modern methods and technologies
of deep learning in computer vision»

Semantic segmentation of images using deep learning

Supported by Intel

Getmanskaya Alexandra, Kustikova Valentina

Content

- Goals
- Semantic segmentation problem statement
- Public datasets
- Quality metrics
- Deep models for semantic segmentation
- Comparison of deep models for semantic segmentation
- Conclusion

Goals

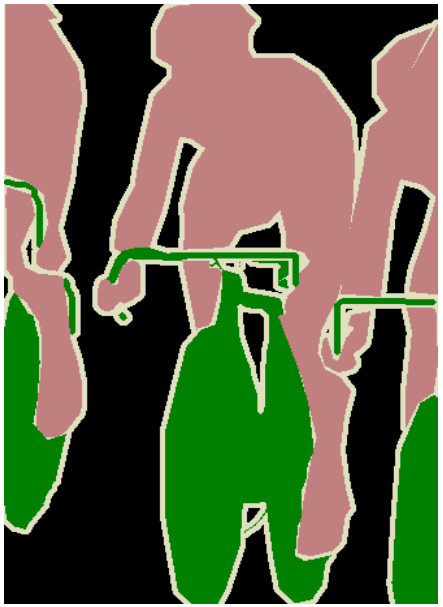
- ***The goal*** is to study deep models for solving problem of semantic segmentation (real-life images, medical images, on-road images)

SEMANTIC SEGMENTATION PROBLEM STATEMENT

Problem statement (1)

- The problem of semantic segmentation is to match each image pixel with the class of objects to which this pixel belongs (different colors correspond to the different classes)

Original image



Groundtruth

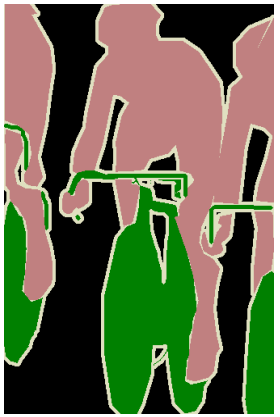
Segmentation result

* The PASCAL Visual Object Classes Homepage [http://host.robots.ox.ac.uk/pascal/VOC].

Problem statement (2)

- The original image is represented by a set of pixel intensities $I = \left(I_{ij}^{k}\right)_{\substack{0 \le i < w \\ 0 \le j < h \\ 0 \le k < 3}}$, where w and h are image width and height, k is a number of color channels of the image
- The set of object classes $C = \{0, 1, \ldots, N-1\}$ is defined, 0 corresponds to the background, the set of class identifiers uniquely corresponds to the set of class names
- It is required to find a mapping

$$\varphi\left(I_{ij}\right) = c$$



PUBLIC DATASETS

Public datasets (1)

Dataset	Number of images in train dataset	Number of images in test dataset	Number of classes
Semantic segmentation of real-life images			
PASCAL VOC 2012 [http://host.robots.ox.ac.uk/pascal/VOC/voc2012]	9 963	1 447	20
ADE20K [http://groups.csail.mit.edu/vision/datasets/ADE20K]	20 210	2 000	150
MS COCO'15 [http://mscoco.org]	80 000	40 000	80
...			

Public datasets (2)

Dataset	Number of images in train dataset	Number of images in test dataset	Number of classes
Semantic segmentation of on-road images			
CamVid [http://mi.eng.cam.ac.uk/research/projects/VideoRec/CamVid]	468	233	11
Cityscapes [https://www.cityscapes-dataset.com]	2 975	500	19
KITTI [http://www.cvlibs.net/datasets/kitti]	200	200	4
Semantic segmentation of indoor scenes			
Sun-RGBD [http://rgbd.cs.princeton.edu]	10 355	2 860	37
NYUDv2 [http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html]	795	645	40

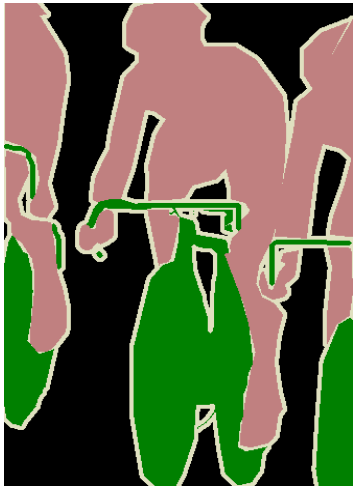
Public datasets (3)

- MS COCO’15 is the largest dataset of real-life images for semantic segmentation
- Cityscapes dataset contains images captured in 50 cities from a DVR on a car moving in urban in various weather conditions
- KITTI benchmark is the dataset and the toolkit for measuring the quality of analyzing on-road scenes (object detection, semantic segmentation, object tracking, lane detection, etc.)
- Sun-RGBD benchmark contains images of indoor scenes (home, office) for solving the tasks of image classification (2 categories), semantic segmentation, 3D reconstruction, high-level scene understanding

PASCAL VOC 2012

- PASCAL VOC 2012 is the most popular dataset
- 20 classes of real-life objects: airplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv/monitor

Original image



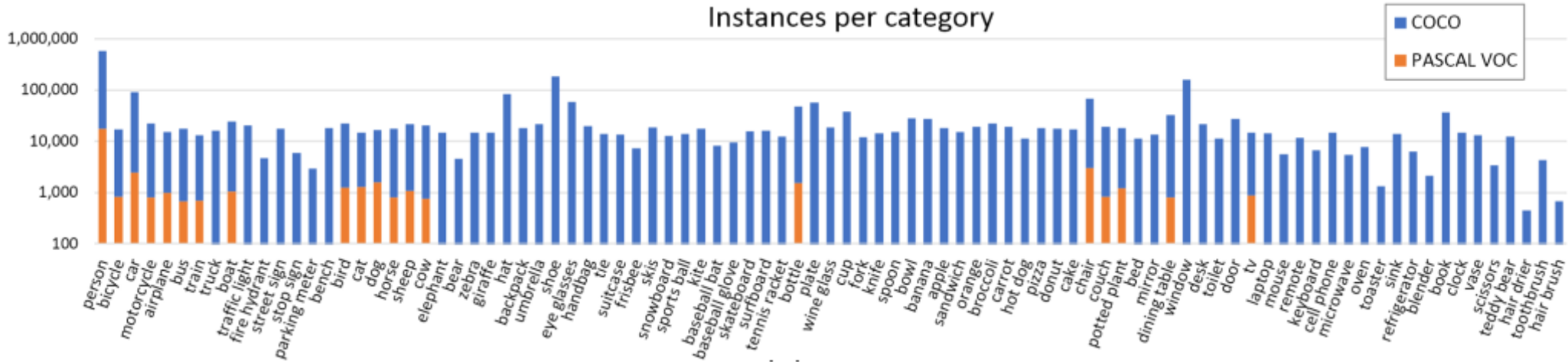
Groundtruth

(different colors correspond to different object classes, also object boundaries are represented)

* The PASCAL Visual Object Classes Homepage [http://host.robots.ox.ac.uk/pascal/VOC].

MS COCO’15

- MS COCO’15 is the largest public dataset of real-life images (similar to PASCAL VOC) by the number of object classes (80 categories) and the number of images; each category contains a significant number of images (approximately equal number of objects for each class)



* Lin T.Y., et al. Microsoft COCO: Common objects in context // Lecture Notes in Computer Science. – Vol. 8693. – 2014. – P. 740-755. – [https://arxiv.org/pdf/1405.0312].

Cityscapes

- Images of on-road scenes from a DVR
- 5 000 images with high quality annotation
- 20 000 images with coarse annotation
- 30 classes combined into 8 groups

Example of high quality annotation

Zurich (Switzerland)

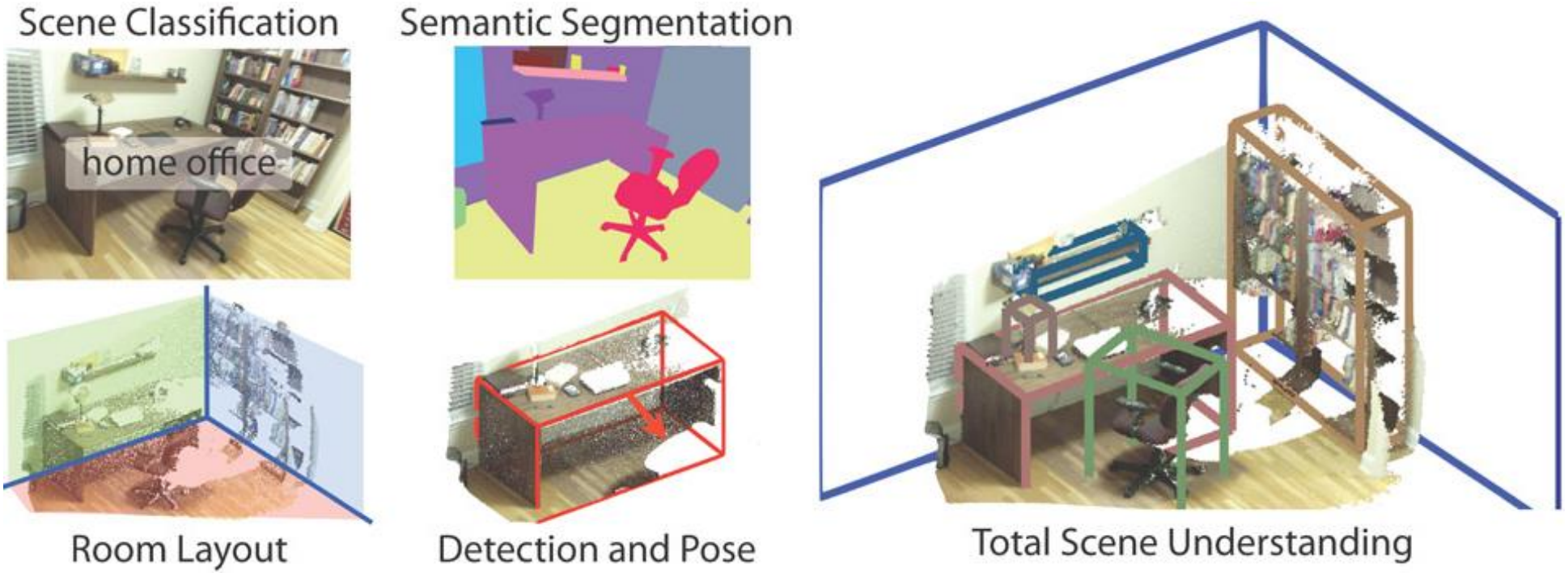
Example of rough annotation

Saarbrücken (Germany)

* The Cityscapes Dataset Homepage [https://www.cityscapes-dataset.com/examples].

SUN RGB-D

- SUN RGB-D contains images and groundtruth of indoor scenes for solving several tasks (examples are represented below)



* Song S., Lichtenberg S.P., Xiao J. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite [https://3dvision.princeton.edu/projects/2015/SUNrgbd/poster.pdf].

