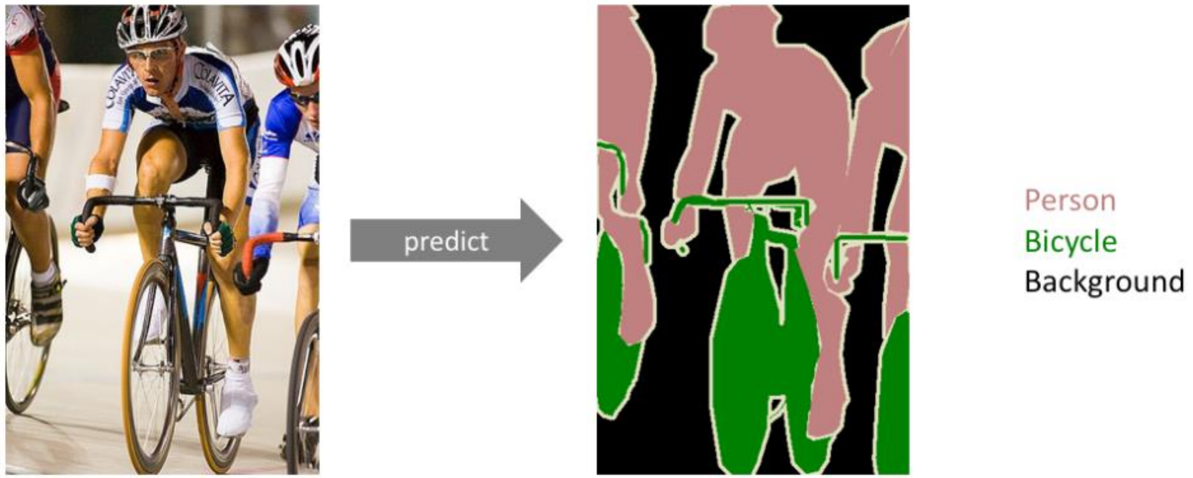
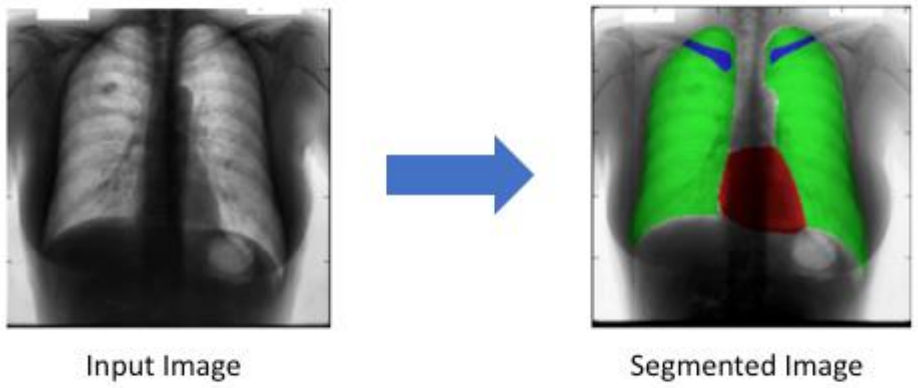
AI for Video Data Processing 2:
Fully Convolutional Networks

Outline

1. Semantic Segmentation

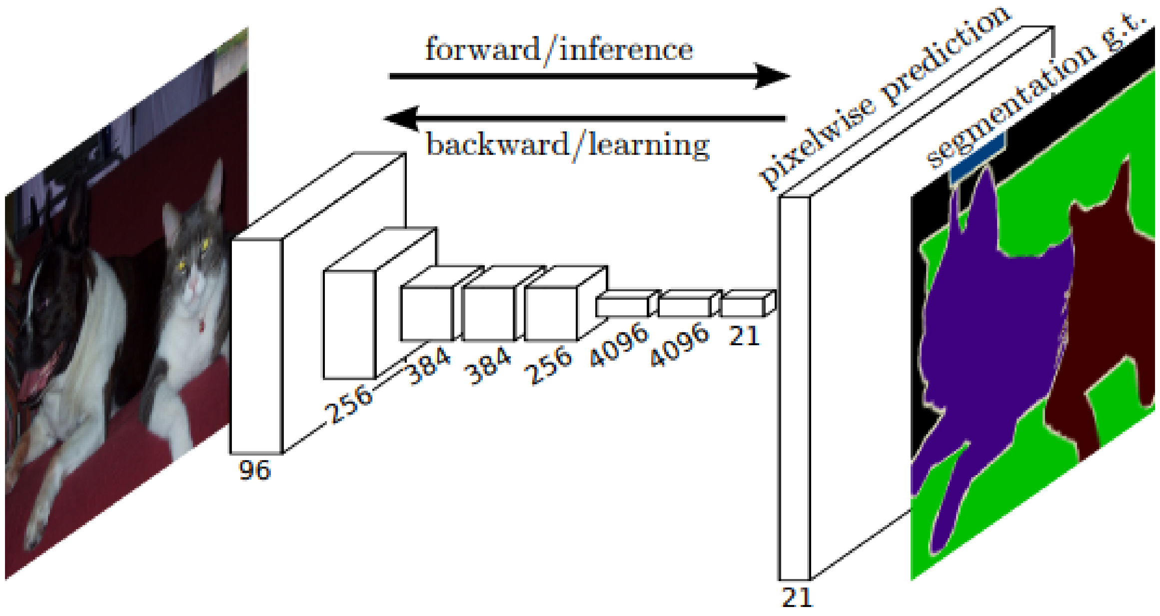
Goal of semantic segmentation: label *each pixel* of an image with a corresponding ***class***





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

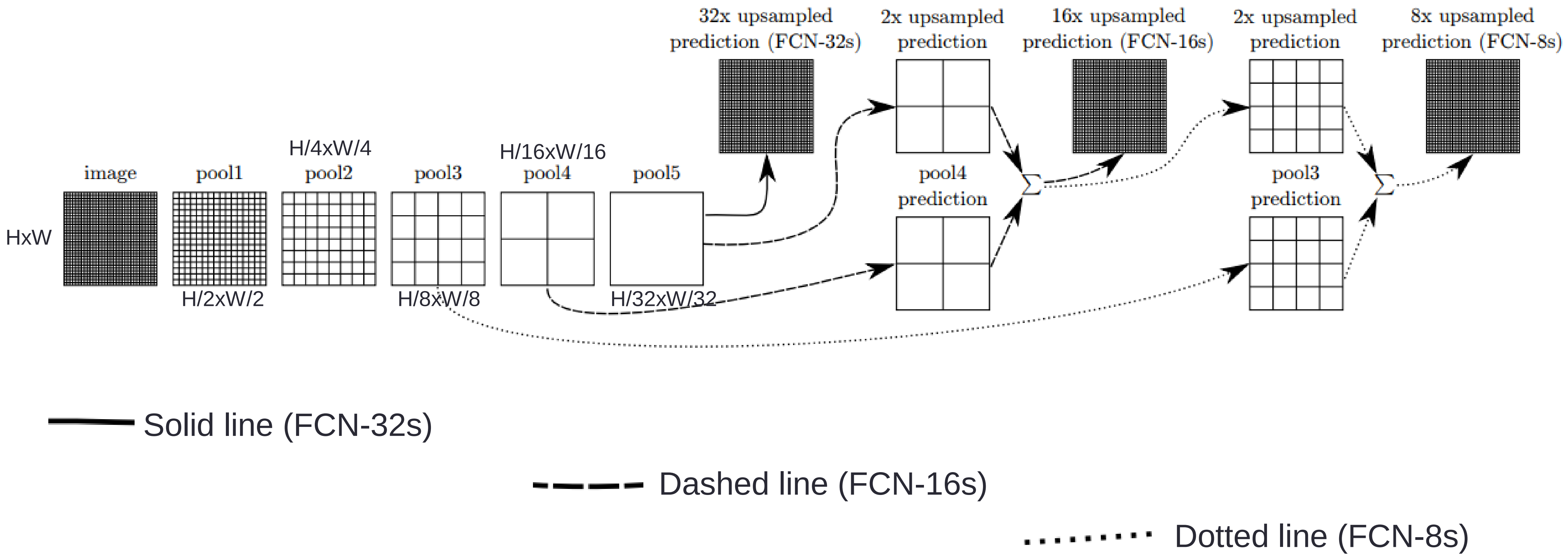
1. FCN



No fully-connected layers

Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431-3440. 2015.

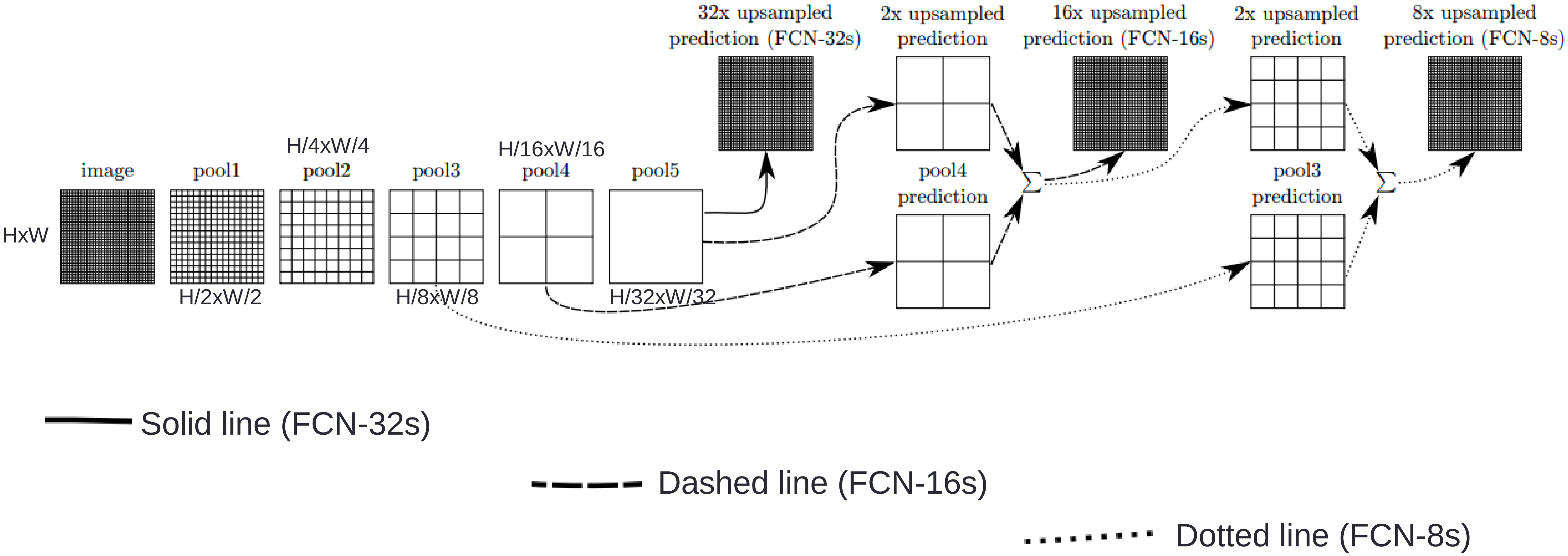
1. FCN



Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431-3440. 2015.

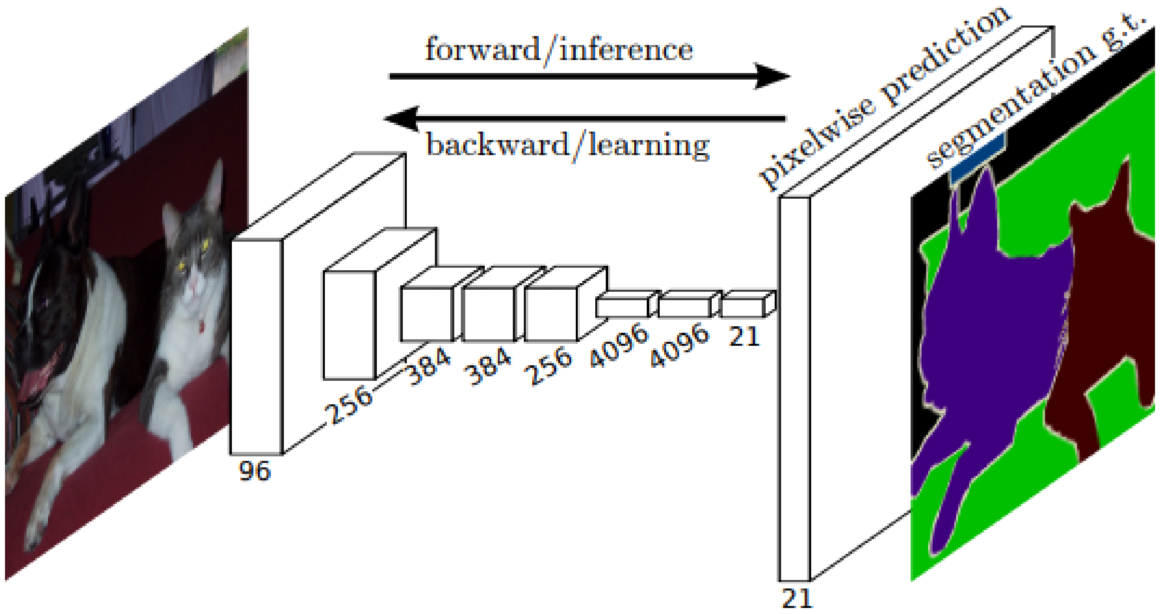
1. FCN

Brainstorm: Let us discuss the differences between FCN-32S, FCN-16S, and FCN-8S. Which one is more accurate?

