DeepLab

김 성 철

목차

1. DeepLab v1

Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs

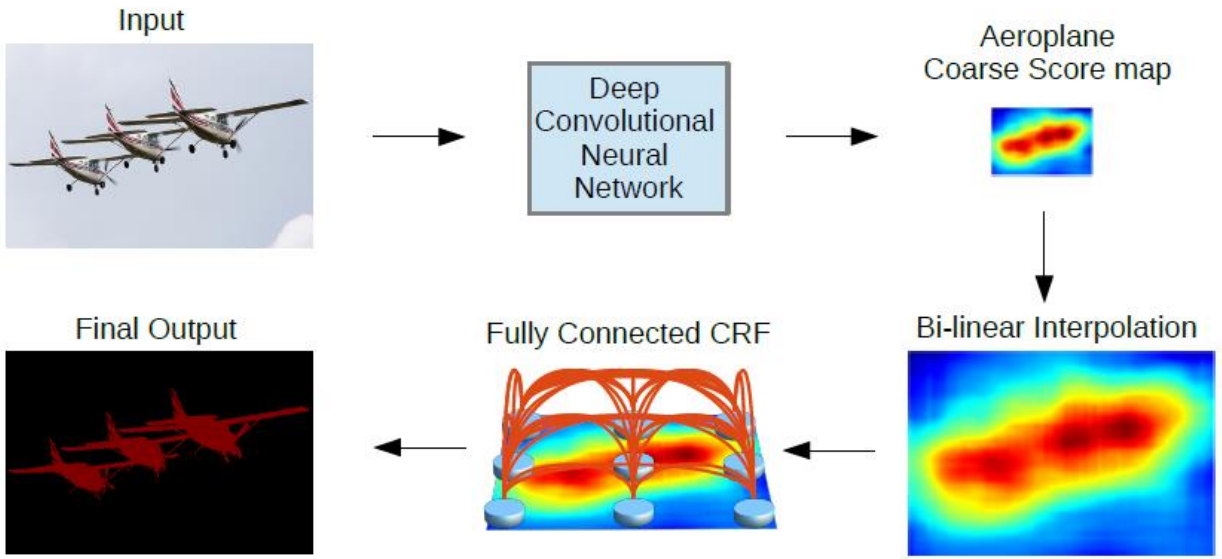
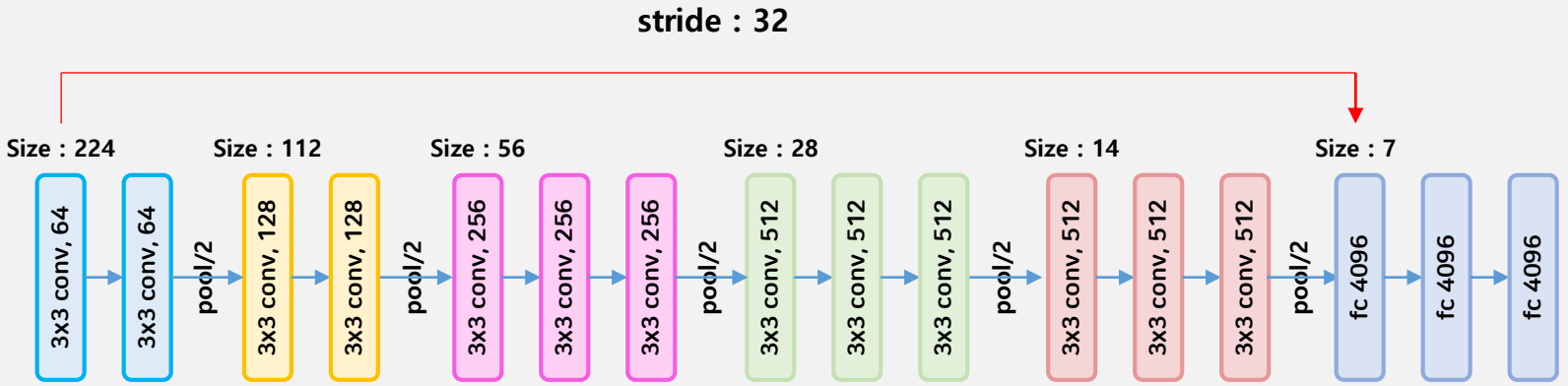


Figure 3: Model Illustration. The coarse score map from Deep Convolutional Neural Network (with fully convolutional layers) is upsampled by bi-linear interpolation. A fully connected CRF is applied to refine the segmentation result. Best viewed in color.