QUALITY METRICS

Quality metrics

- Pixel accuracy
- Mean pixel accuracy over classes
- Intersection over Union (IoU) or Jaccard index
- Dice index или F1-score

Pixel accuracy

- ***Pixel accuracy*** is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where $TP + TN$ is a number of correctly classified pixels (true positives + true negatives),

and $TP + TN + FP + FN$ is a total number of pixels

		Prediction: **True**	Prediction: **False**
Groundtruth	**True**	TP	FN
	False	FP	TN

Mean pixel accuracy over classes

- Pixel accuracy shows the number of correctly classified pixels
- Pixel accuracy is not representative in the case of class imbalance

- Therefore, ***mean pixel accuracy*** is introduced. This metric calculates the pixel accuracy for each class separately and then calculates mean value over the number of classes

Intersection over Union (1)

- ***Intersection over Union metric*** (***IoU)*** or Jaccard index

$$IoU = \frac{TP}{TP + FP + FN}$$

where TP is a number of correctly classified pixels (true positives), FP is a number of pixels that the method has been classified as belonging to the class, but they do not belong (false positives), FN is a number of pixels that belong to the class, but the method has been classified them as not belonging to the class (false negatives)

		Prediction	
		True	**False**
Groundtruth	**True**	TP	FN
	False	FP	TN

Intersection over Union (2)

- Usually, the mean value of the IoU metric (mean IoU) for all classes on a complete dataset is calculated
- Mean IoU can be calculated as a weighted mean of values obtained for individual classes. Weights are assigned equal to the number of pixels of each class
- When calculating the IoU metric, the background class may not be taken into account
- Pixels on the object boundaries may not be taken into metric calculation or taken into the metric with a lower weight

Dice index

- ***Dice index*** or F1-score is as follows:

$$DICE = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

- Dice index differs from the Jaccard index by one coefficient
- Dice index and Jaccard index are related by the formulas:

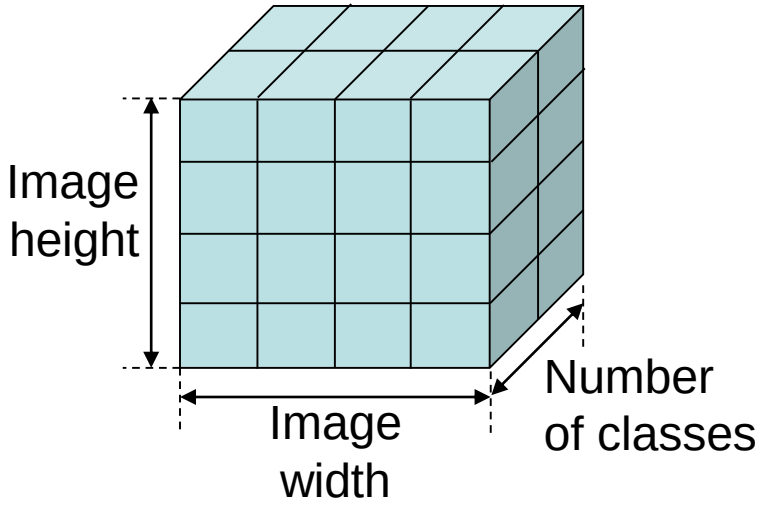
$$IoU = \frac{DICE}{2 - IoU}, \qquad DICE = \frac{2 \cdot IoU}{1 + IoU}$$

- It is not required to calculate both metrics, it is enough to calculate one of them

DEEP MODELS FOR SEMANTIC SEGMENTATION

The problem of using deep models for semantic segmentation (1)

- Solving the problem of semantic segmentation, the model output should be a three-dimensional tensor, tensor elements correspond to the confidence of each pixel belonging to a certain class



- ***How to provide at the output a tensor whose spatial dimensions coincide with the resolution of the original image?***

The problem of using deep models for semantic segmentation (2)

- Methods of obtaining an output tensor, whose spatial dimension coincides with the resolution of the original image:
  - Interpolation
  - Encoder-decoder architecture
  - Probabilistic graph methods, in particular, conditional random fields (CRF)
- Interpolation is the most simple way, but it does not allow to obtain high segmentation quality, especially for small objects and at the object boundaries
- Two other methods are more perspective in terms of semantic segmentation quality

Deep models (1)

- ***FCNs, SegNet, U-Net (2015)***

Fully convolutional networks

  - Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].
  - Badrinarayanan V., Kendall A., Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. – 2015. – [https://arxiv.org/pdf/1511.00561.pdf], [https://ieeexplore.ieee.org/document/7803544].
  - Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation. – 2015. – [https://arxiv.org/pdf/1505.04597.pdf], [https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28].

Deep models (2)

- ***PSPNet (2016)***
  - Zhao H., Shi J., Qi X., Wang X., Jia J. Pyramid scene parsing network. – 2016. – [https://arxiv.org/pdf/1612.01105.pdf], [https://ieeexplore.ieee.org/document/8100143].
- ***ICNet (2017)***
  - Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].

Feature pyramids

Deep models (3)

Using CRF and searching for the alternative for CRF to speed up calculations

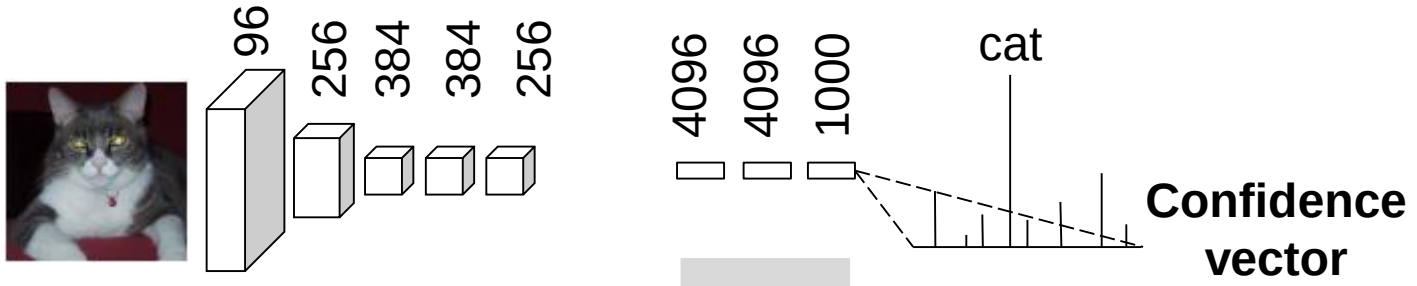
- ***DeepLab-v1,*-v2, *-v3, *v3+ (2014-2018)***
  - Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. – 2014. – [https://arxiv.org/pdf/1412.7062.pdf].
  - Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. – 2017. – [https://arxiv.org/pdf/1606.00915.pdf], [https://ieeexplore.ieee.org/document/7913730].
  - Chen L.-C., Papandreou G., Schroff F., Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. – 2017. – [https://arxiv.org/pdf/1706.05587.pdf].
  - Chen L.-C., Zhu Y., Papandreou G., Schoff F., Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. – 2018. – [https://arxiv.org/pdf/1802.02611.pdf].

FCN (1)

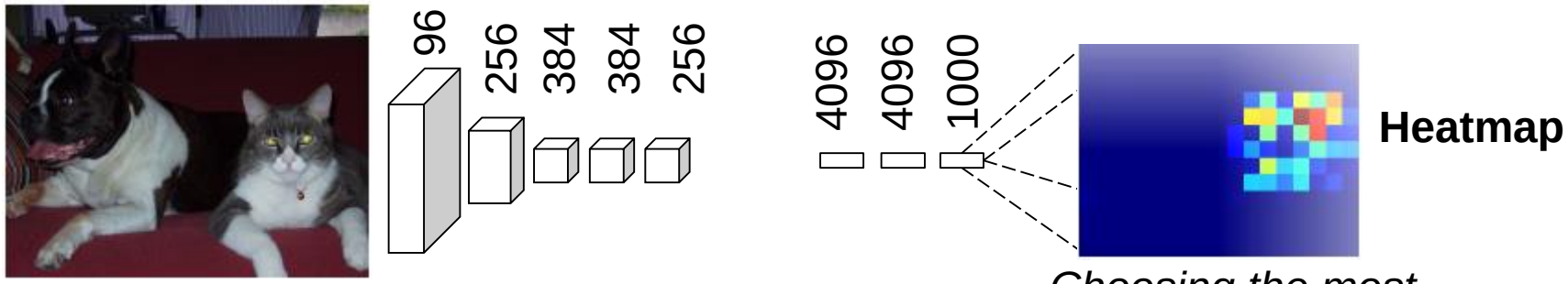
- FCNs (Fully Convolutional Networks) are models whose goal is to adapt classification convolutional networks (AlexNet, VGG, GoogLeNet) to solve the problem of semantic segmentation
  - Classification models receive a fixed-resolution image as input
  - Classification models return confidences of belonging the image of available object classes
  - We replace fully connected layers with convolutional ones to apply the model to images of arbitrary resolution
  - Therefore, we applied a sliding window to obtain a confidence vector for each pixel

* Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].

FCN (2)



Replacing fully connected layers with fully convolutional ones (another interpretation of features)



Choosing the most probable class index

* Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].

FCN (3)

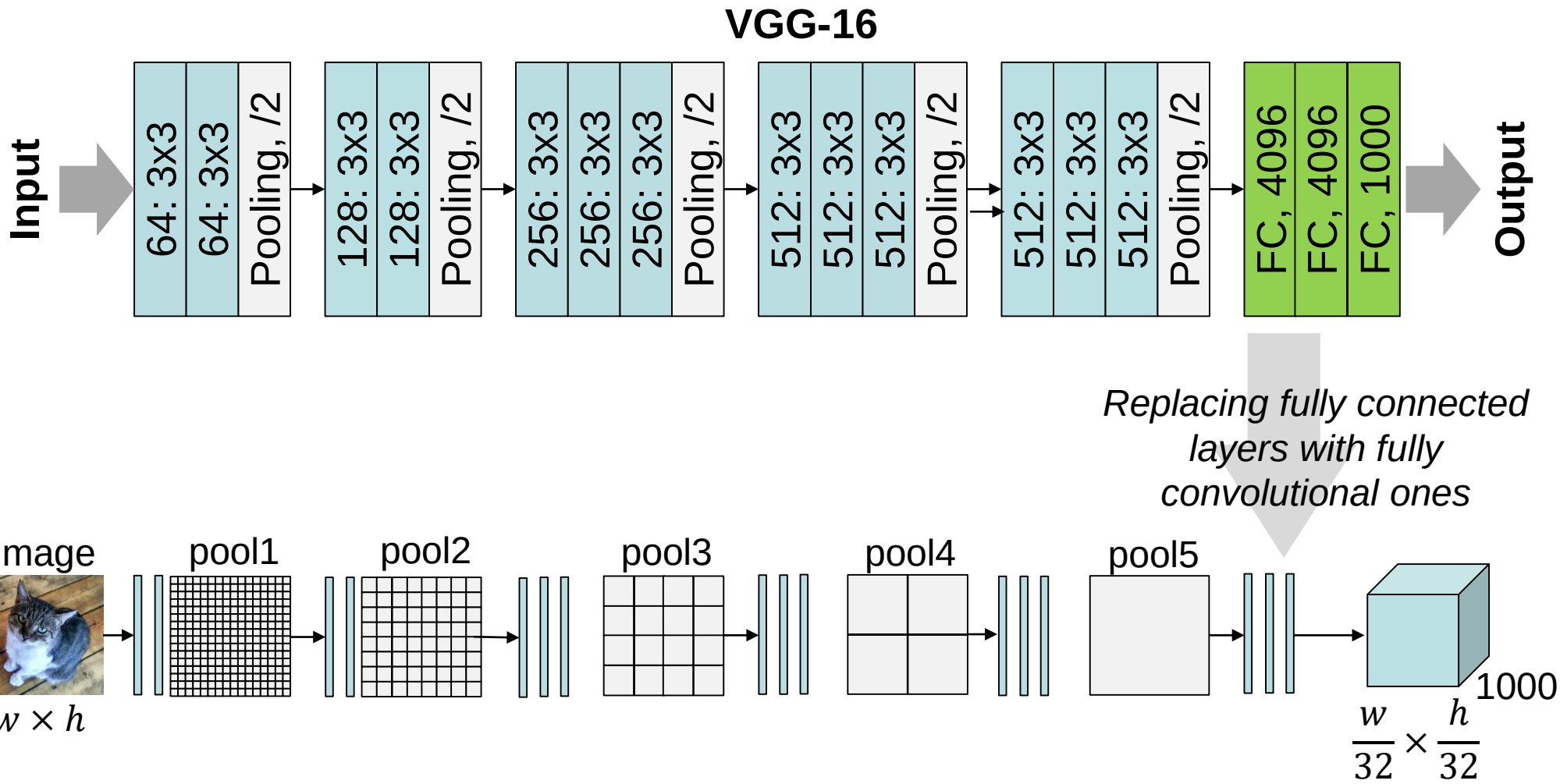
- Fully connected layers are converted to fully convolutional ones using one-dimensional convolutions (the kernel size is 1x1). The layers remain the same
- The input image may be of arbitrary resolution
- A three-dimensional tensor is the output of the deep model, the number of channels correspond to the number of object classes, and the spatial dimensions correspond to the number of possible positions of sliding window on the original image
- Choosing a class with the maximum confidence for each position allows to construct a heatmap of the image, which is a result of semantic segmentation, but it has lower resolution