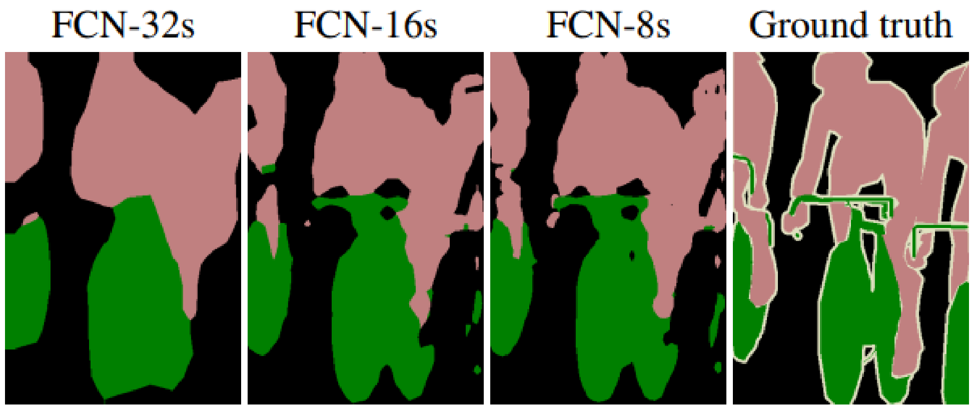


Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431-3440. 2015.

1. FCN



No fully-connected layers



	pixel acc.	mean acc.	mean IU	f.w. IU
FCN-32s-fixed	83.0	59.7	45.4	72.0
FCN-32s	89.1	73.3	59.4	81.4
FCN-16s	90.0	75.7	62.4	83.0
FCN-8s	**90.3**	**75.9**	**62.7**	**83.2**

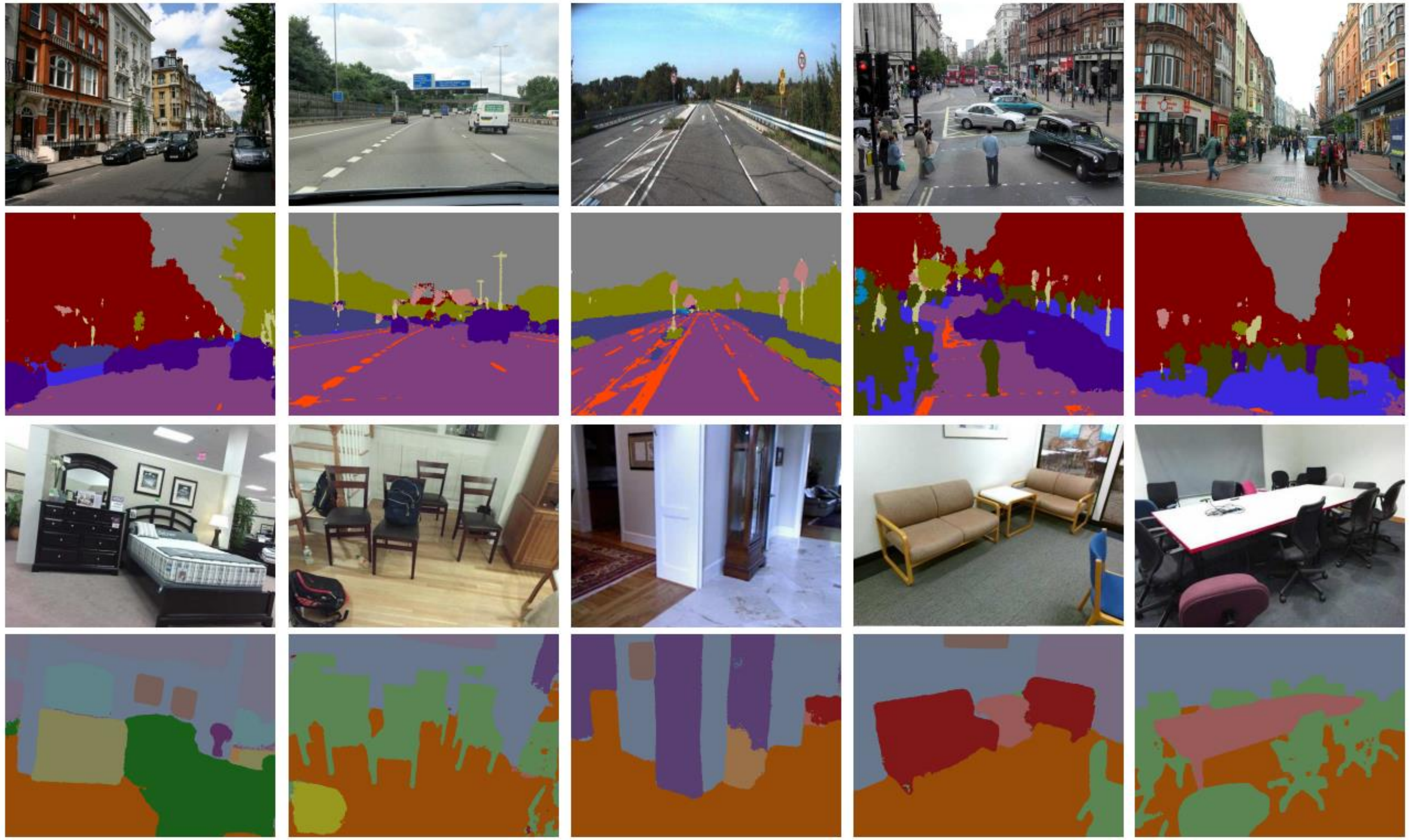
Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431-3440. 2015.

1. FCN

Compatible with many CNN architectures

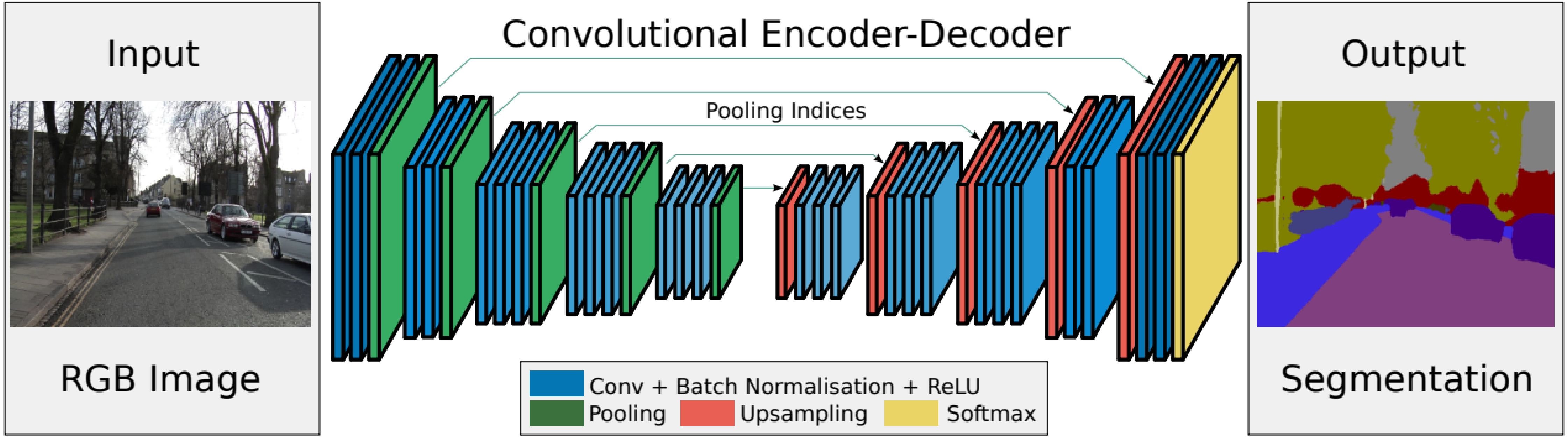
	FCN-AlexNet	FCN-VGG16	FCN-GoogLeNet[4]
mean IU	39.8	**56.0**	42.5
forward time	50 ms	210 ms	59 ms
conv. layers	8	16	22
parameters	57M	134M	6M
rf size	355	404	907
max stride	32	32	32

2. SegNet



Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 39, no. 12 (2017): 2481-2495.

2. SegNet



- 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

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 39, no. 12 (2017): 2481-2495.

2. SegNet

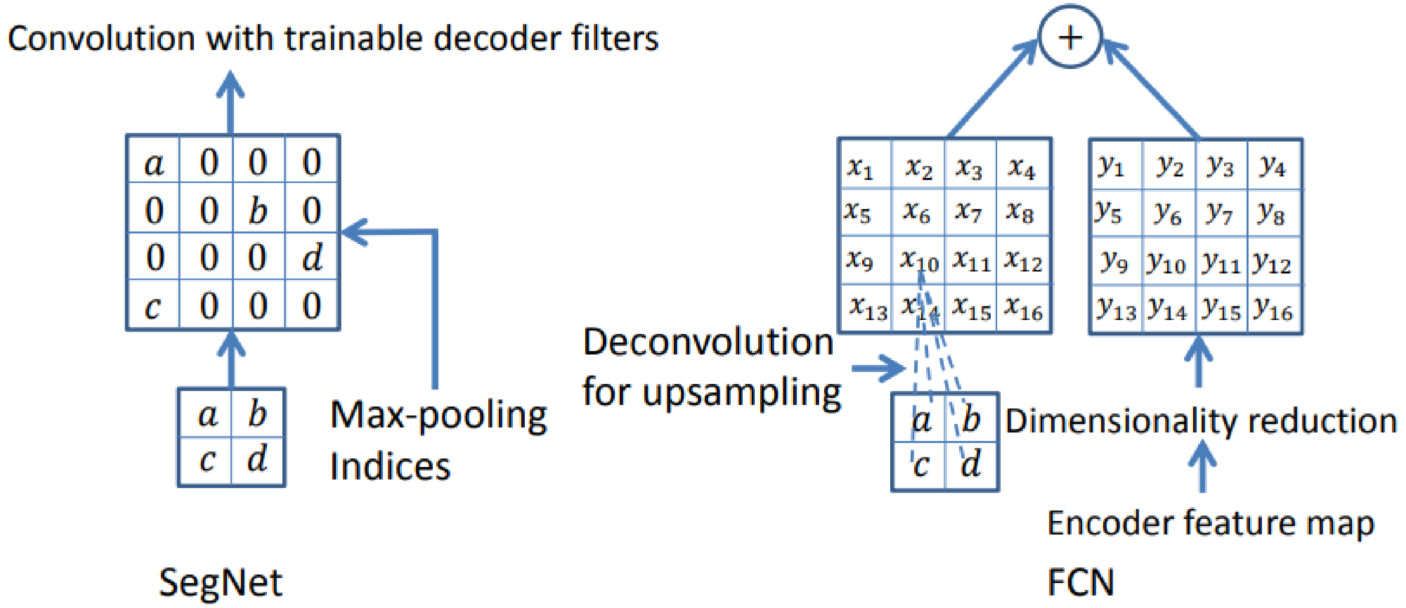


Fig. 3. An illustration of SegNet and FCN [2] decoders. a, b, c, d correspond to values in a feature map. SegNet uses the max pooling indices to upsample (without learning) the feature map(s) and convolves with a trainable decoder filter bank. FCN upsamples by learning to deconvolve the input feature map and adds the corresponding encoder feature map to produce the decoder output. This feature map is the output of the max-pooling layer (includes sub-sampling) in the corresponding encoder. Note that there are no trainable decoder filters in FCN.

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 39, no. 12 (2017): 2481-2495.