1. DeepLab v1


stride : 32
Size : 224
Size : 112
Size : 56
Size : 28
Size : 14
Size : 7
3x3 conv, 64
3x3 conv, 64
pool/2
3x3 conv, 128
3x3 conv, 128
pool/2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
pool/2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool/2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool/2
fc 4096
fc 4096
fc 4096

1. DeepLab v1

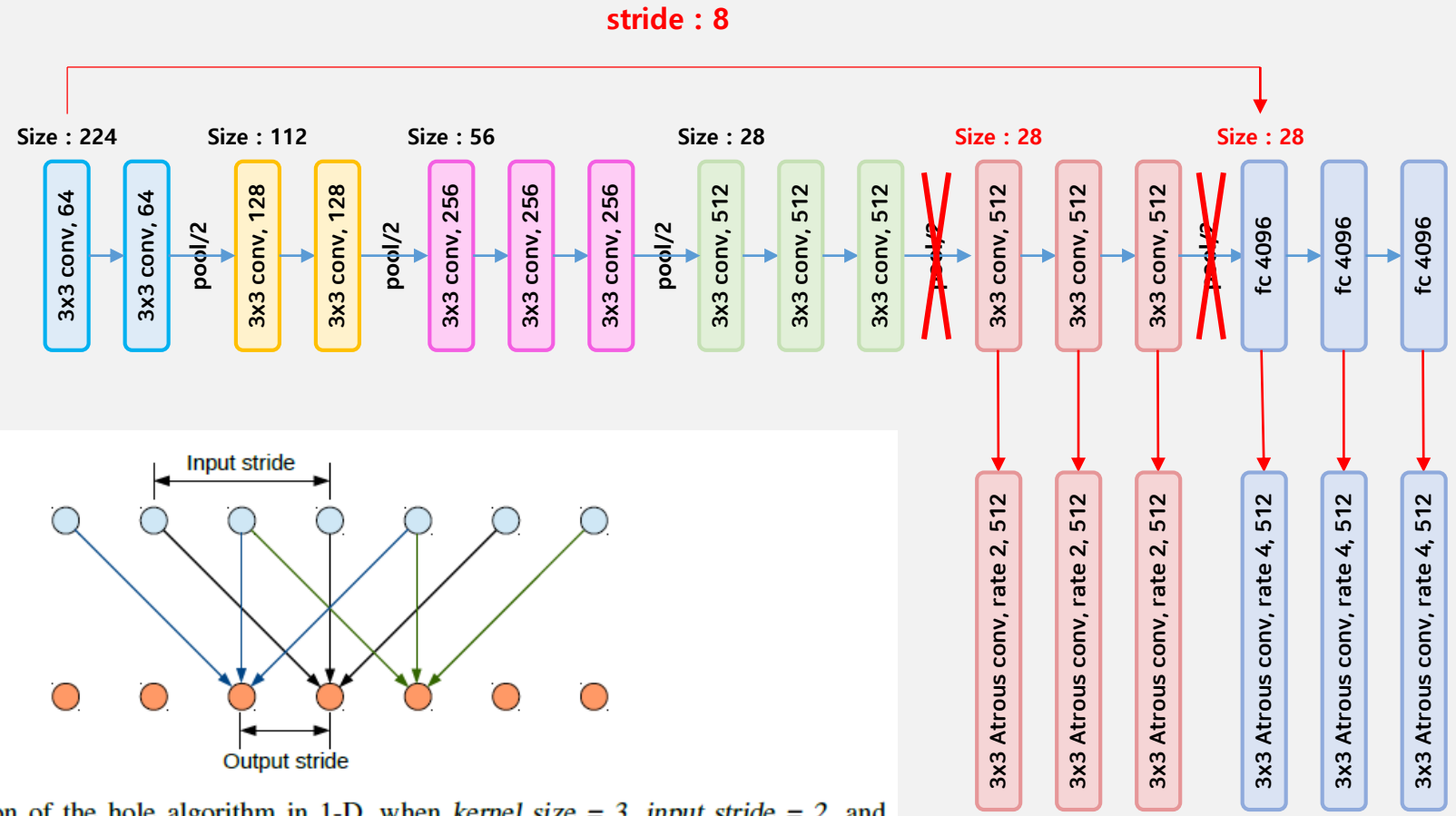
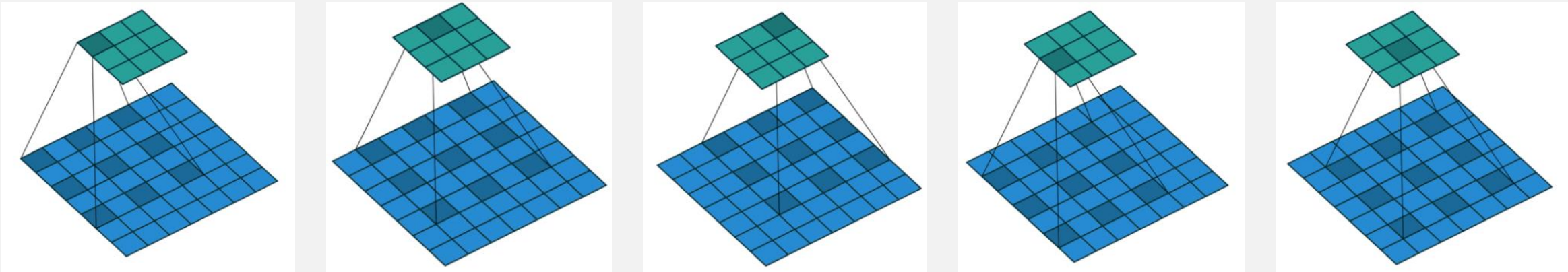