FCN (4)

- Increasing the resolution of feature maps, including the output heatmap, is implemented using deconvolutions (backwards or transposed convolutions)
- To improve the quality of the final heatmap, it is proposed to use feature maps obtained on the intermediate layers of the model, i.e. low-level features
- Authors of the FCN model* used AlexNet, VGG, GoogLeNet as basic models
- VGG-16 allowed to achieve the best results, so FCN, based on VGG-16, is further considered

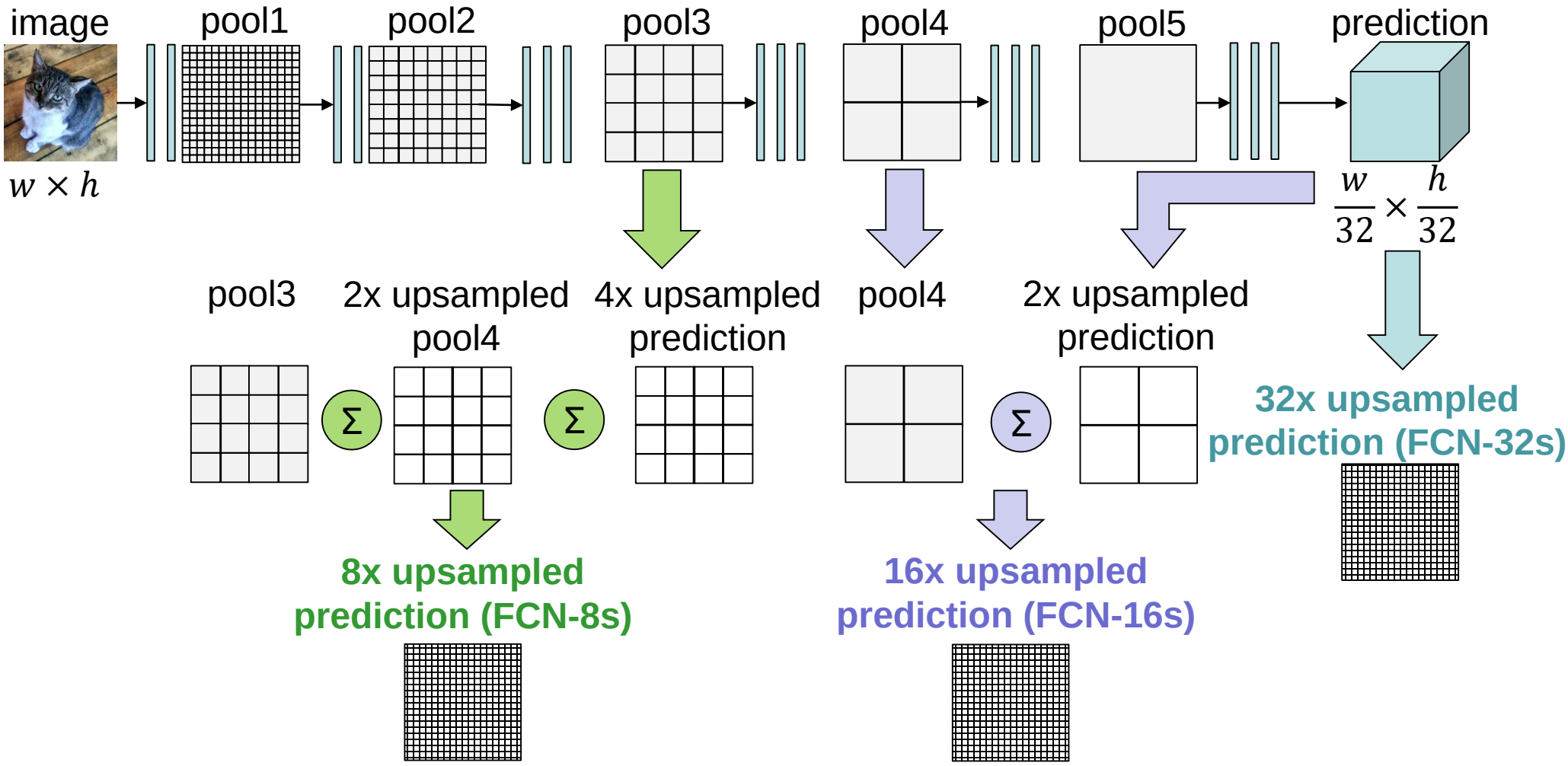
* Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].

FCN (5)



* Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].

FCN (6)



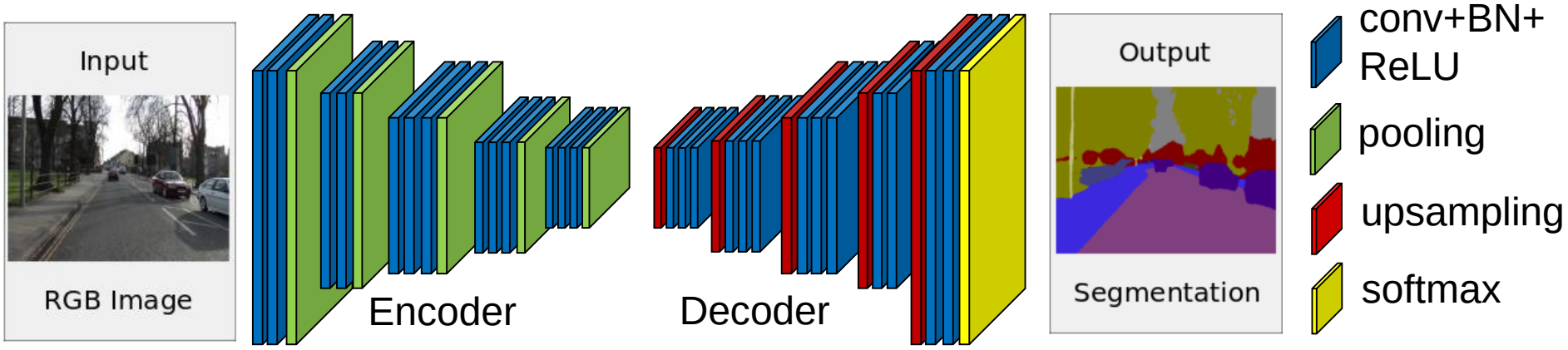
* Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].

FCN (7)

- Replacing the fully connected layers with the fully convolutional ones in VGG-16:
  - FC 4096 → Conv 4096, 1x1
  - FC 4096 → Conv 4096, 1x1
  - FC 1000 → Conv 1000, 1x1
- After replacing we can process images of arbitrary resolution $w \times h$, and construct the final heatmap of the shape $\frac{w}{32} \times \frac{h}{32} \times 1000$
- The spatial dimension of the output is increased by applying upsampling with stride 32. The coarse segmentation is constructed (FCN-32s model)
- More accurate segmentation results are obtained when using features from the intermediate layers (models FCN-16s, FCN-8s)

SegNet (1)

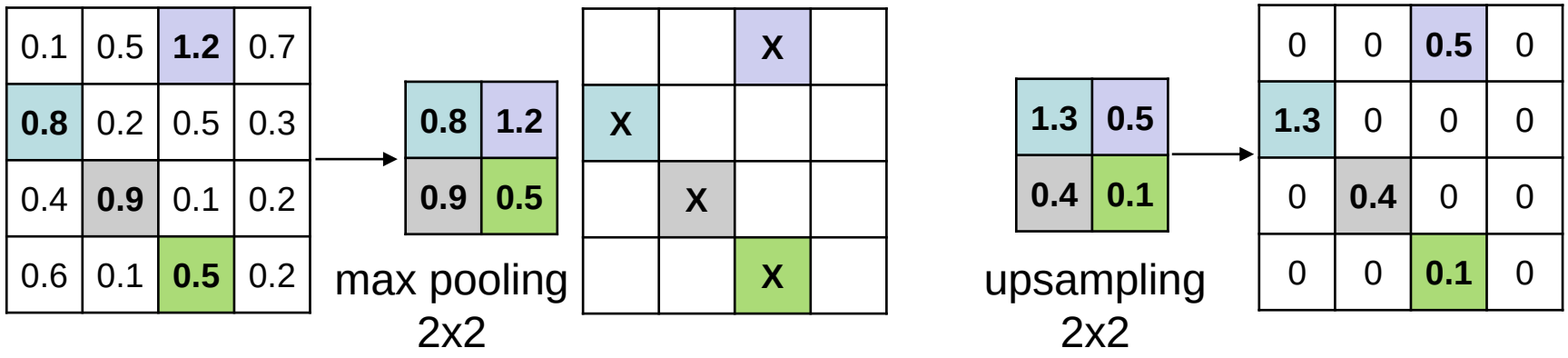
- ***SegNet*** is a deep model for semantic segmentation based on the encoder-decoder architecture
- The goal is to create an efficient deep model for semantic segmentation of on-road and indoor images



* Badrinarayanan V., Kendall A., Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. – 2015. – [https://arxiv.org/pdf/1511.00561.pdf], [https://ieeexplore.ieee.org/document/7803544].