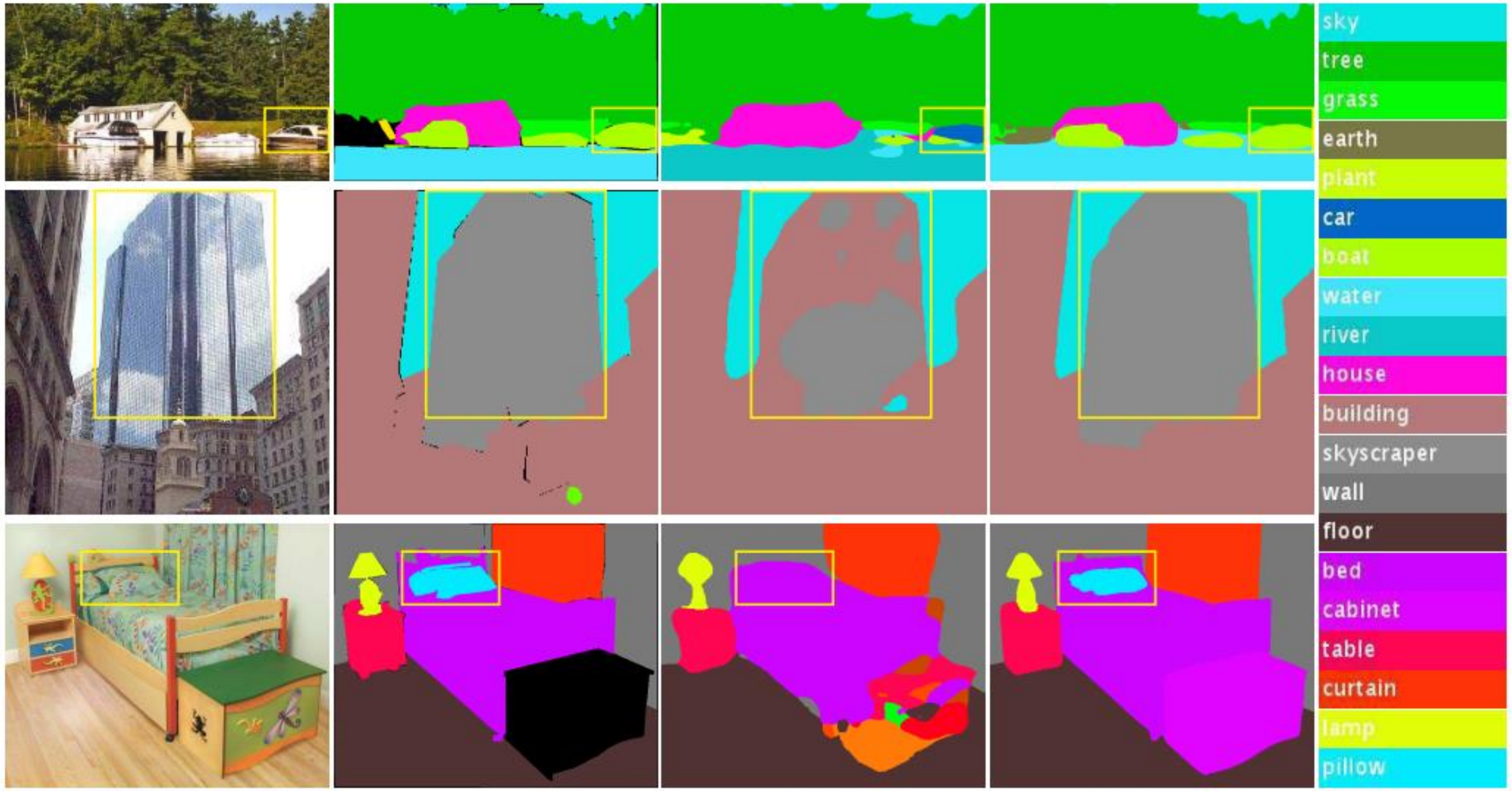
2. SegNet

Semantic segmentation on the CamVid test set for autonomous driving

Network/Iterations	40K				80K				>80K				Max iter
	G	C	mIoU	BF	G	C	mIoU	BF	G	C	mIoU	BF	
SegNet	88.81	59.93	50.02	35.78	89.68	69.82	57.18	42.08	90.40	71.20	60.10	46.84	140K
DeepLab-LargeFOV [3]	85.95	60.41	50.18	26.25	87.76	62.57	53.34	32.04	88.20	62.53	53.88	32.77	140K
DeepLab-LargeFOV-denseCRF [3]	not computed								89.71	60.67	54.74	40.79	140K
FCN	81.97	54.38	46.59	22.86	82.71	56.22	47.95	24.76	83.27	59.56	49.83	27.99	200K
FCN (learnt deconv) [2]	83.21	56.05	48.68	27.40	83.71	59.64	50.80	31.01	83.14	64.21	51.96	33.18	160K
DeconvNet [4]	85.26	46.40	39.69	27.36	85.19	54.08	43.74	29.33	89.58	70.24	59.77	52.23	260K

Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 39, no. 12 (2017): 2481-2495.

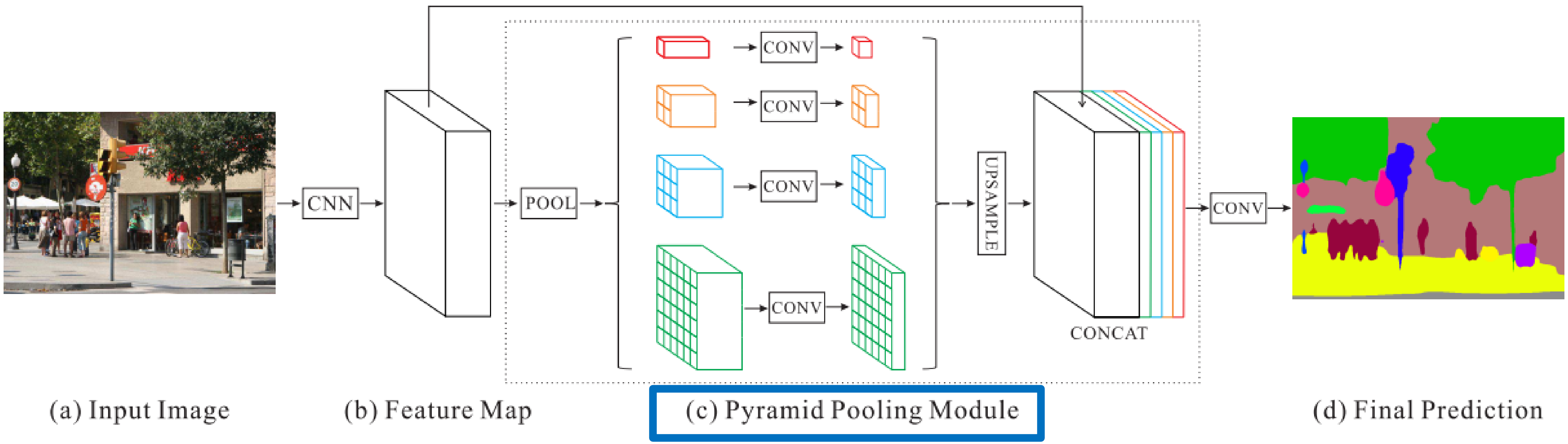
3. PSPNet



(a) Image (b) Ground Truth (c) FCN (d) PSPNet (e) ColorMap

Zhao, Hengshuang, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. "Pyramid scene parsing network." *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881-2890. 2017.

3. PSPNet

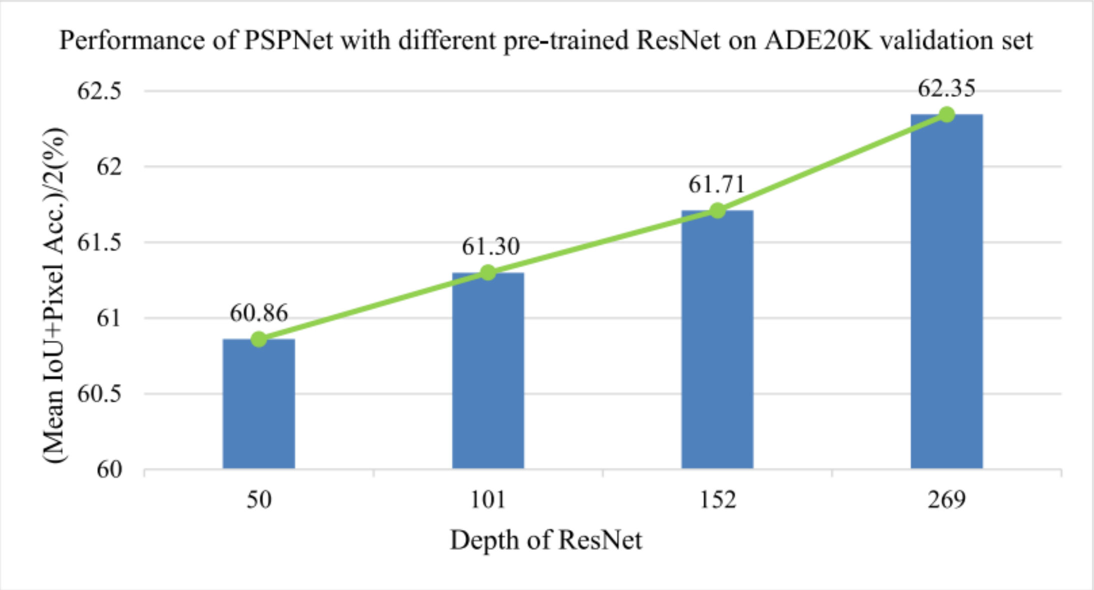


Zhao, Hengshuang, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. "Pyramid scene parsing network." *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881-2890. 2017.

3. PSPNet

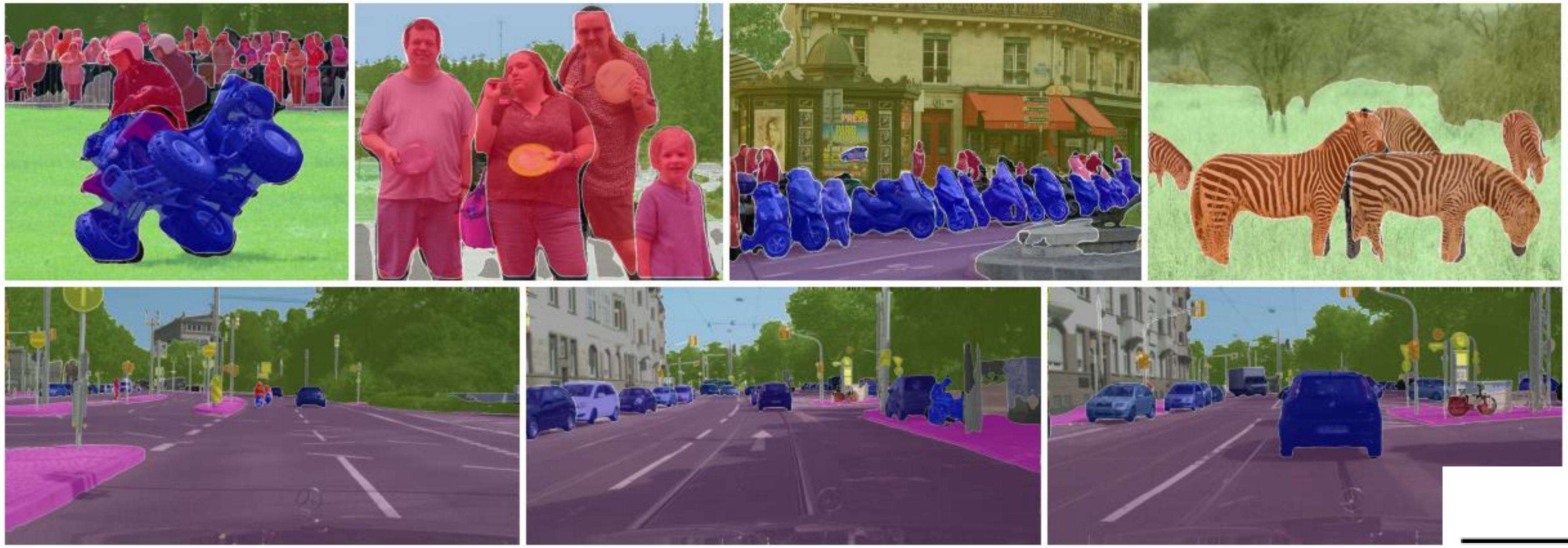
Rank	Team Name	Final Score (%)
1	**Ours**	**57.21**
2	Adelaide	56.74
3	360+MCG-ICT-CAS_SP	55.56
-	(our single model)	(55.38)
4	SegModel	54.65
5	CASIA_IVA	54.33
-	DilatedNet [40]	45.67
-	FCN [26]	44.80
-	SegNet [2]	40.79

Table 5. Results of ImageNet scene parsing challenge 2016.