Figure 1: Illustration of the hole algorithm in 1-D, when *kernel_size* = 3, *input_stride* = 2, and *output_stride* = 1.

1. DeepLab v1



1. DeepLab v1

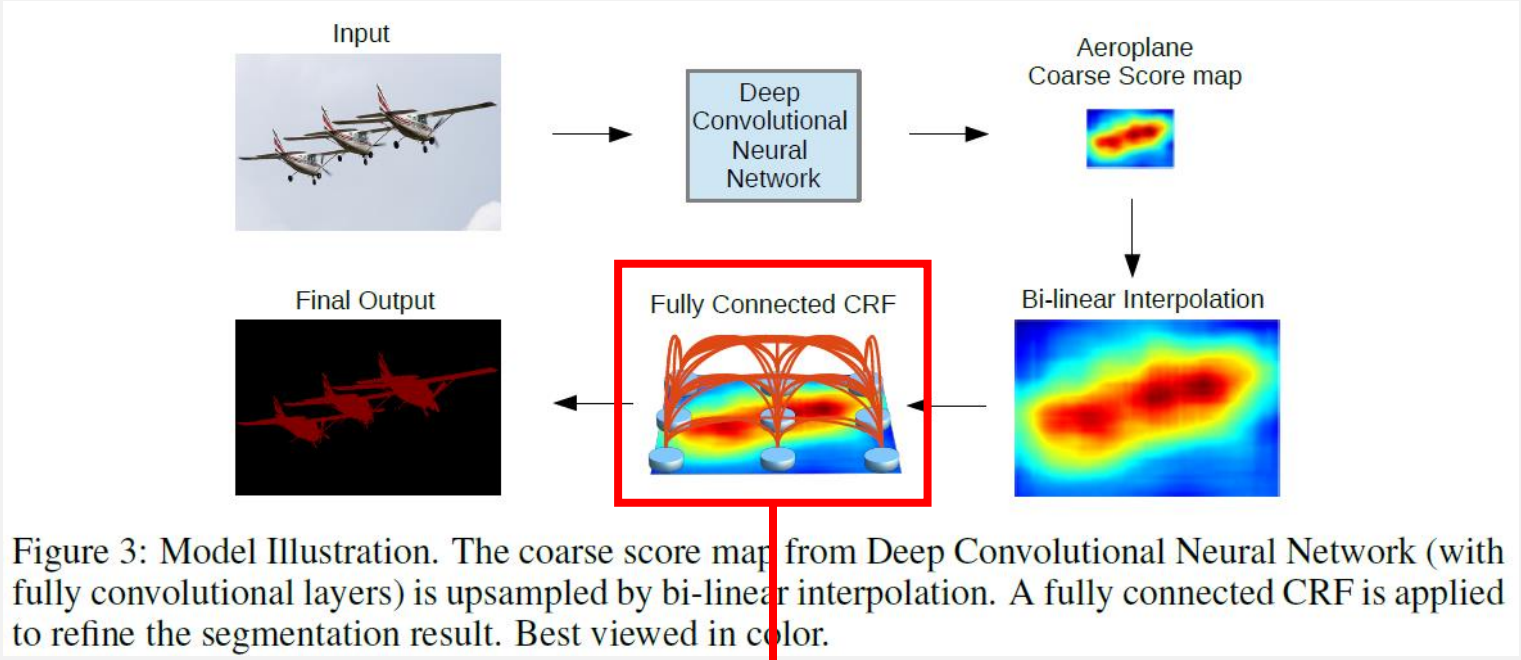
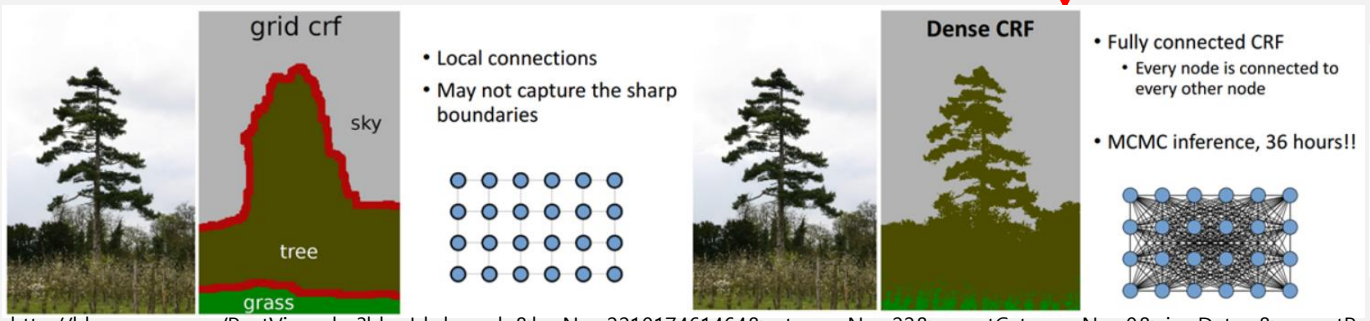


Figure 3: Model Illustration. The coarse score map from Deep Convolutional Neural Network (with fully convolutional layers) is upsampled by bi-linear interpolation. A fully connected CRF is applied to refine the segmentation result. Best viewed in color.



$$E(\boldsymbol{x}) = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j)$$

$$w_1 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\alpha^2} - \frac{||I_i - I_j||^2}{2\sigma_\beta^2}\right) + w_2 \exp\left(-\frac{||p_i - p_j||^2}{2\sigma_\gamma^2}\right)$$

http://blog.naver.com/PostView.nhn?blogId=laonple&logNo=221017461464&categoryNo=22&parentCategoryNo=0&viewDate=&currentPage=2&postListTopCurrentPage=&from=postList&userTopListOpen=true&userTopListCount=10&userTopListManageOpen=false&userTopListCurrentPage=2