SegNet (2)

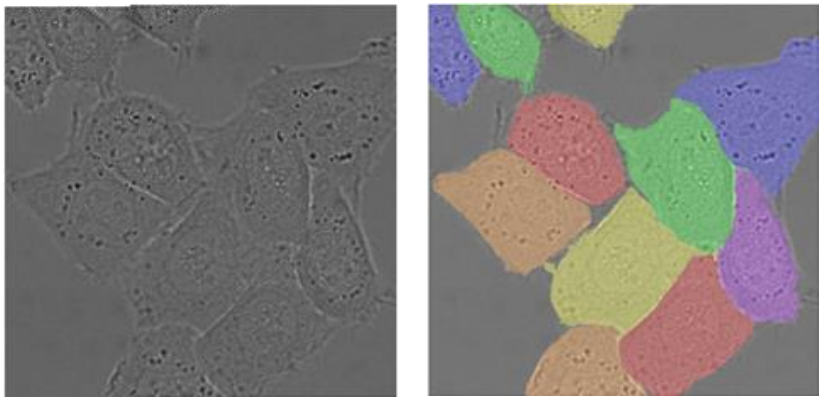
- The encoder contains the convolutional part of the VGG-16 network
- The decoder is constructed in a mirror-wise manner to the encoder:
  - Each convolutional layer in the encoder corresponds to the convolutional layer in the decoder in the reverse order
  - Each pooling operation corresponds to the upsampling operation. The indices of the max pooling on each layer of the encoder are stored and used in the decoder for the upsampling



U-Net (1)

- The authors of U-Net propose the model and learning strategy based on the increasing the size of the dataset by image transformation (data augmentation) for more efficient use of the small set of annotated samples
- U-Net shows high segmentation quality for neuronal structures in electron microscopic stacks

Original image



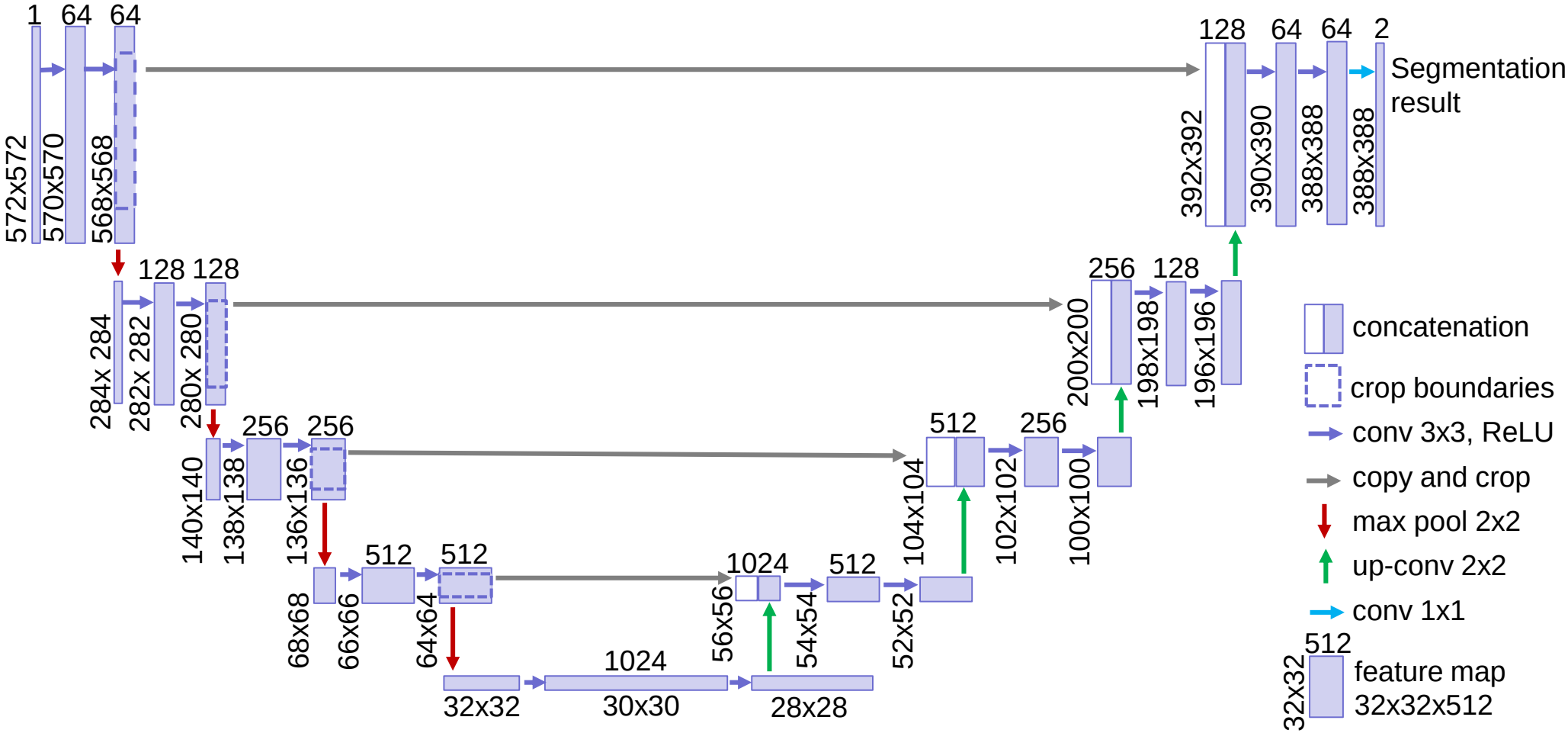
Neuronal structures

* Ronneberger O., Fischer P., Brox T. U-Net: Convolutional networks for biomedical image segmentation. – 2015. – [https://arxiv.org/pdf/1505.04597.pdf].

U-Net (2)

- The U-Net model consists of two parts:
  - ***Contracting path*** is a convolutional network represented by a sequence of blocks which contain two 3x3 convolutions (no padding), followed by ReLU activation function and max pooling with the kernel 2x2 and stride 2
  - ***Expansive path*** is a convolutional network represented by a sequence of blocks which contain upsampling, upper convolution (2x2 convolution reducing the number of channels by half), concatenation with the corresponding feature map from the contracting path, two 3x3 convolutions, followed by ReLU activation function

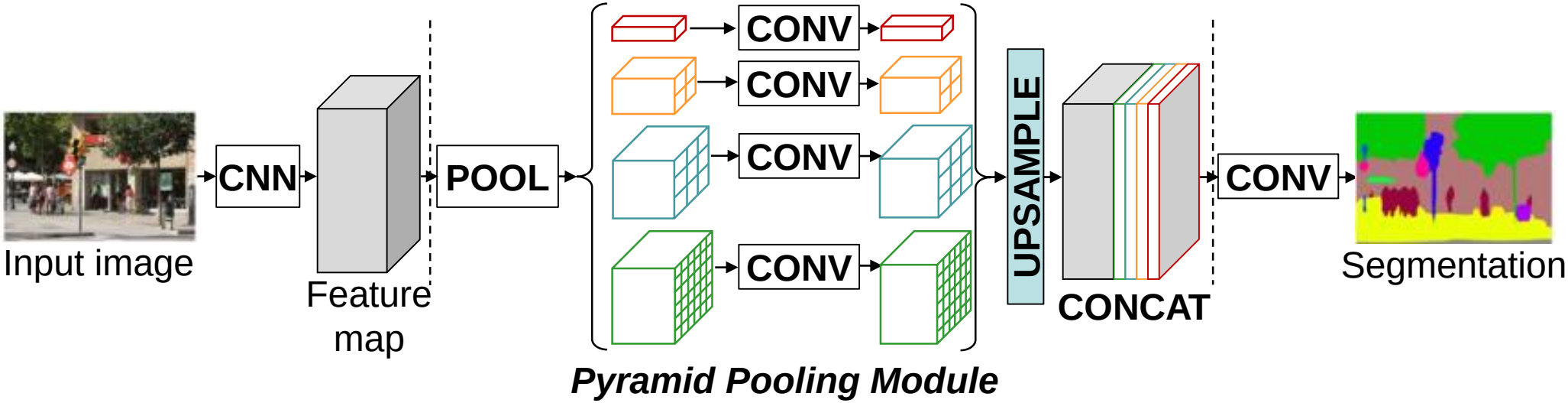
U-Net (3)



* Ronneberger O., Fischer P., Brox T. U-Net: Convolutional networks for biomedical image segmentation. – 2015. – [https://arxiv.org/pdf/1505.04597.pdf].

PSPNet (1)

- PSPNet (Pyramid Scene Parsing) is a model that constructs a pyramid of feature maps of different scales
- PSPNet was the best model at ImageNet Scene Parsing Challenge 2016, PASCAL VOC 2012 and Cityscapes in 2016



* Zhao H., Shi J., Qi X., Wang X., Jia J. Pyramid scene parsing network. – 2016. – [https://arxiv.org/pdf/1612.01105.pdf], [https://ieeexplore.ieee.org/document/8100143].

PSPNet (2)

- ***Feature map***
  - To extract features, the convolutional part with dilated convolutions of the ResNet model is used
- ***Pyramid Pooling Module***
  - Pooling (POOL)
    - Red map: the result of global pooling for each channel of the feature map (the “coarsest” level)
    - Orange map: the result of pooling by regions obtained when dividing the feature map into 2x2 blocks
    - Blue map: the result of pooling by regions obtained when dividing the feature map into 3x3 blocks
    - Green map: the result of pooling for each channel by regions obtained when dividing the feature map into 6x6 blocks

PSPNet (3)

- Intermediate convolutions (set of CONV layers)
  - Convolutions with 1x1 kernels to reduce the number of channels, i.e. reducing the representation of the context to $\frac{1}{N}$ from the original one, where N is the number of pyramid levels
  - In the presented example $N = 4$, if the number of channels of the input feature map is 2048, then the number of channels at the output of each pyramid level is 512
- Upsampling (UPSAMPLE)
  - Upsampling supposes using of bilinear interpolation to increase the dimension of feature maps to the original one
- Concatenation of feature maps (CONCAT)
  - Concatenation of the original feature map with the feature maps obtained after upsampling

❑ ***Segmentation result***

- Final convolution (CONV)

PSPNet (4)