Zhao, Hengshuang, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. "Pyramid scene parsing network." *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881-2890. 2017.

4. Panoptic FPN

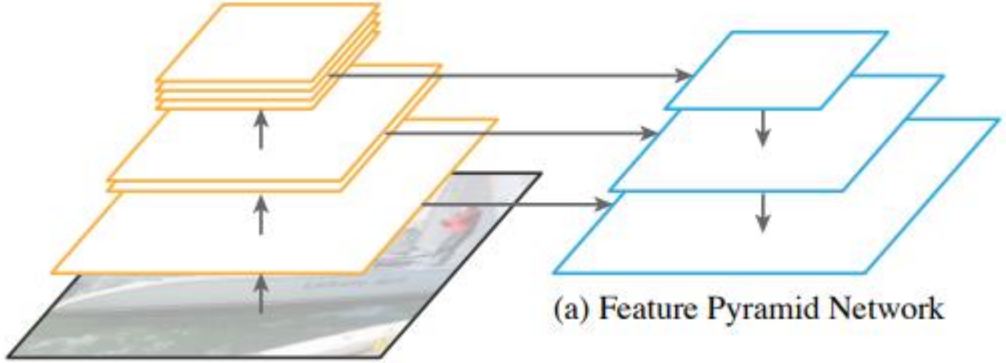


Cityscapes Semantic Segmentation

	backbone	mIoU
DeeplabV3 [11]	ResNet-101-D8	77.8
PSANet101 [59]	ResNet-101-D8	77.9
Mapillary [5]	WideResNet-38-D8	79.4
DeeplabV3+ [12]	X-71-D16	79.6
Semantic FPN	ResNet-101-FPN	77.7
Semantic FPN	ResNeXt-101-FPN	79.1

Kirillov, Alexander, Ross Girshick, Kaiming He, and Piotr Dollár. "Panoptic feature pyramid networks." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6399-6408. 2019.

4. Panoptic FPN



(a) Feature Pyramid Network

Multi-scale deep features

(better than regular CNN features)

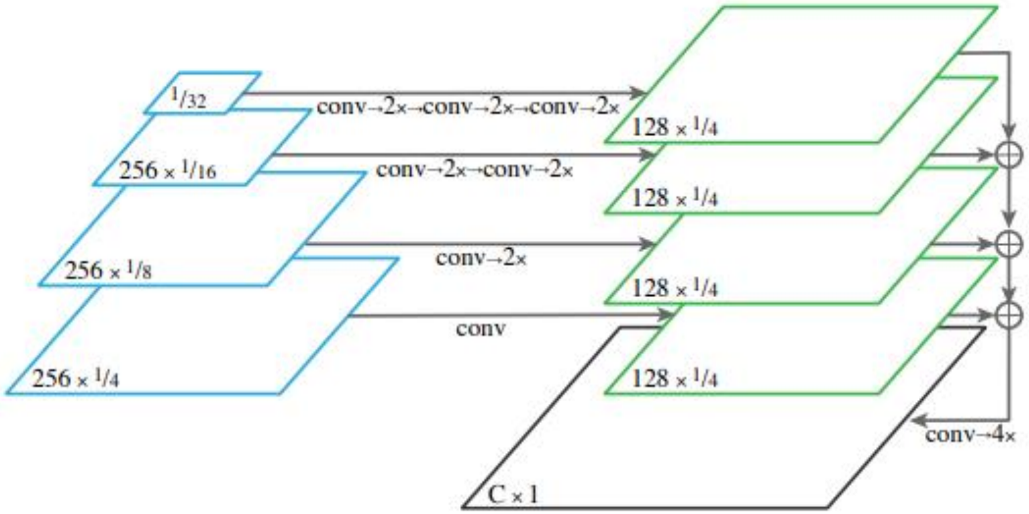
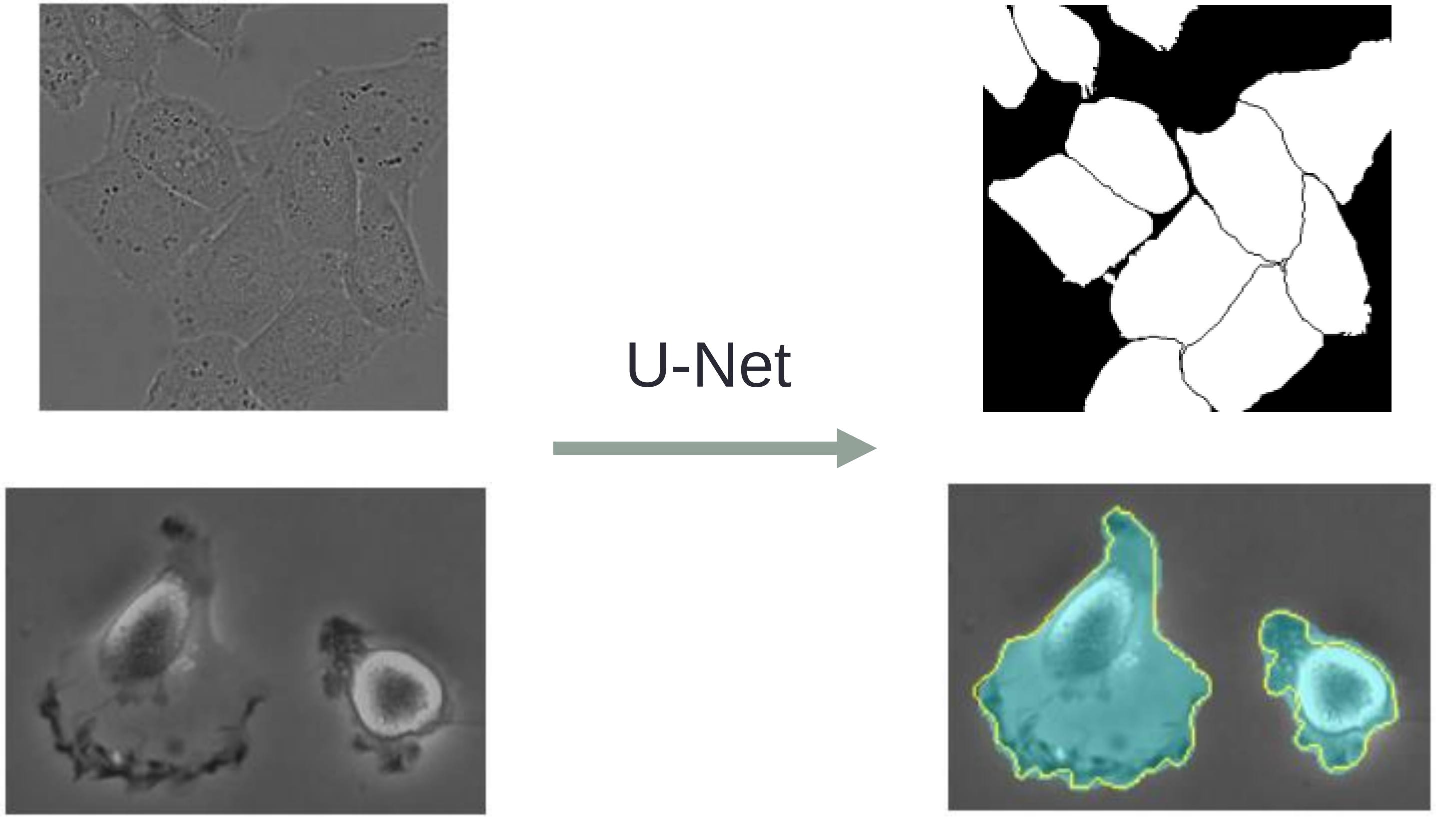


Figure 3: **Semantic segmentation branch.** Each FPN level (left) is upsampled by convolutions and bilinear upsampling until it reaches 1/4 scale (right), theses outputs are then summed and finally transformed into a pixel-wise output.

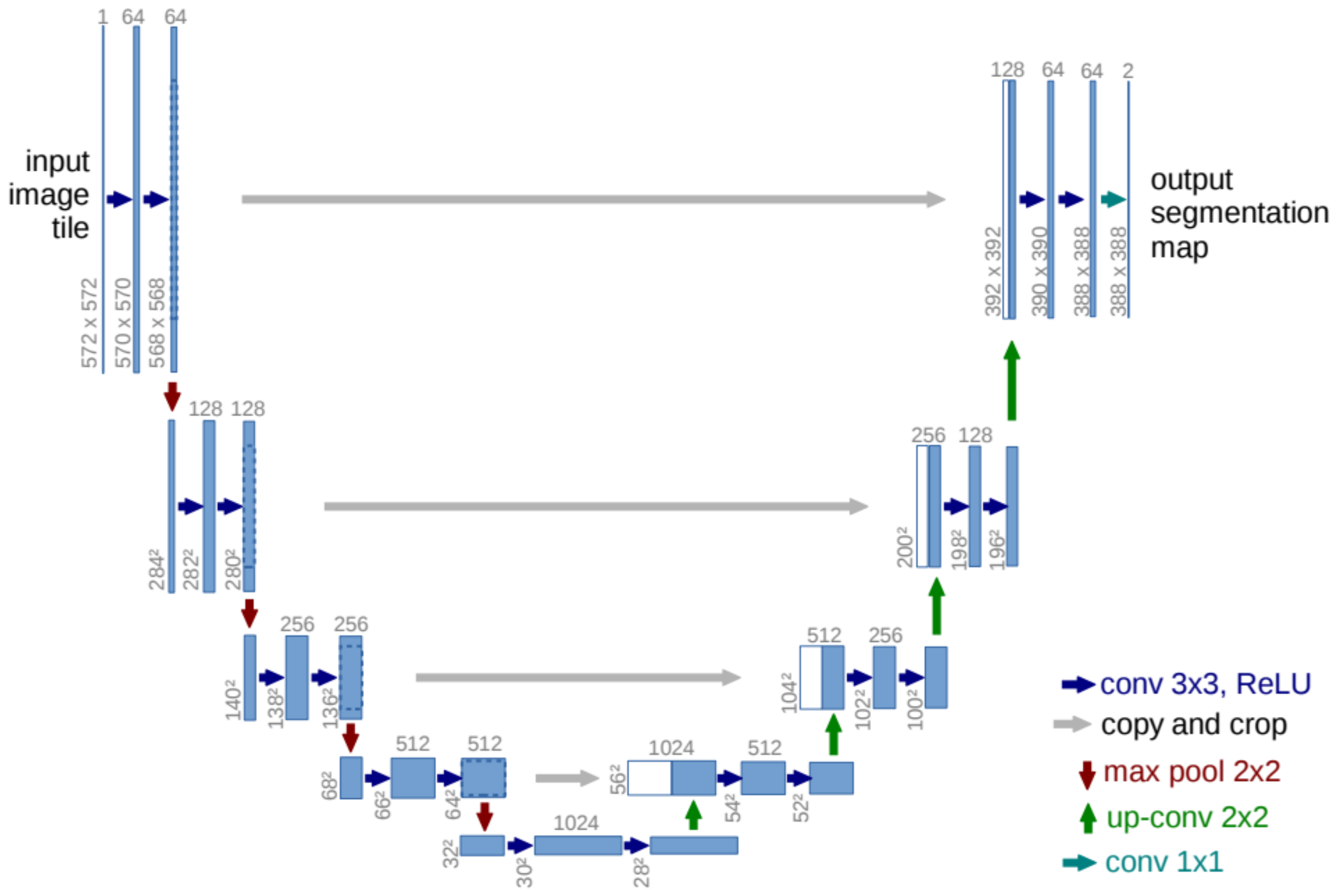
Kirillov, Alexander, Ross Girshick, Kaiming He, and Piotr Dollár. "Panoptic feature pyramid networks." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6399-6408. 2019.

5. U-Net



Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234-241. Springer, Cham, 2015.

5. U-Net



Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234-241. Springer, Cham, 2015.

5. U-Net

Table 2. Segmentation results (IOU) on the ISBI cell tracking challenge 2015.

Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	**0.9203**	**0.7756**

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234-241. Springer, Cham, 2015.