1. DeepLab v1

Method	mean IOU (%)
DeepLab	59.80
DeepLab-CRF	63.74
DeepLab-MSc	61.30
DeepLab-MSc-CRF	65.21
DeepLab-7x7	64.38
DeepLab-CRF-7x7	67.64
DeepLab-LargeFOV	62.25
DeepLab-CRF-LargeFOV	67.64
DeepLab-MSc-LargeFOV	64.21
DeepLab-MSc-CRF-LargeFOV	68.70

(a)

Method	mean IOU (%)
MSRA-CFM	61.8
FCN-8s	62.2
TTI-Zoomout-16	64.4
DeepLab-CRF	66.4
DeepLab-MSc-CRF	67.1
DeepLab-CRF-7x7	70.3
DeepLab-CRF-LargeFOV	70.3
DeepLab-MSc-CRF-LargeFOV	71.6

(b)

Table 1: (a) Performance of our proposed models on the PASCAL VOC 2012 'val' set (with training in the augmented 'train' set). The best performance is achieved by exploiting both multi-scale features and large field-of-view. (b) Performance of our proposed models (with training in the augmented 'trainval' set) compared to other state-of-art methods on the PASCAL VOC 2012 'test' set.

2. DeepLab v2

DeepLab v1 vs. DeepLab v2 (DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs)

➢ **공통점**

- Atrous Convolution
- Fully Connected CRF

➢ **차이점**

- Multiple Scale 처리방법 (ASPP)
- Back bone network (VGG-16 / ResNet-101)

2. DeepLab v2

Multiple Scale → Atrous Spatial Pyramid Pooling (ASPP)

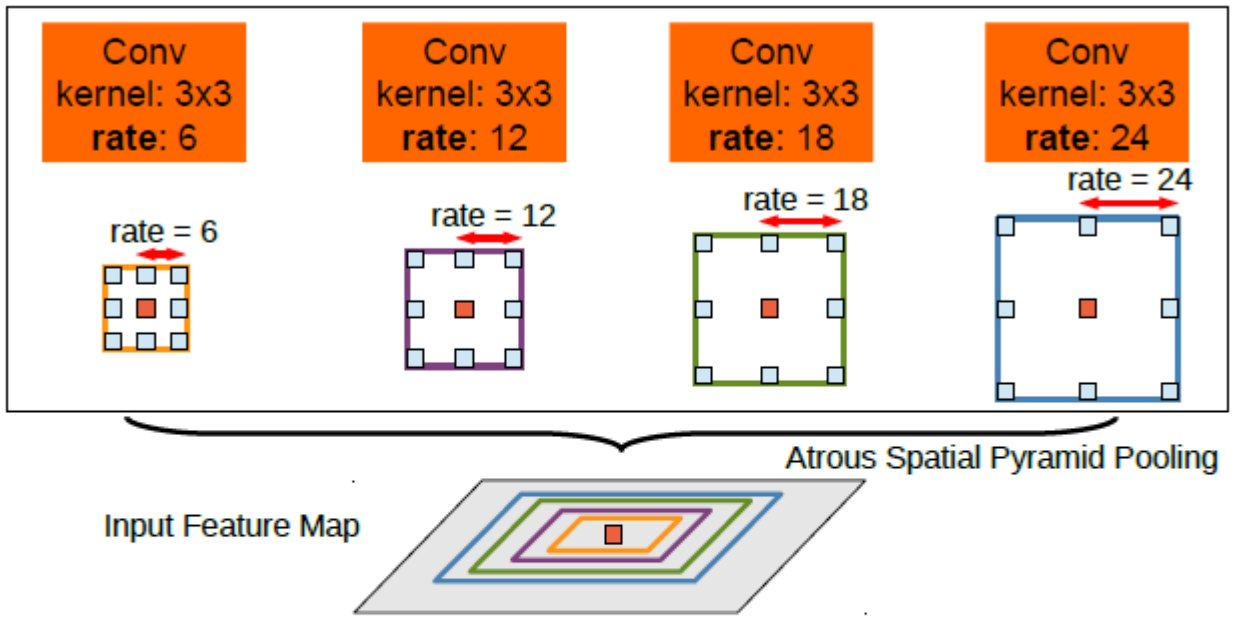
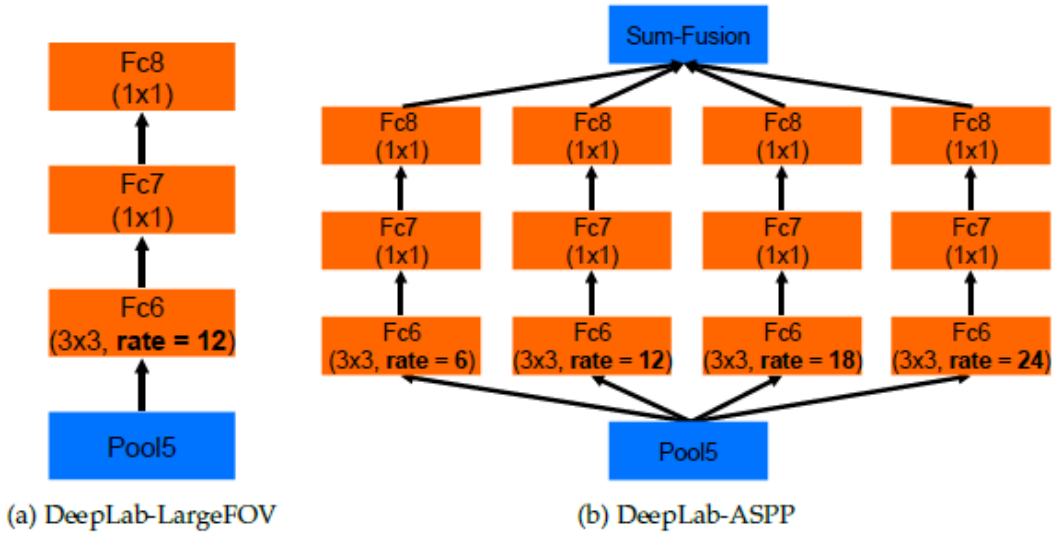