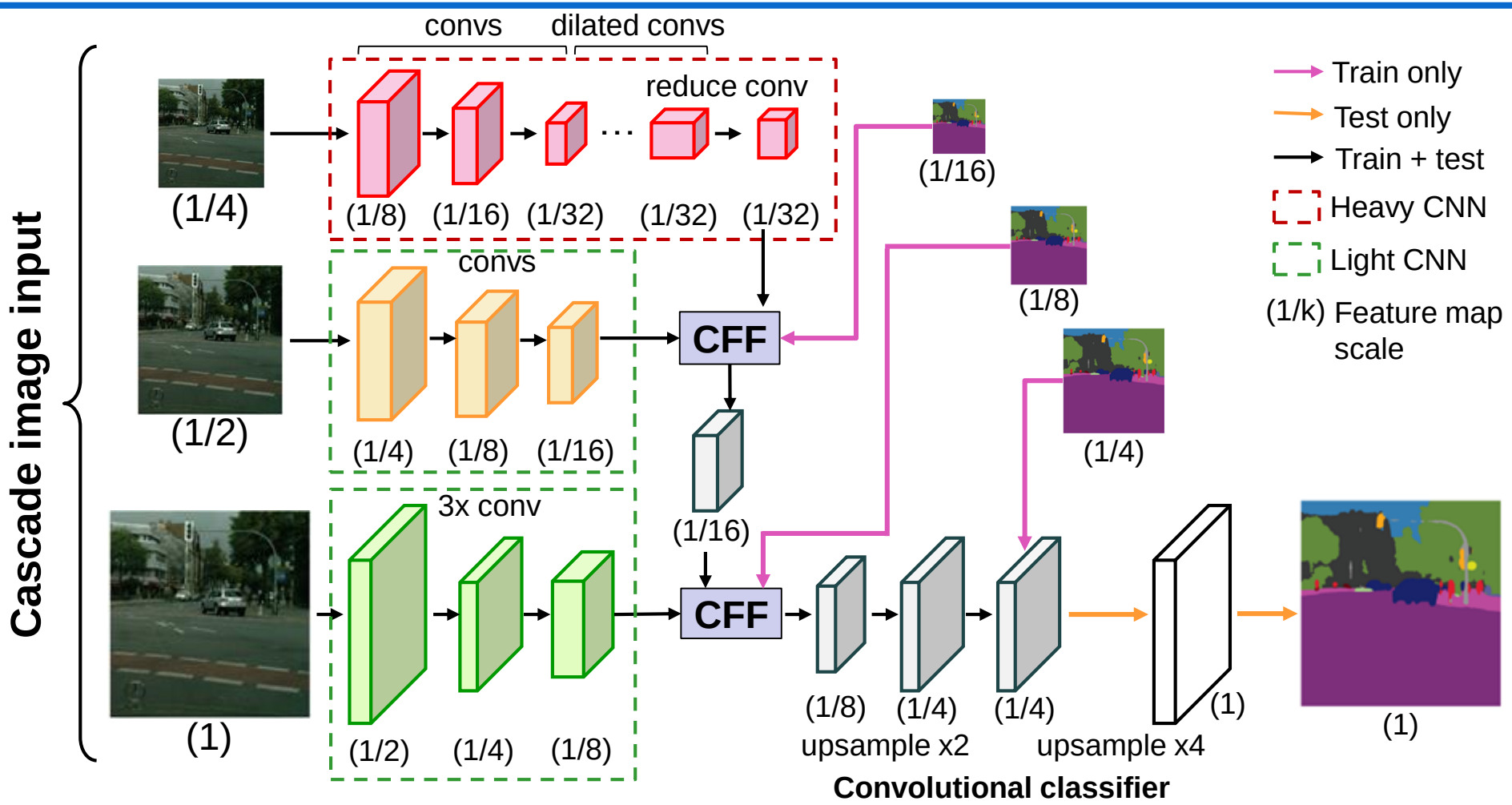
- Learning features:
  - The auxiliary loss function from the intermediate layer of the model is introduced
  - The auxiliary loss function helps to optimize training, while the main loss function is fully responsible for solving semantic segmentation problem
  - To balance the contribution of the auxiliary loss, a weight coefficient is introduced

ICNet (1)

- ICNet (Image Cascade Network) is a model for semantic segmentation in real time (on a single GPU), which is based on the a cascade of feature maps constructed for different scales of the original image
- The model input is a pyramid of image scales
- For each image in the pyramid, feature map using convolutional networks is provided:
  - The larger image in the pyramid, the simpler its convolution model is used
  - Constructing a feature map on each subsequent scale, features from smaller scales are used

* Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].

ICNet (2)



* Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].

ICNet (3)

- The convolutional network on each layer reduces the spatial dimensions of the feature map, or does not change them
- Concatenation of feature maps from adjacent scales is provided using ***Cascade Feature Fusion*** (CFF) module
- CFF module allows to restore and improve the result of segmentation with less computational cost

- Further, we will consider the structure of the CFF module and the scheme of its working during training and testing the model

* Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].

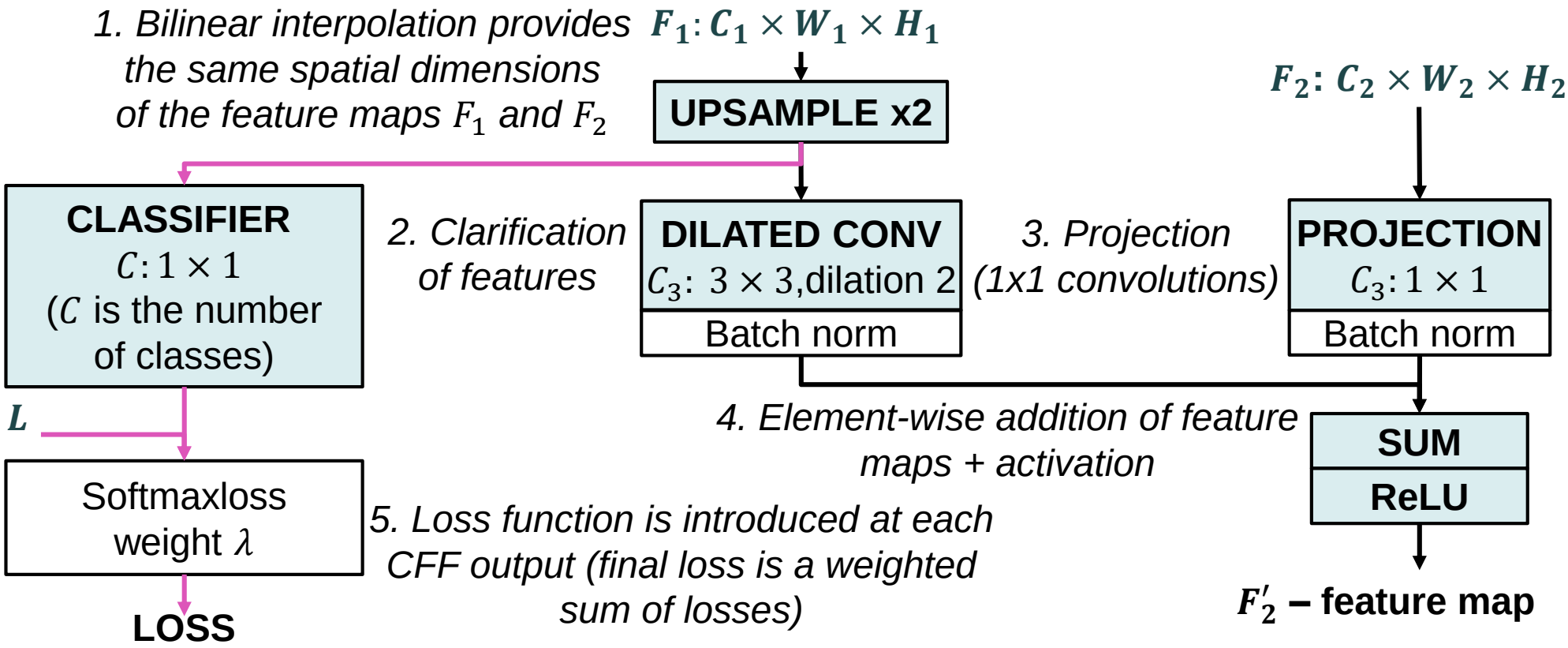
ICNet (4)

- CFF module has three inputs:
  - The feature map F_1 of the size $C_1 \times W_1 \times H_1$ (it is used during training and testing)
  - The feature map F_2 of the size $C_2 \times W_2 \times H_2$ (it is used during training and testing). Spatial size of F_2 in two times larger than spatial size of F_1
  - Image annotation L of the size $1 \times W_2 \times H_2$ (it is used during training)
- Combining the feature maps F_1 and F_2, the combined feature map F_2' is constructed, which is taking into account at the next (larger) scale

* Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].

ICNet (5)

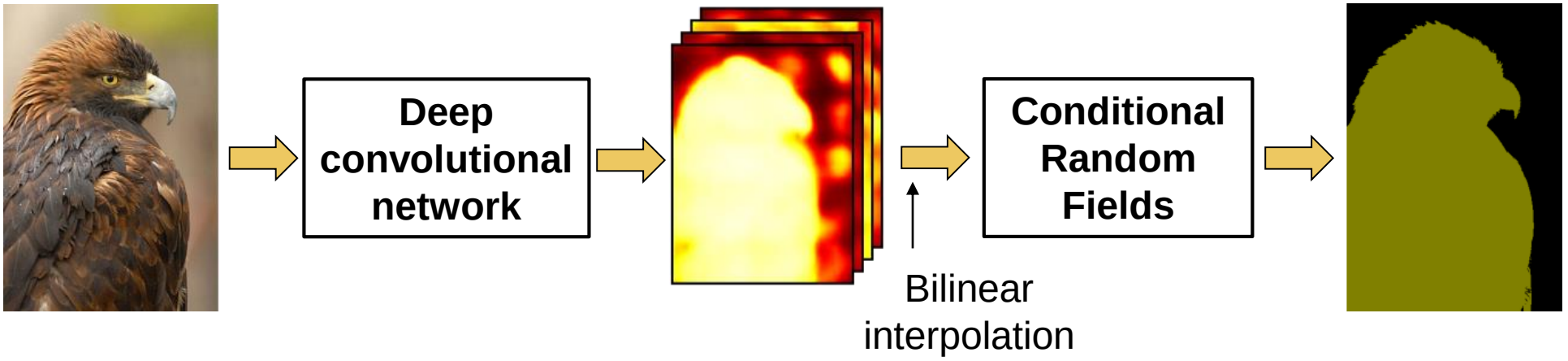
- ***Cascade Feature Fusion***:



* Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].

DeepLab-v1 (1)

- DeepLab-v1 is one of the well-known methods of semantic segmentation, based on the construction of a deep convolutional model to obtain a coarse map of segments and the subsequent using of ***conditional random fields*** (CRF) to refine the results