6. DeepLabv3

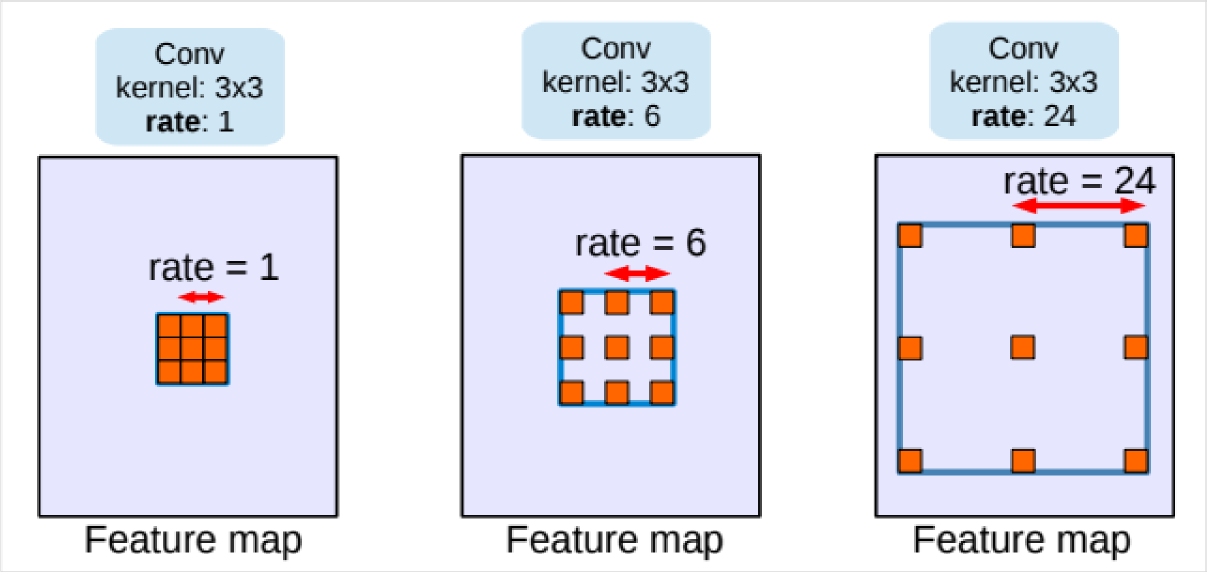
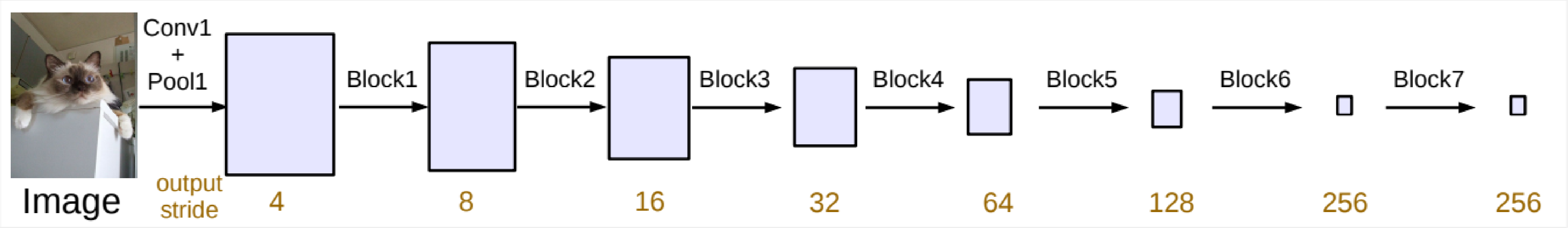


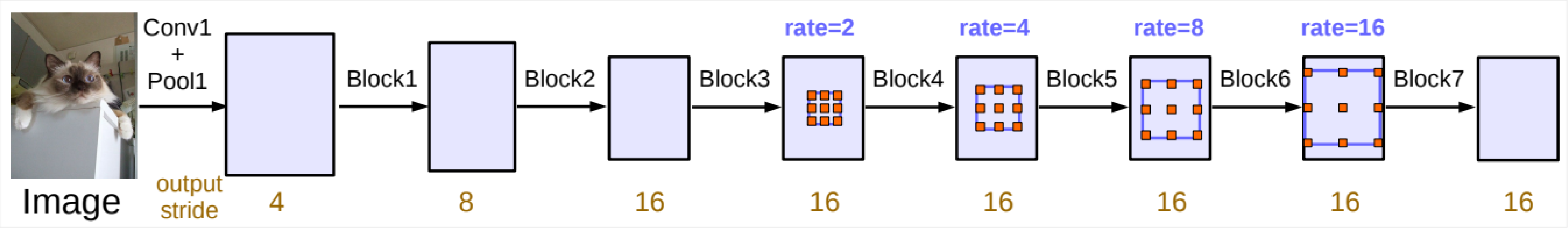
Figure 1. Atrous convolution with kernel size 3×3 and different rates. Standard convolution corresponds to atrous convolution with $rate = 1$. Employing large value of atrous rate enlarges the model's field-of-view, enabling object encoding at multiple scales.

Chen, Liang-Chieh, George Papandreou, Florian Schroff, and Hartwig Adam. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).

6. DeepLabv3



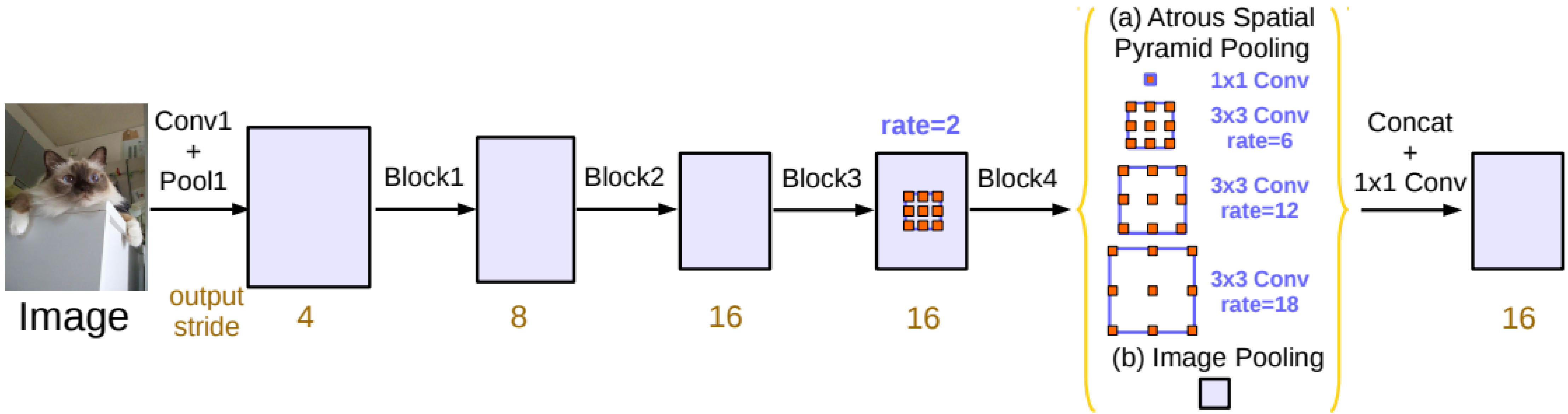
(a) Going deeper without atrous convolution.



(b) Going deeper with atrous convolution. Atrous convolution with $rate > 1$ is applied after block3 when $output\_stride = 16$.

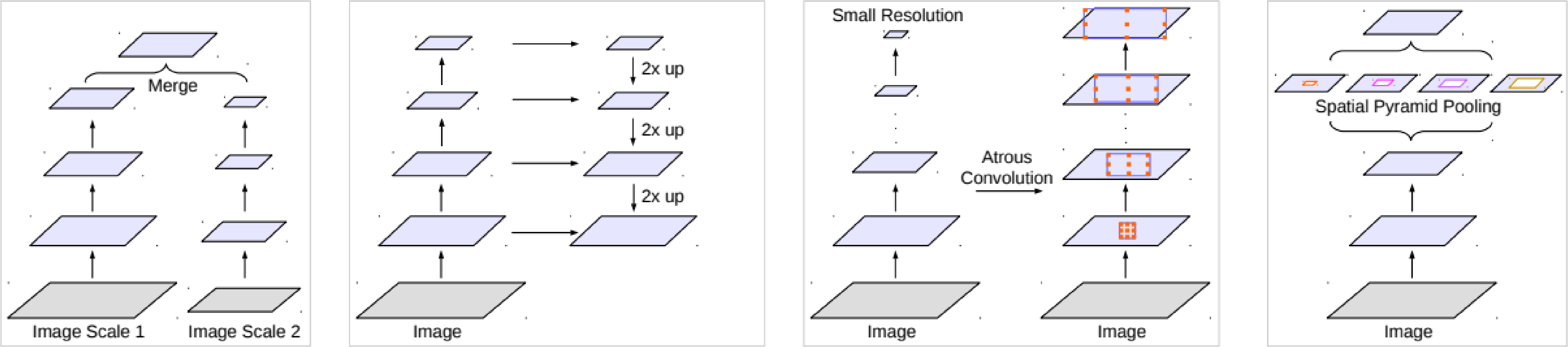
Chen, Liang-Chieh, George Papandreou, Florian Schroff, and Hartwig Adam. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).

6. DeepLabv3



Chen, Liang-Chieh, George Papandreou, Florian Schroff, and Hartwig Adam. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).

1-6. Semantic Segmentation Summary: multi-scale context



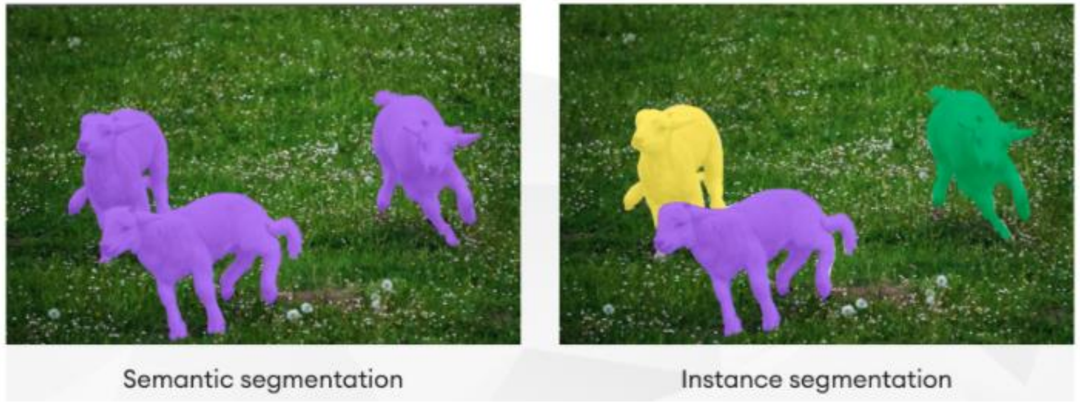
(a) Image Pyramid (b) Encoder-Decoder (c) Deeper w. Atrous Convolution (d) Spatial Pyramid Pooling

Figure 2. Alternative architectures to capture multi-scale context.

Chen, Liang-Chieh, George Papandreou, Florian Schroff, and Hartwig Adam. "Rethinking atrous convolution for semantic image segmentation." arXiv preprint arXiv:1706.05587 (2017).