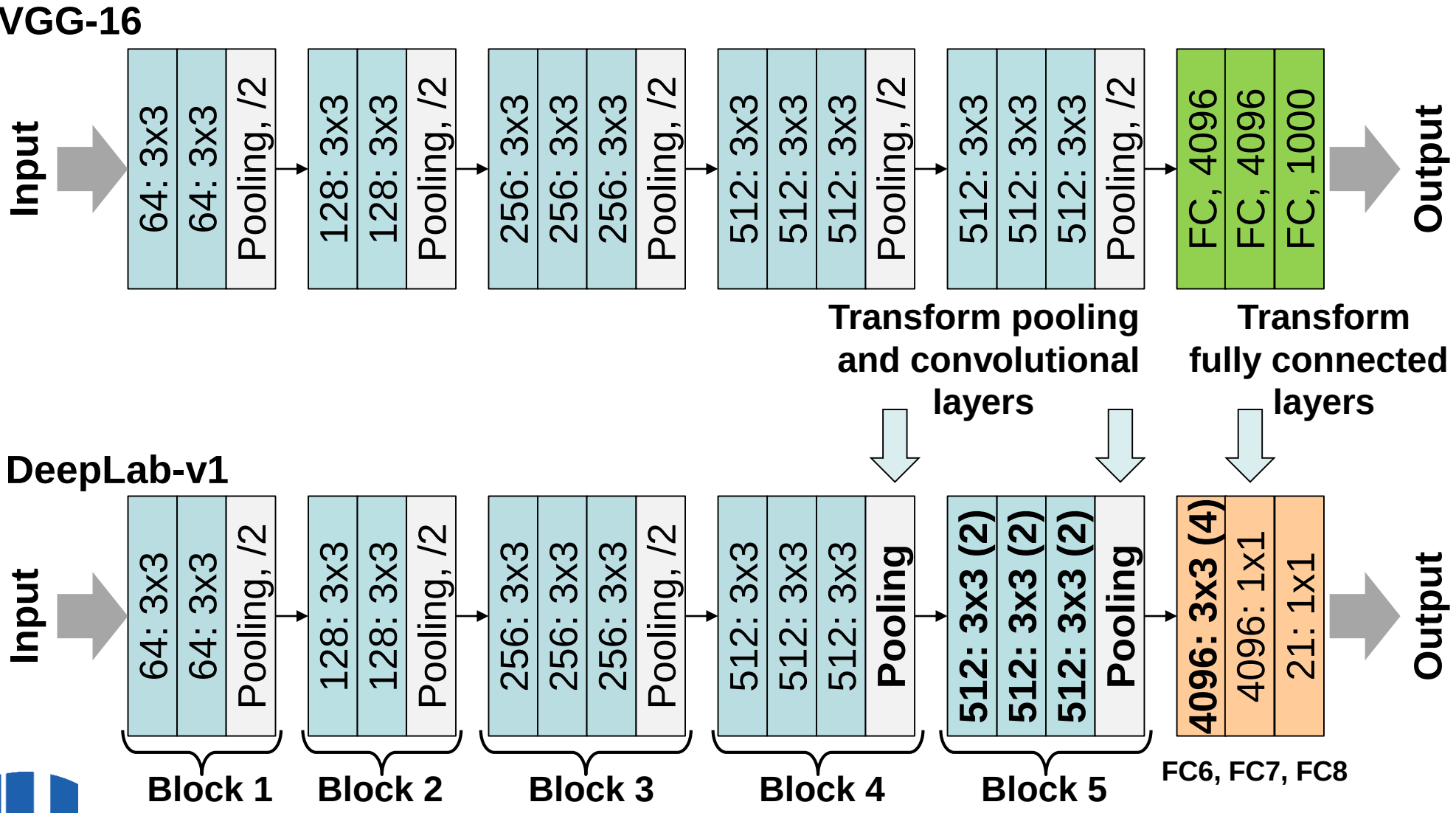


* Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. – 2014. – [https://arxiv.org/pdf/1412.7062.pdf].

DeepLab-v1 (2)

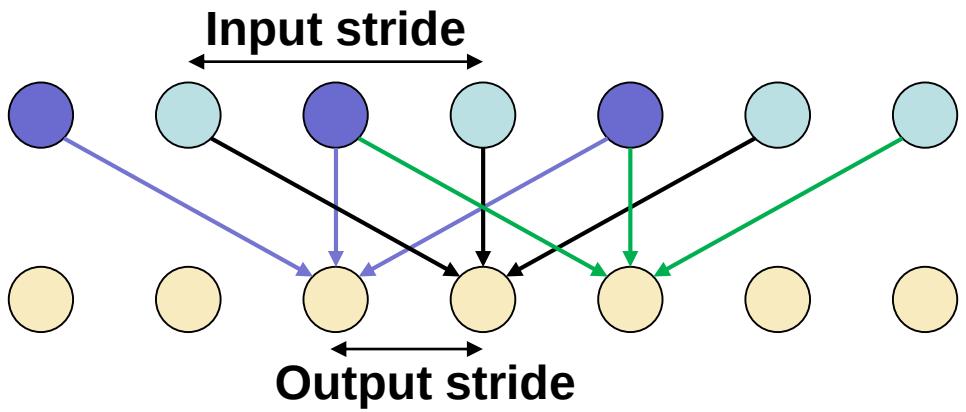
- The convolutional network is based on VGG-16, trained for image classification on ImageNet
- The main differences:
  - The fully connected layers are converted into the fully convolutional ones, therefore, the network input is an image of arbitrary resolution
  - The network input spatial size is 513x513 pixels
  - The network output is a vector of the size 21, which corresponds to the number of classes in PASCAL VOC including background (instead of 1 000)
  - For the last two pooling layers, the down-sampling is removed and the convolutional layers following the pooling layers are modified (the last 3 convolutions and the first fully connected layer)

DeepLab-v1 (3)


VGG-16
Input
64: 3x3
64: 3x3
Pooling, /2
128: 3x3
128: 3x3
Pooling, /2
256: 3x3
256: 3x3
256: 3x3
Pooling, /2
512: 3x3
512: 3x3
512: 3x3
Pooling, /2
512: 3x3
512: 3x3
512: 3x3
Pooling, /2
FC, 4096
FC, 4096
FC, 1000
Output
Transform pooling and convolutional layers
Transform fully connected layers
DeepLab-v1
Input
64: 3x3
64: 3x3
Pooling, /2
128: 3x3
128: 3x3
Pooling, /2
256: 3x3
256: 3x3
256: 3x3
Pooling, /2
512: 3x3
512: 3x3
512: 3x3
Pooling
512: 3x3 (2)
512: 3x3 (2)
512: 3x3 (2)
Pooling
4096: 3x3 (4)
4096: 1x1
21: 1x1
Output
Block 1
Block 2
Block 3
Block 4
Block 5
FC6, FC7, FC8

DeepLab-v1 (4)

- Feature extraction with the hole (‘atrous’) algorithm:
  - The kernel size of convolutions do not change
  - Kernels are superimposed on the feature map with holes
  - Atrous rate of the kernel for three convolutional layers is 2, for the first fully convolutional layer is 4
- Illustration of the hole algorithm (atrous rate is 2):



* Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. – 2014. – [https://arxiv.org/pdf/1412.7062.pdf].

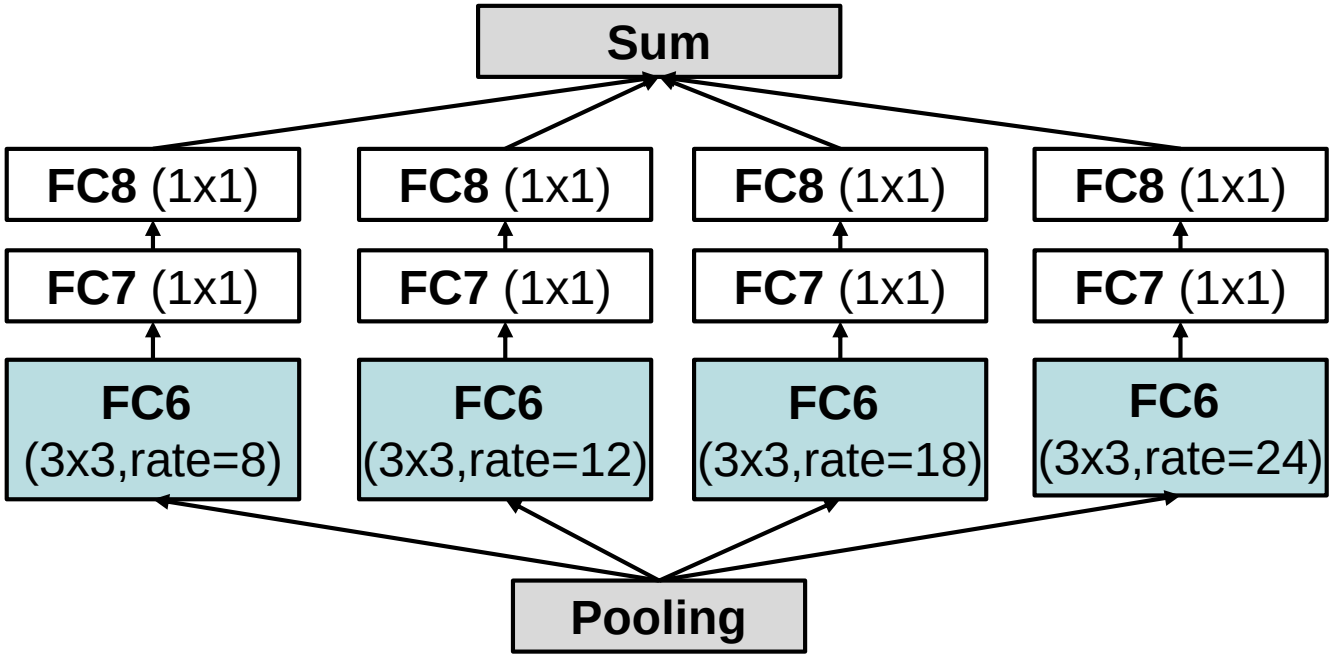
DeepLab-v2 (1)

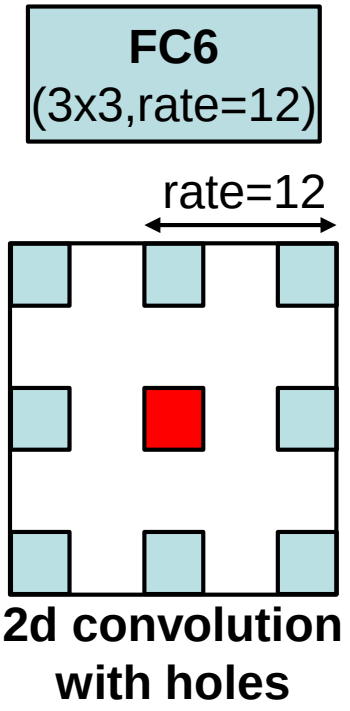
- DeepLab-v2 is a modification of DeepLab-v1 designed to improve the model performance
- It solves the problem of different object scales during segmentation of objects belonging to the same classes
- The classical approach for solving this problem is image scaling and aggregating of feature maps at different scales
- To implement this approach, ***Atrous Spatial Pyramid Pooling*** (ASPP) is introduced
- ASPP combines the results of applying convolutions with holes of different sizes to the feature map

* Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. – 2017. – [https://arxiv.org/pdf/1606.00915.pdf], [https://ieeexplore.ieee.org/document/7913730].

DeepLab-v2 (2)

- The structure of ASPP
  (FC6, FC7, FC8 – fully convolutional layers):





* Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. – 2017. – [https://arxiv.org/pdf/1606.00915.pdf], [https://ieeexplore.ieee.org/document/7913730].

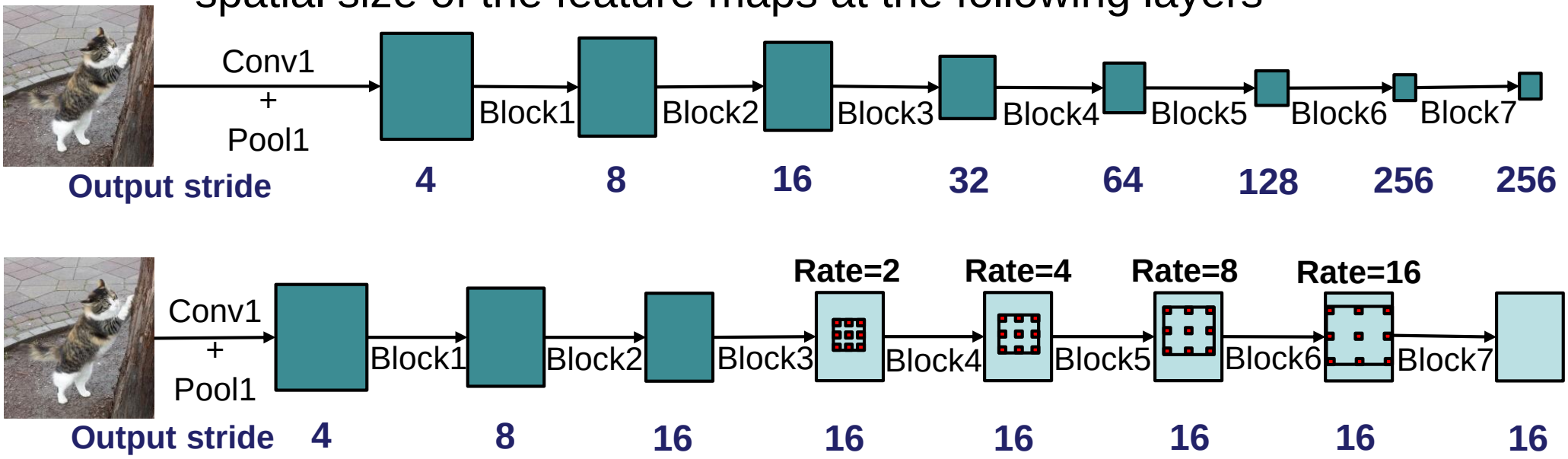
DeepLab-v3 (1)

- DeepLab-v3 is an improvement of the DeepLab-v2 model
- To solve the problem of different object scales, special modules based on convolutions with holes are proposed
- These modules are organized in cascade or parallel manner to capture context from different scales using different atrous rate
- The module organized in parallel manner is an extension of the atrous spatial pyramid pooling

* Chen L.-C., Papandreou G., Schroff F., Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. – 2017. – [https://arxiv.org/pdf/1706.05587.pdf].

DeepLab-v3 (2)

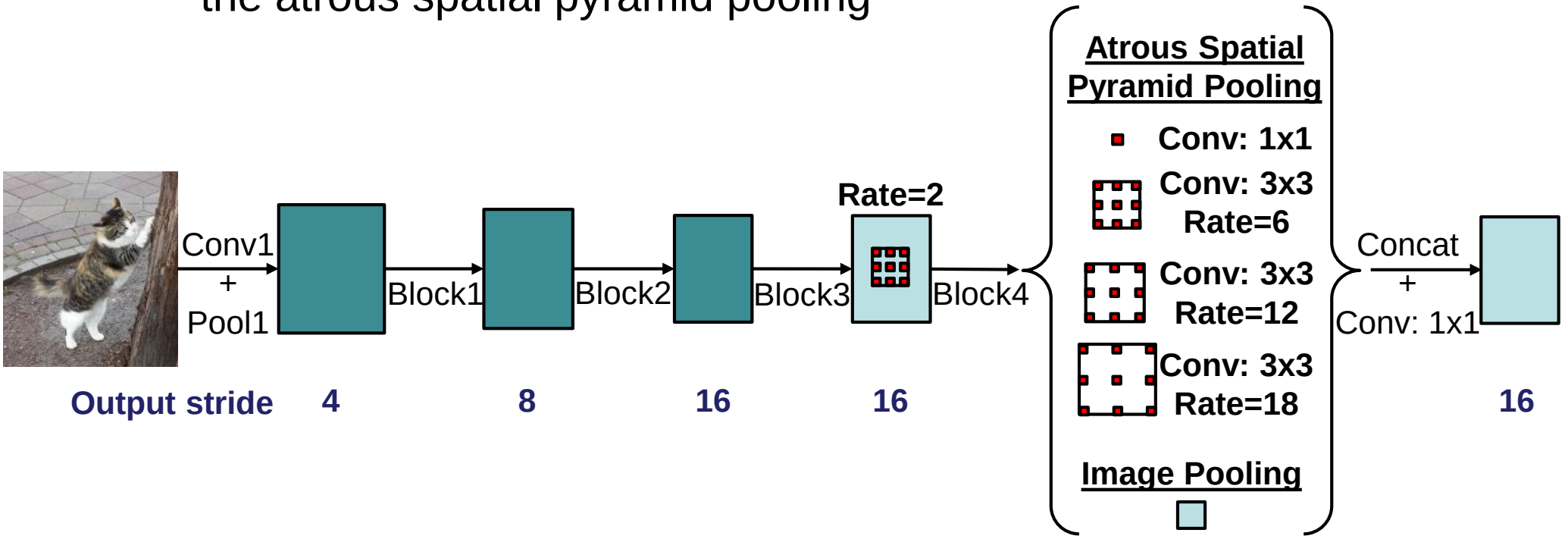
- The cascade module:
  - The model is a sequence of residual blocks
  - The standard convolutions in the last residual blocks are replaced with convolutions with holes, to keep the original spatial size of the feature maps at the following layers



* Chen L.-C., Papandreou G., Schroff F., Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. – 2017. – [https://arxiv.org/pdf/1706.05587.pdf].

DeepLab-v3 (3)

- The parallel module:
  - The image-level features (Image Pooling block) supplement the atrous spatial pyramid pooling



* Chen L.-C., Papandreou G., Schroff F., Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. – 2017. – [https://arxiv.org/pdf/1706.05587.pdf].

DeepLab-v3 (4)

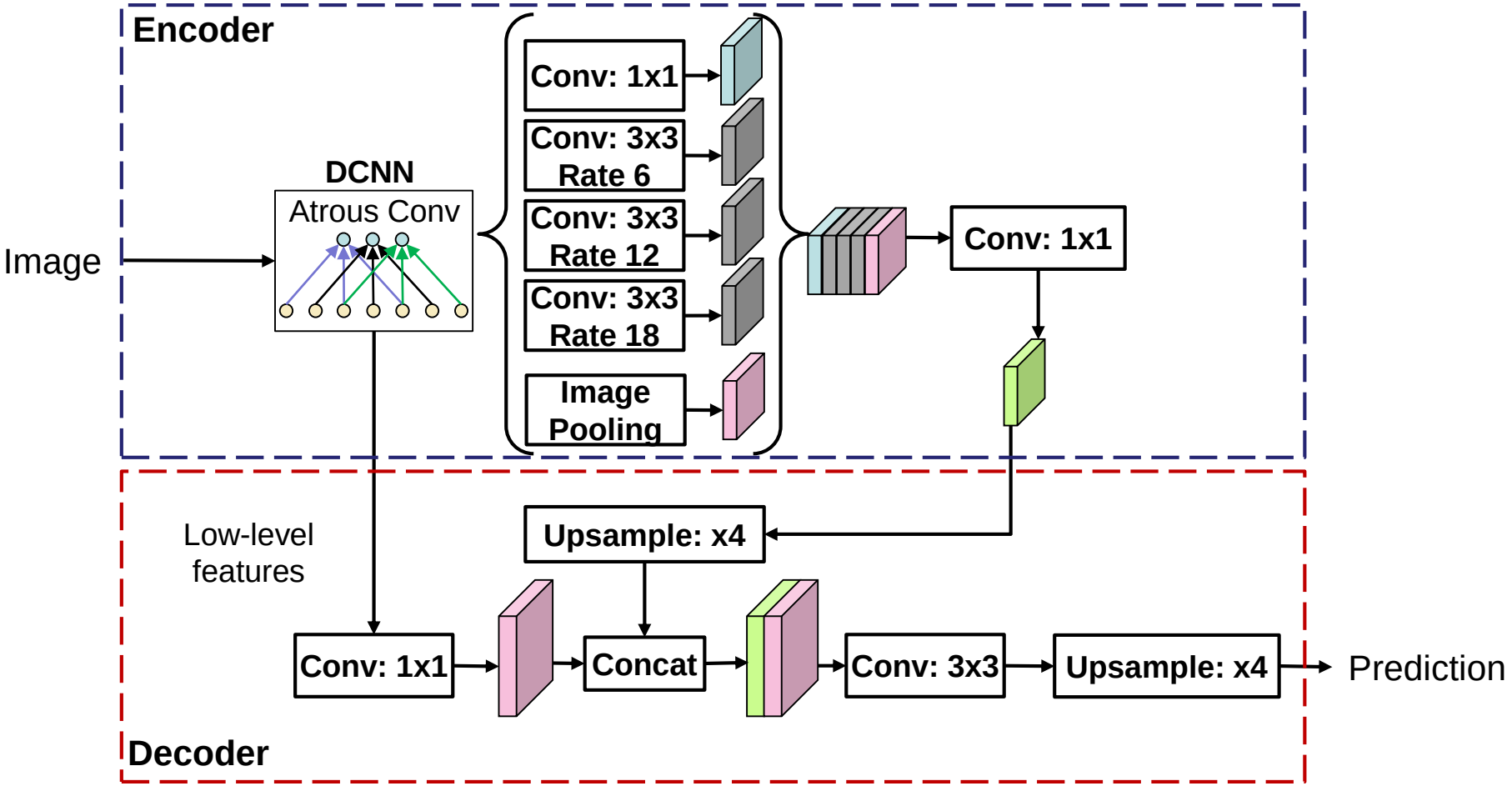
- ASPP:
  - To extract image-level features, the following transformations are performed:
    - Calculation of the global average pooling for the last model feature map
    - 1x1 convolution, 256 channels
    - Batch normalization
    - Bilinear interpolation of the feature map to provide the same spatial size of all feature maps
  - Feature maps from all branches of the pyramid are concatenated, 1x1 convolutions (256 channels), batch normalization and final 1x1 convolution are applied

DeepLab-v3+ (1)

- DeepLab-v3+ is a modification of the DeepLab-v3 model, aimed at the improving quality of object boundaries segmentation
- The model is based on the encoder-decoder architecture:
  - The encoder consists of the basic part of the DeepLab-v3 model (all transformations to the final one-dimensional convolution)
  - The decoder consists of convolutions and upsamplings applied to the image-level feature map and the output of the spatial pyramid pooling

* Chen L.-C., Zhu Y., Papandreou G., Schoff F., Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. – 2018. – [https://arxiv.org/pdf/1802.02611.pdf].

DeepLab-v3+ (2)



* Chen L.-C., Zhu Y., Papandreou G., Schoff F., Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. – 2018. – [https://arxiv.org/pdf/1802.02611.pdf].

DeepLab-v3+ (3)

- To implement the encoder, DeepLab-v3, ResNet-101 or Xception is used (the corresponding experimental results are given in the paper*)
- To optimize calculations, convolutions with 3x3 kernels are converted to the standard depthwise separable convolutions:
  - Each convolution is represented by a depthwise and pointwise convolutions
  - The depthwise convolution involves splitting the feature map into channels, applying 3x3 convolution of the depth 1 to each channel, and concatenating the channels
  - The pointwise convolution is a convolution of the shape 1x1x<channels_number>

* Chen L.-C., Zhu Y., Papandreou G., Schoff F., Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. – 2018. – [https://arxiv.org/pdf/1802.02611.pdf].

COMPARISON OF DEEP MODELS FOR SEMANTIC SEGMENTATION

Comparison of deep models for semantic segmentation (1)

- The problem is semantic segmentation of on-road images
- The test dataset is Cityscapes [https://www.cityscapes-dataset.com]
- Quality metric is mean Intersection over Union (mean IoU)

- Comparison* “quality-performance” is qualitative, since the given experiments collected from the original papers and obtained on different test infrastructures

- Results of semantic segmentation for another datasets are available by link**

* Real-Time Semantic Segmentation on Cityscapes test [https://paperswithcode.com/sota/real-time-semantic-segmentation-on-cityscapes].
** Semantic Segmentation [https://paperswithcode.com/task/semantic-segmentation/latest].