7. Instance Segmentation

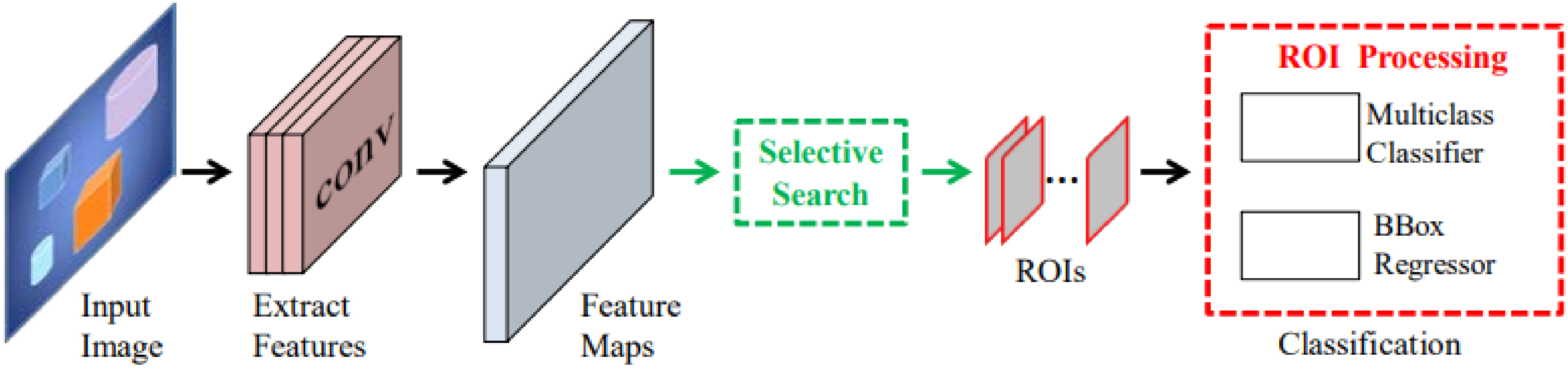
Object Detection Semantic Segmentation **Instance Segmentation**



https://towardsdatascience.com/single-stage-instance-segmentation-a-review-1eeb66e0cc49
https://blog.superannotate.com/guide-to-semantic-segmentation/

7. Instance Segmentation

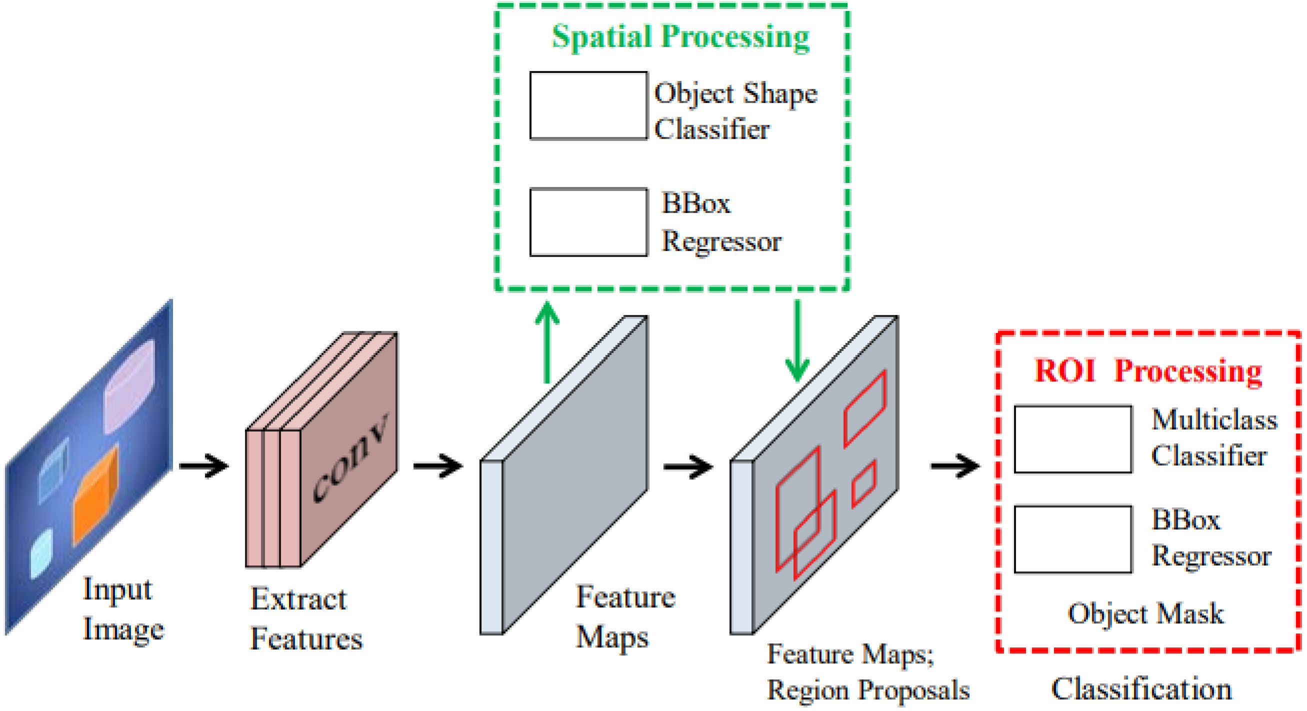
7.1 Classification of mask proposals



Hafiz, Abdul Mueed, and Ghulam Mohiuddin Bhat. "A survey on instance segmentation: state of the art." *International Journal of Multimedia Information Retrieval* 9, no. 3 (2020): 171-189.

7. Instance Segmentation

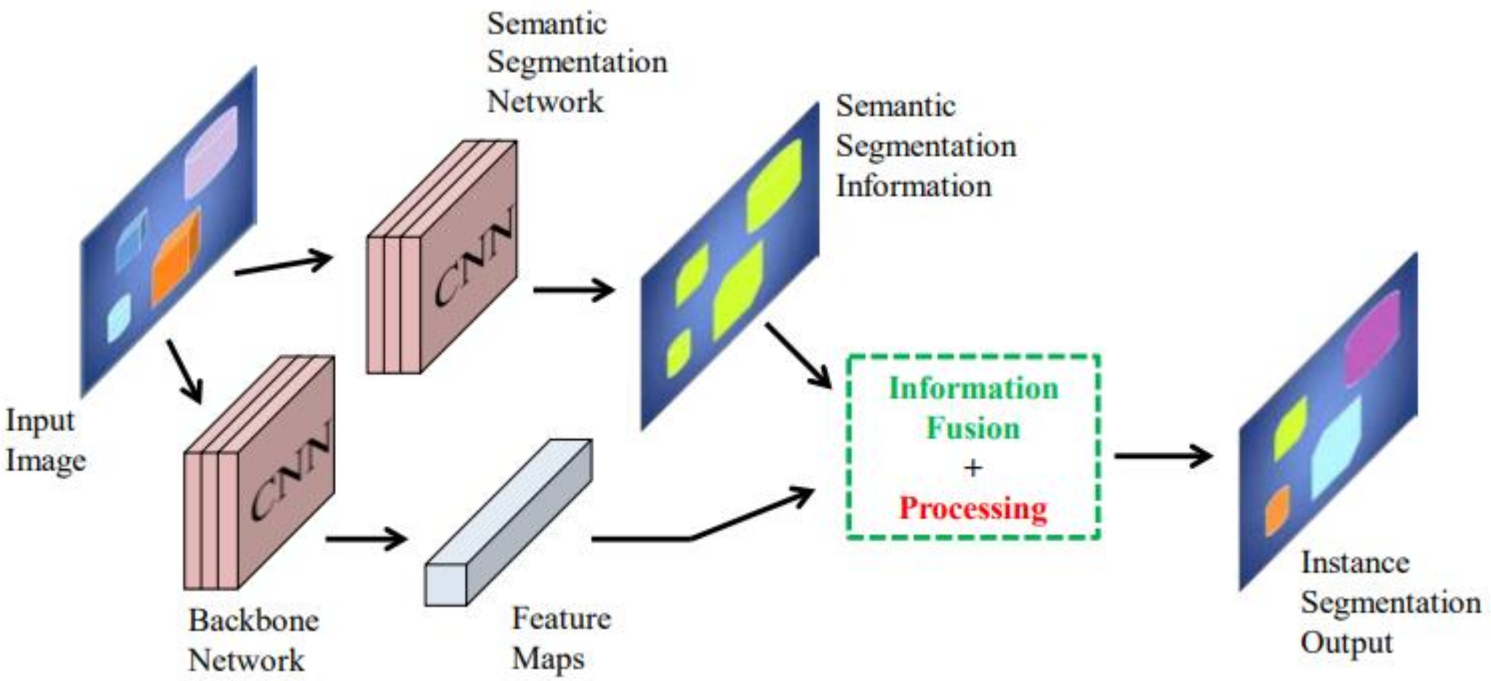
7.2 Detection followed by segmentation



Hafiz, Abdul Mueed, and Ghulam Mohiuddin Bhat. "A survey on instance segmentation: state of the art." *International Journal of Multimedia Information Retrieval* 9, no. 3 (2020): 171-189.

7. Instance Segmentation

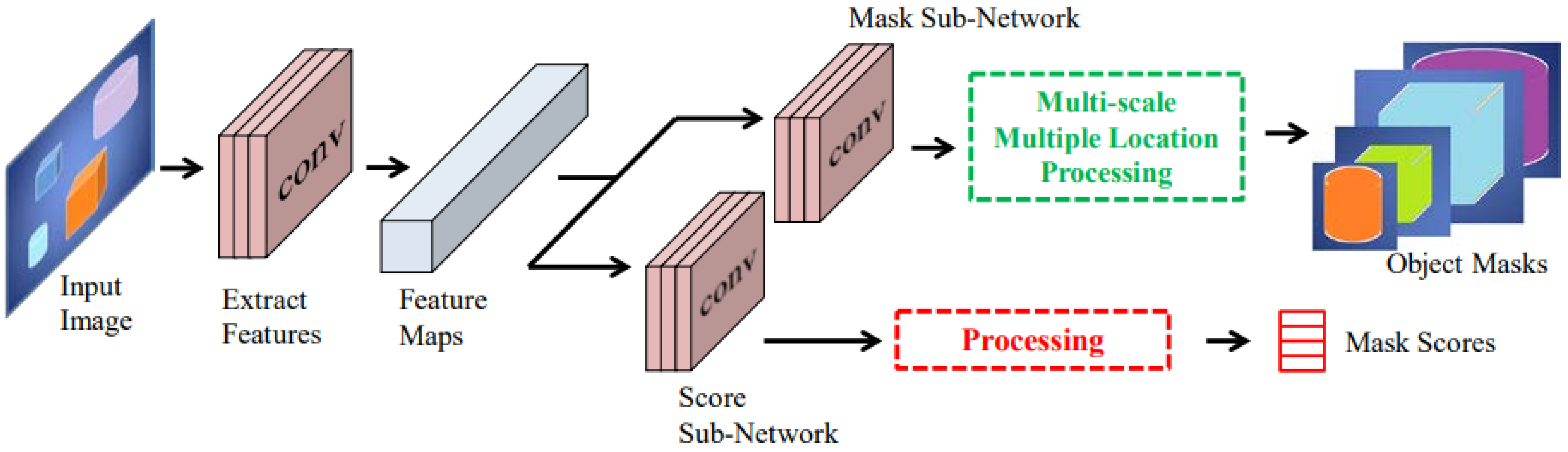
7.3 Labelling pixels followed by clustering



Hafiz, Abdul Mueed, and Ghulam Mohiuddin Bhat. "A survey on instance segmentation: state of the art." *International Journal of Multimedia Information Retrieval* 9, no. 3 (2020): 171-189.

7. Instance Segmentation

7.4 Dense sliding window methods



Hafiz, Abdul Mueed, and Ghulam Mohiuddin Bhat. "A survey on instance segmentation: state of the art." *International Journal of Multimedia Information Retrieval* 9, no. 3 (2020): 171-189.

7. Instance Segmentation

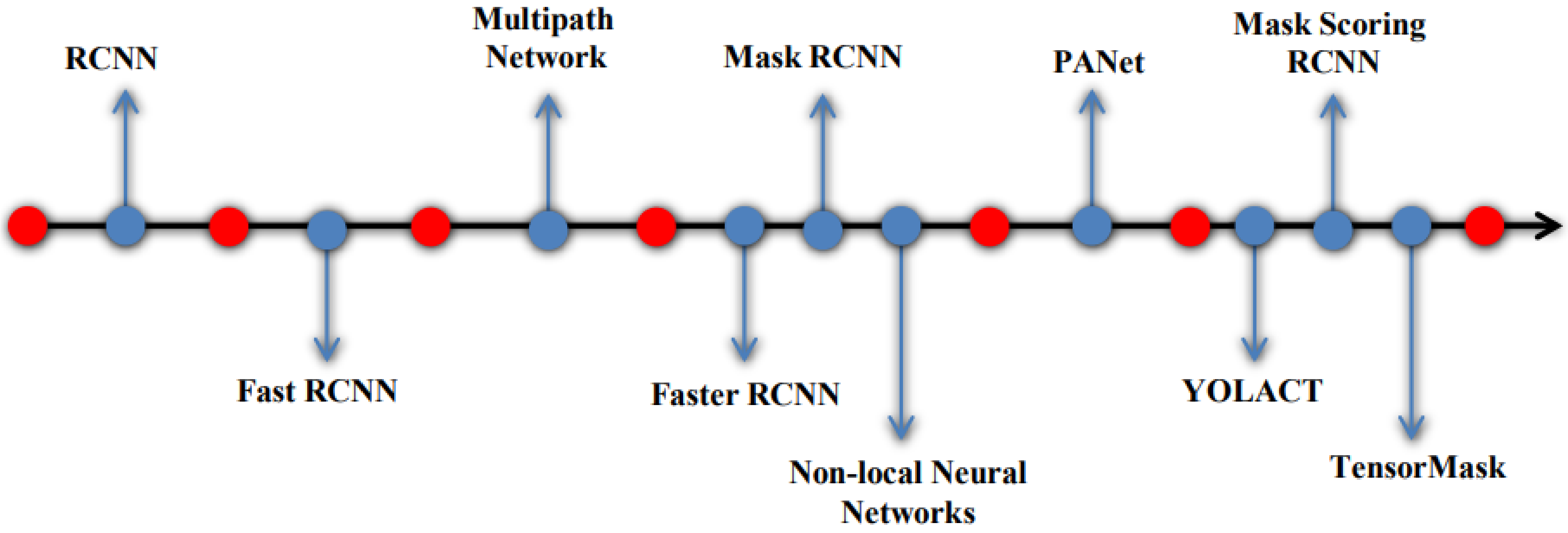


Figure 6. Timeline for notable techniques in instance segmentation

Hafiz, Abdul Mueed, and Ghulam Mohiuddin Bhat. "A survey on instance segmentation: state of the art." *International Journal of Multimedia Information Retrieval* 9, no. 3 (2020): 171-189.