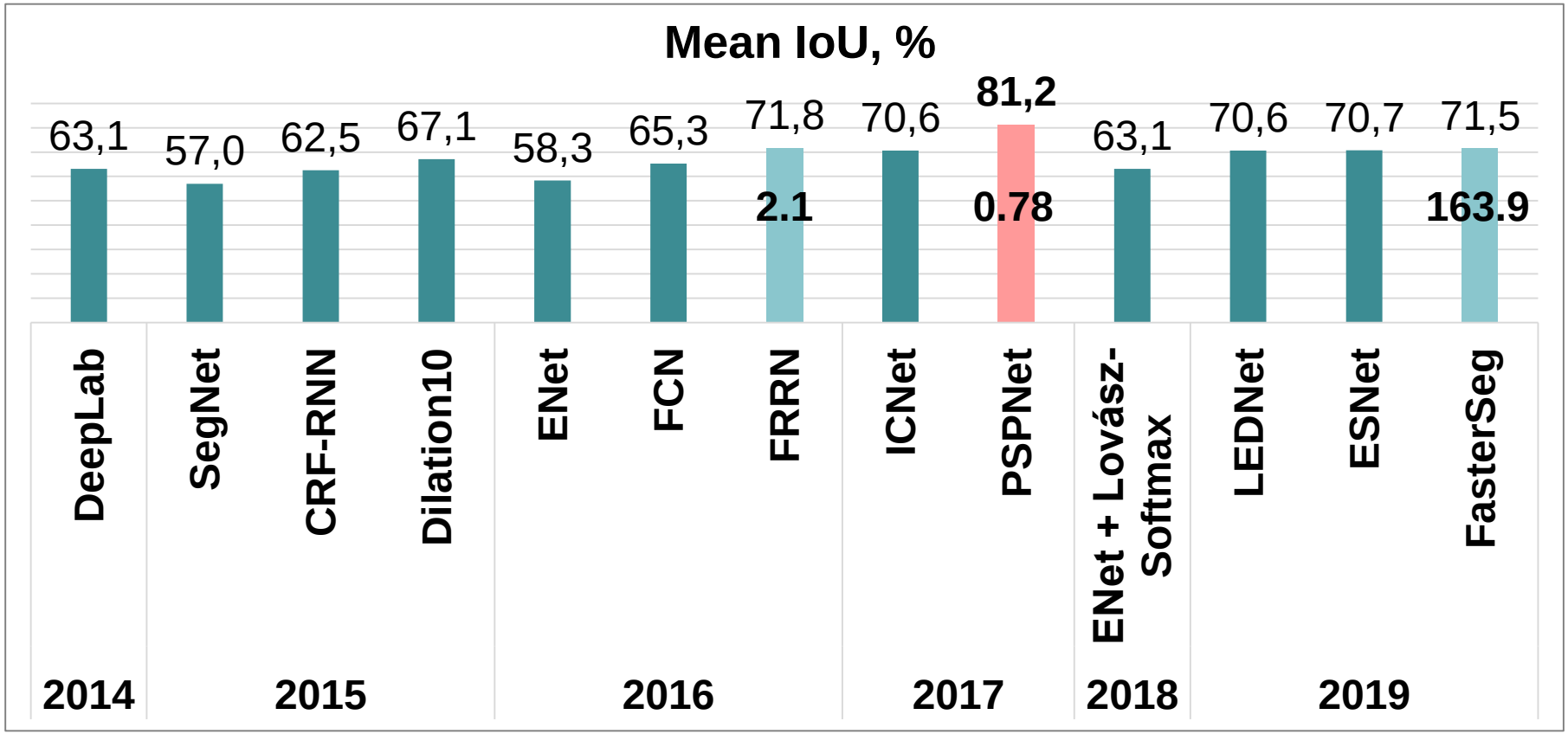
Comparison of deep models for semantic segmentation (2)

Model	Year	Mean IoU, %	FPS	Time, ms
DeepLab	2014	63.1	0.25	4000
SegNet	2015	57.0	16.7	60
CRF-RNN	2015	62.5	1.4	700
Dilation10	2015	67.1	0.25	4000
ENet	2016	58.3	76.9	13
FCN	2016	65.3	2	500
FRRN	2016	**71.8**	2.1	469
ICNet	2017	70.6	30.3	33
PSPNet	2017	**81.2**	**0.78**	**1288**
ENet + Lovász-Softmax	2018	63.1	76.9	13
LEDNet	2019	70.6	71	14
ESNet	2019	70.7	63	16
FasterSeg	2019	**71.5**	163.9	**6.1**

* Real-Time Semantic Segmentation on Cityscapes test [https://paperswithcode.com/sota/real-time-semantic-segmentation-on-cityscapes].

Comparison of deep models for semantic segmentation (3)

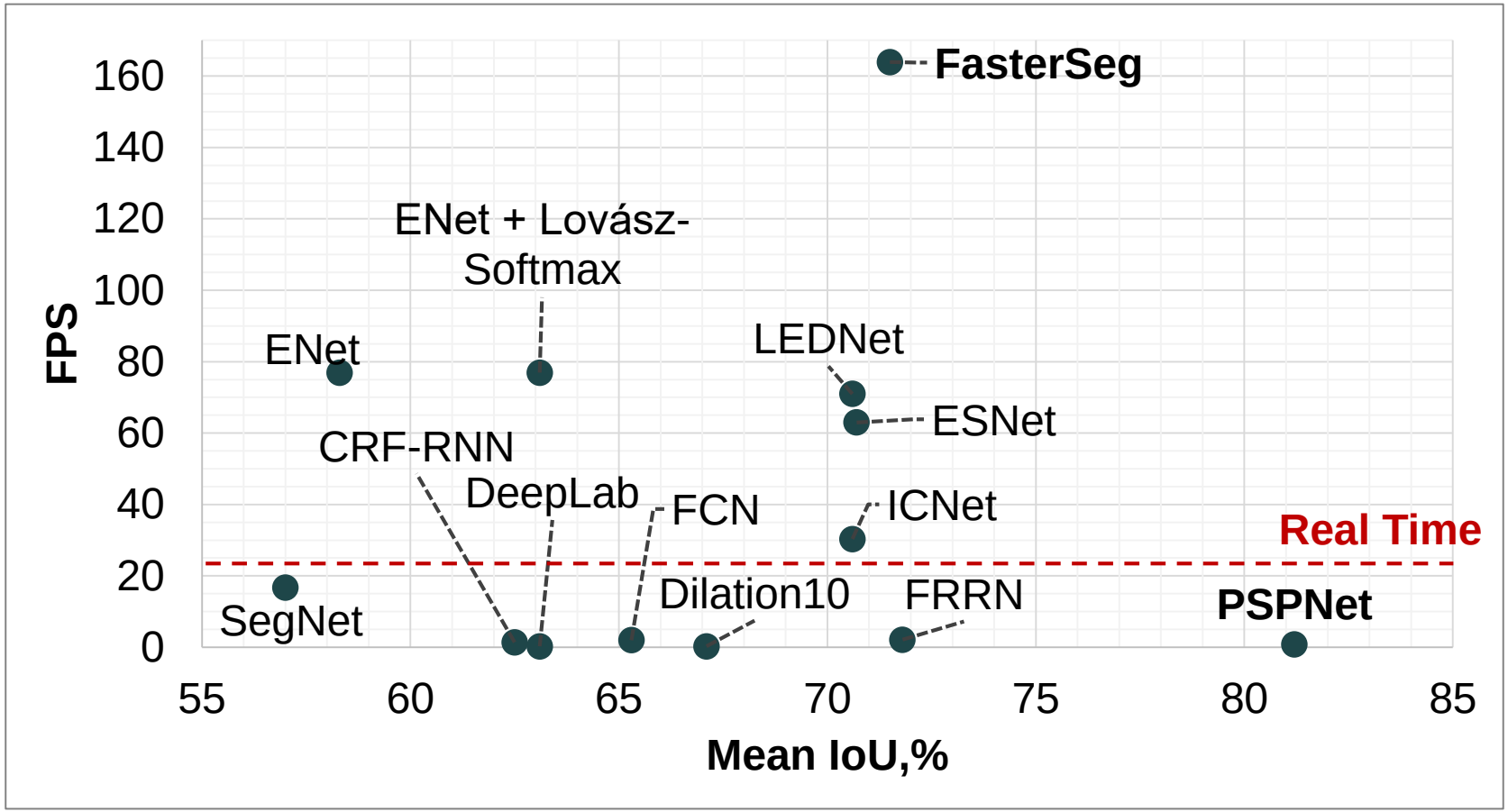
- Mean IoU for the selected models:



- ***For 2017-2019 years, the quality varies from ~70 to ~81%, and the best model is the slowest one***

Comparison of deep models for semantic segmentation (4)

- ***An effective model is a compromise between quality and performance***



Conclusion

- Models for semantic segmentation are not limited to those discussed in the lecture
- The main problem constructing segmentation models is to obtain an output whose spatial resolution is equal to the input image spatial resolution
- The considered models solve this problem in different ways. As a rule, the decision greatly affects the performance
- ***The optimal model is a compromise between quality and complexity***
  - Quality is determined by the requirements for solving a practical problem
  - Complexity is determined by the available computational resources and inference time requirements

Literature (1)

- Long J., Shelhamer E., Darrel T. Fully Convolutional Networks for Semantic Segmentation. – 2015. – [https://arxiv.org/pdf/1411.4038.pdf], [https://ieeexplore.ieee.org/document/7298965].
- Badrinarayanan V., Kendall A., Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. – 2015. – [https://arxiv.org/pdf/1511.00561.pdf], [https://ieeexplore.ieee.org/document/7803544].
- Ronneberger O., Fischer P., Brox T. U-net: Convolutional networks for biomedical image segmentation. – 2015. – [https://arxiv.org/pdf/1505.04597.pdf], [https://link.springer.com/chapter/10.1007/978-3-319-24574-4_28].

Literature (2)

- Zhao H., Shi J., Qi X., Wang X., Jia J. Pyramid scene parsing network. – 2016. – [https://arxiv.org/pdf/1612.01105.pdf], [https://ieeexplore.ieee.org/document/8100143].
- Zhao H., Qi X., Shen X., Shi J., Jia J. ICNet for Real-Time Semantic Segmentation on High-Resolution Images. – 2017. – [https://arxiv.org/pdf/1704.08545.pdf], [https://link.springer.com/chapter/10.1007/978-3-030-01219-9_25].
- Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. – 2014. – [https://arxiv.org/pdf/1412.7062.pdf].

Literature (3)

- Chen L.-C., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. – 2017. – [https://arxiv.org/pdf/1606.00915.pdf], [https://ieeexplore.ieee.org/document/7913730].
- Chen L.-C., Papandreou G., Schroff F., Adam H. Rethinking Atrous Convolution for Semantic Image Segmentation. – 2017. – [https://arxiv.org/pdf/1706.05587.pdf].
- Chen L.-C., Zhu Y., Papandreou G., Schoff F., Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. – 2018. – [https://arxiv.org/pdf/1802.02611.pdf].

Authors

- **Turlapov Vadim Evgenievich,** Dr., Prof., department of computer software and supercomputer technologies
  vadim.turlapov@itmm.unn.ru
- **Vasiliev Engeny Pavlovich,** lecturer, department of computer software and supercomputer technologies
  evgeny.vasiliev@itmm.unn.ru
- **Getmanskaya Alexandra Alexandrovna,** lecturer, department of computer software and supercomputer technologies
  alexandra.getmanskaya@itmm.unn.ru
- **Kustikova Valentina Dmitrievna**
  Phd, assistant professor, department of computer software and supercomputer technologies
  valentina.kustikova@itmm.unn.ru