7. Instance Segmentation: PANet

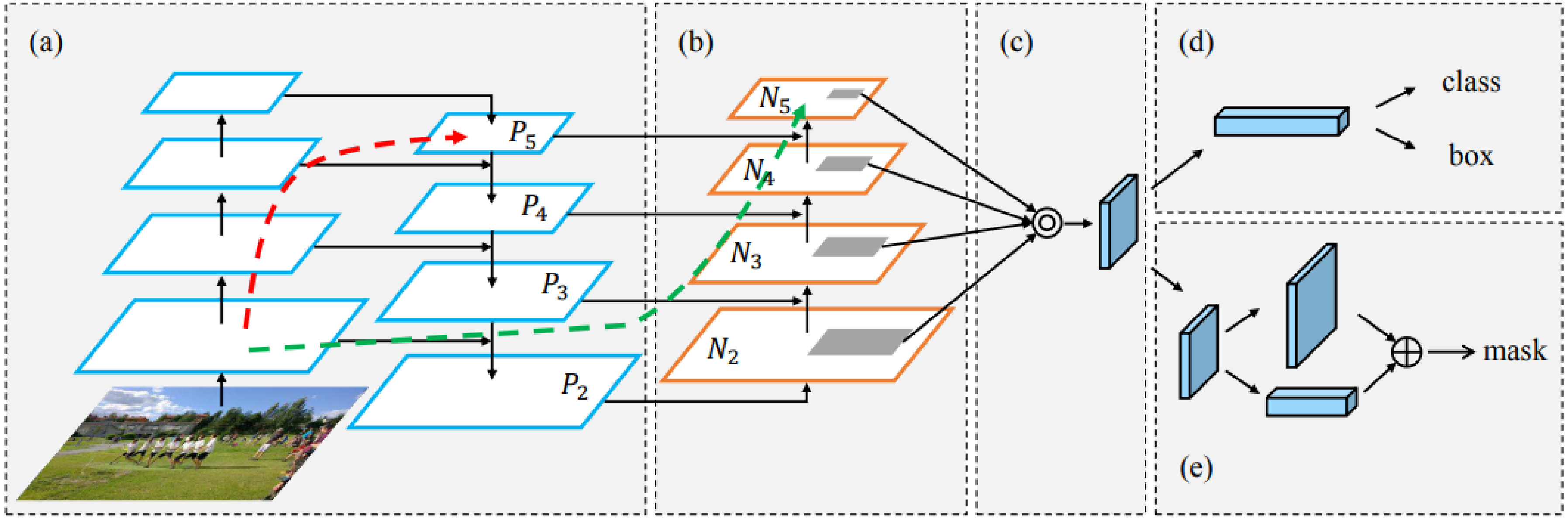


Figure 1. Illustration of our framework. (a) FPN backbone. (b) Bottom-up path augmentation. (c) Adaptive feature pooling. (d) Box branch. (e) Fully-connected fusion. Note that we omit channel dimension of feature maps in (a) and (b) for brevity.

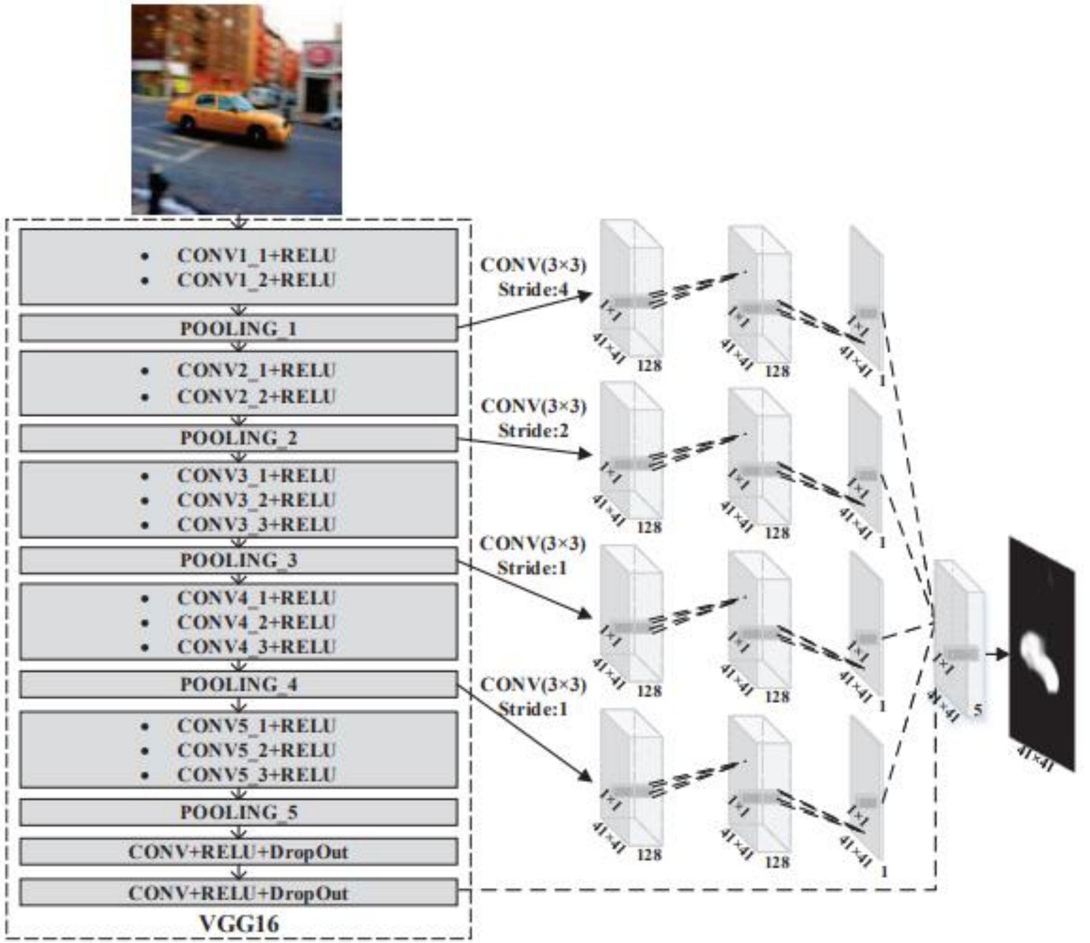
(c) Adaptive feature pooling: pooling features from all levels for each proposal and fusing them

Liu, Shu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. "Path aggregation network for instance segmentation." *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8759-8768. 2018.

8. Saliency

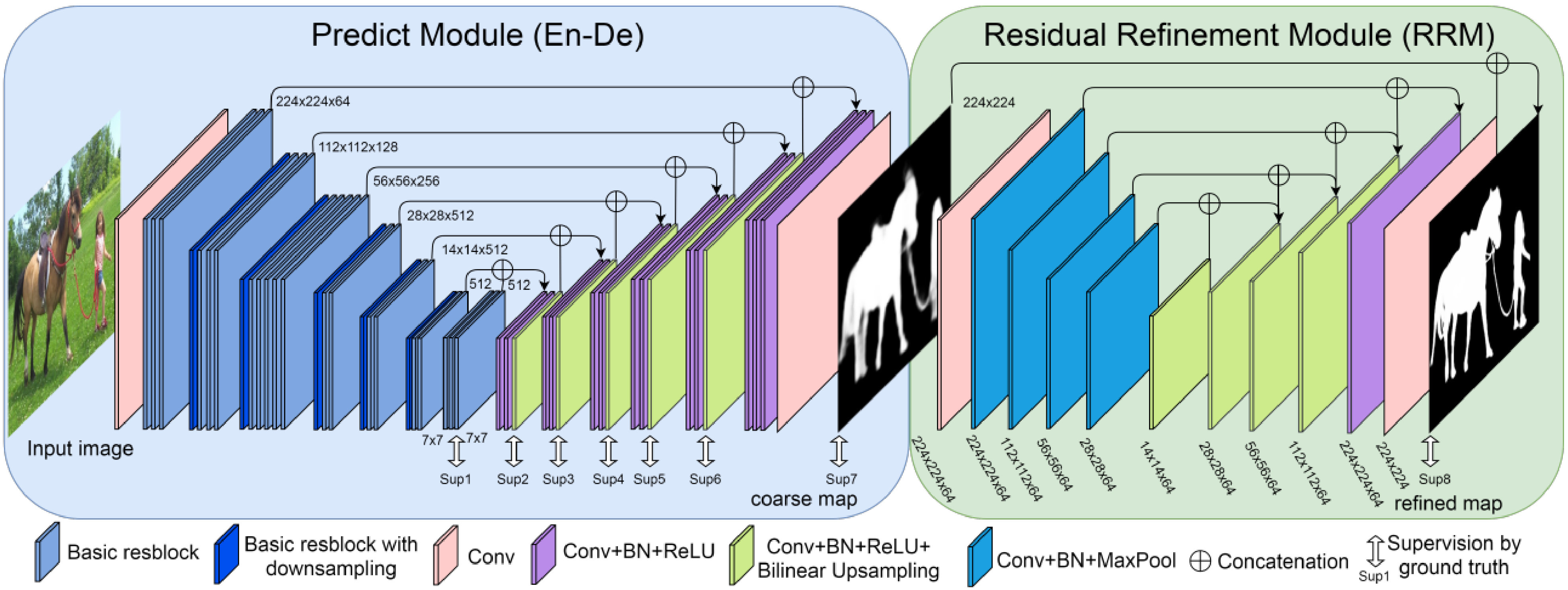
8. Saliency: DCL

- The stacked feature maps (5 channels) are fed into a final convolutional layer with a 1 × 1 kernel
- A single output channel: the inferred saliency map
- The *sigmoid activation function* is used in the final layer



Li, Guanbin, and Yizhou Yu. "Deep contrast learning for salient object detection." *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 478-487. 2016.

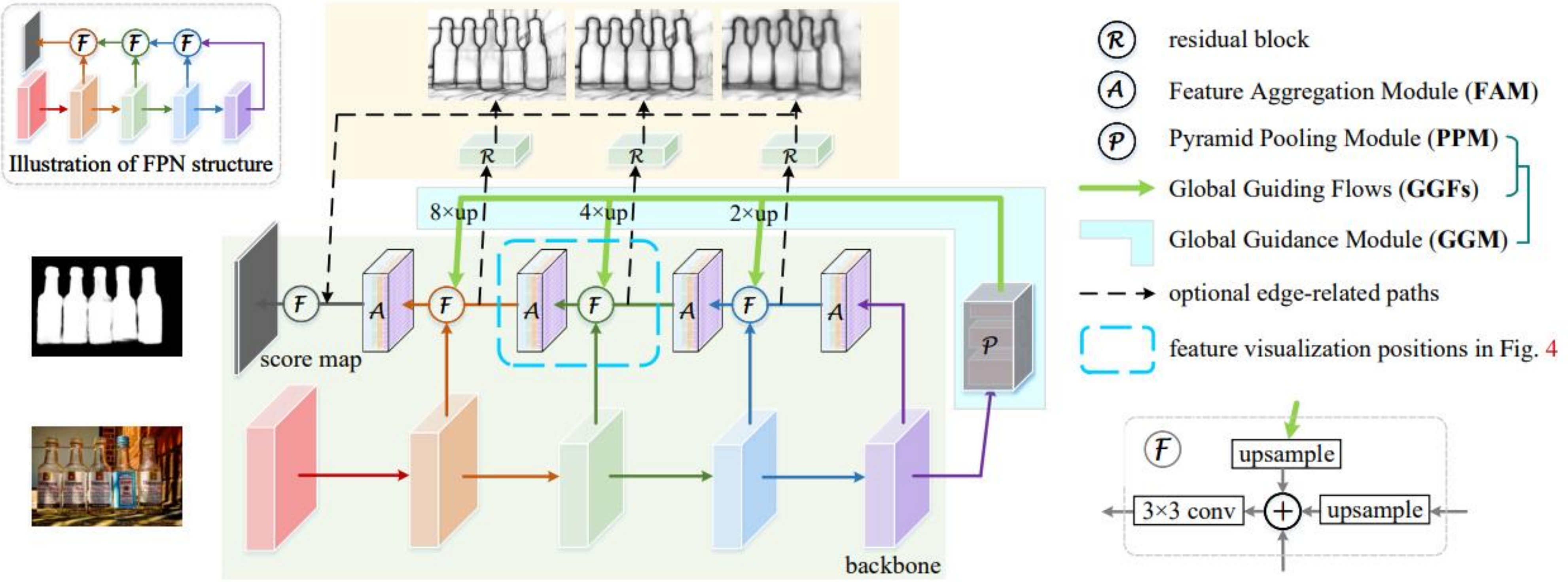
8. Saliency: BASNet



Qin, Xuebin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. "Basnet: Boundary-aware salient object detection." *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7479-7489. 2019.

8. Saliency: PoolNet

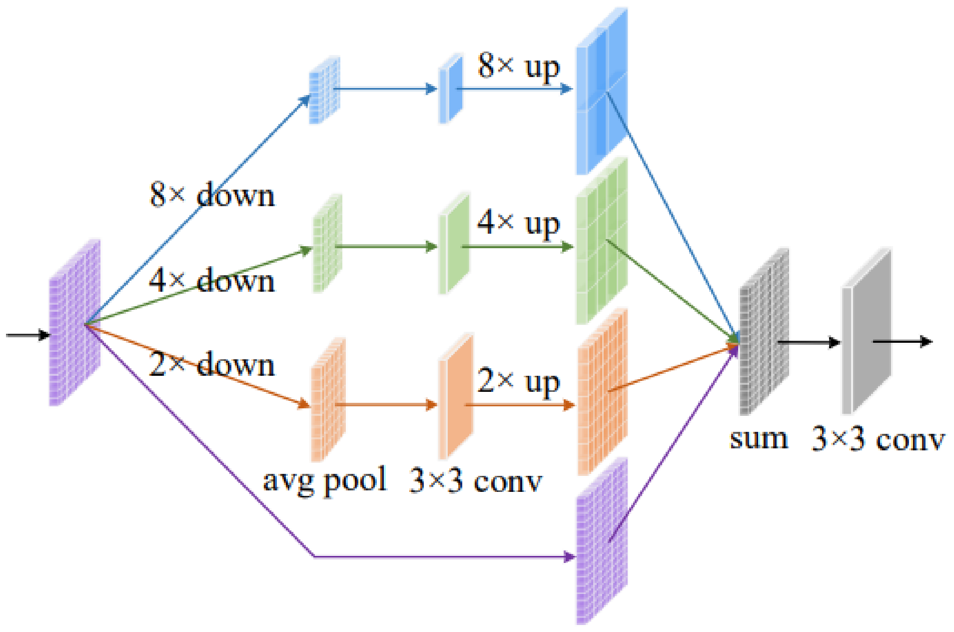
Coarse-level features from the GGM



30 FPS when processing a 300×400 image

Liu, Jiang-Jiang, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. "A simple pooling-based design for real-time salient object detection." *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3917-3926. 2019.

8. Saliency: PoolNet

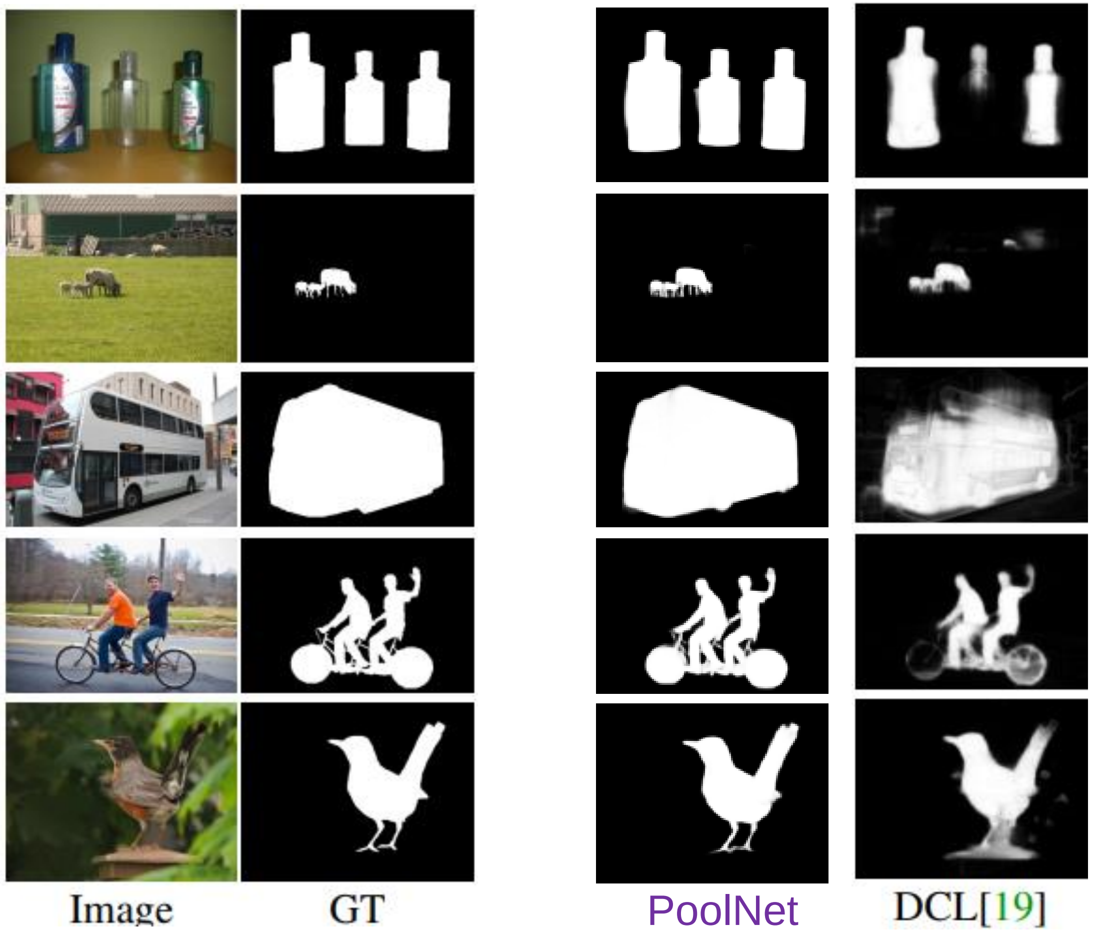


Feature Aggregation Module (FAM)

- FAM is to make the coarse-level semantic information well fused with the fine-level features from the top-down pathway.
- By adding FAMs after the fusion operations in the top-down pathway, coarse-level features from the GGM can be seamlessly merged with features at various scales.

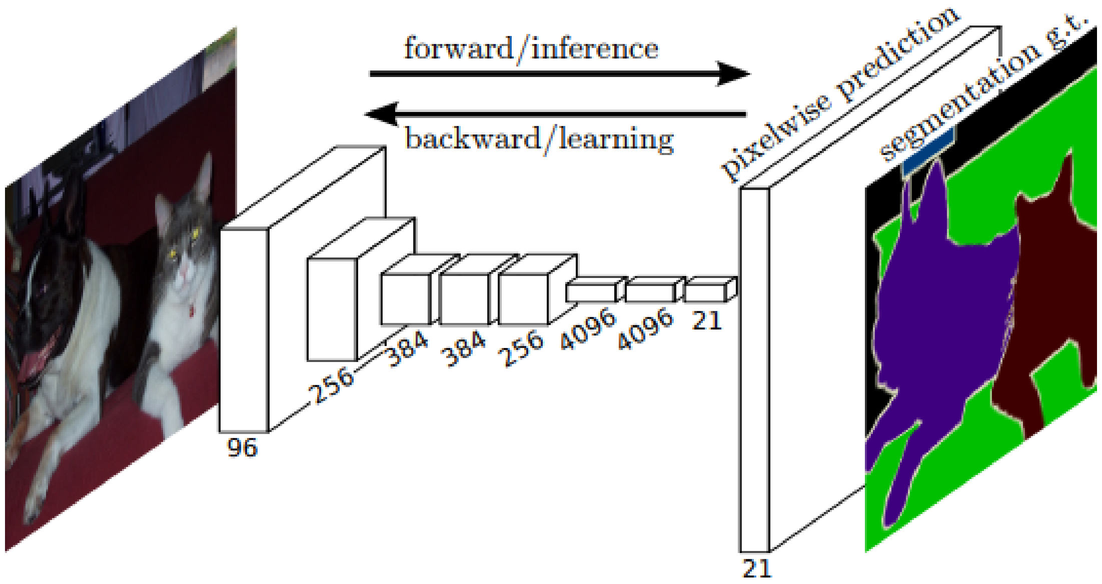
Liu, Jiang-Jiang, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. "A simple pooling-based design for real-time salient object detection." *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3917-3926. 2019.

8. Saliency: PoolNet



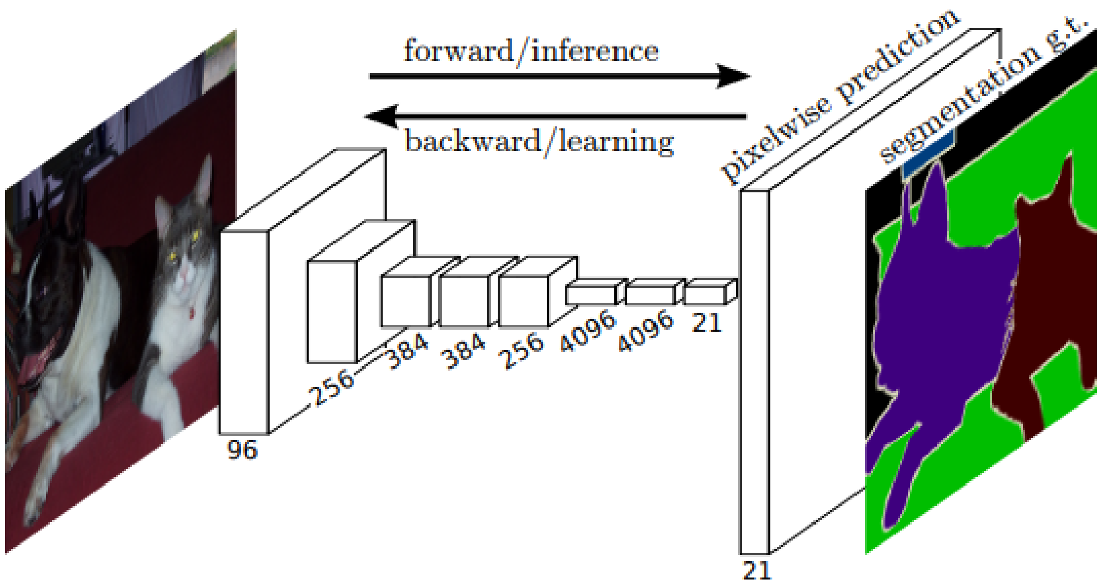
Liu, Jiang-Jiang, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. "A simple pooling-based design for real-time salient object detection." *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3917-3926. 2019.

9. Loss for semantic segmentation



Brainstorm: How to compute the loss for semantic segmentation?

9. Loss for semantic segmentation



Each pixel output: 21 channels for 21 classes

Softmax for each pixel

Cross-entropy loss for each pixel

SUM/Avg over all pixels

9. Loss for binary segmentation and saliency

Binary Cross-Entropy (BCE) loss:

$$\ell_{bce} = -\sum_{(r,c)} [G(r,c)\log(S(r,c)) + (1-G(r,c))\log(1-S(r,c))]$$

Structural Similarity (SSIM) loss:

$$\ell_{ssim} = 1 - \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

(a) Ori_image (b) Ground_truth

Intersection over Union (IoU) loss:

$$\ell_{iou} = 1 - \frac{\sum_{r=1}^{H}\sum_{c=1}^{W} S(r,c)G(r,c)}{\sum_{r=1}^{H}\sum_{c=1}^{W} [S(r,c)+G(r,c)-S(r,c)G(r,c)]}$$