Fig. 4: Atrous Spatial Pyramid Pooling (ASPP). To classify the center pixel (orange), ASPP exploits multi-scale features by employing multiple parallel filters with different rates. The effective Field-Of-Views are shown in different colors.

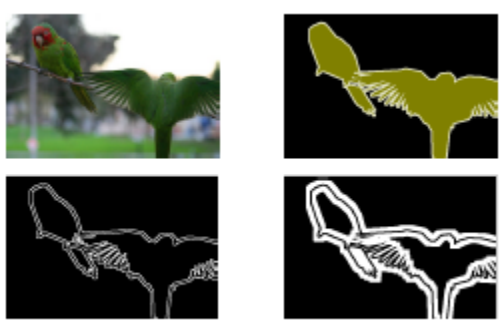


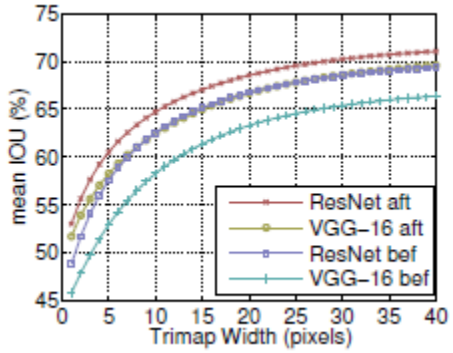
(a) DeepLab-LargeFOV (b) DeepLab-ASPP

Fig. 7: DeepLab-ASPP employs multiple filters with different rates to capture objects and context at multiple scales.

2. DeepLab v2

Back bone network → VGG-16 / ResNet-101





(a) (b)

Fig. 10: (a) Trimap examples (top-left: image. top-right: ground-truth. bottom-left: trimap of 2 pixels. bottom-right: trimap of 10 pixels). (b) Pixel mean IOU as a function of the band width around the object boundaries when employing VGG-16 or ResNet-101 before and after CRF.

2. DeepLab v2

Method	mIOU
DeepLab-CRF-LargeFOV-COCO [58]	72.7
MERL_DEEP_GCRF [88]	73.2
CRF-RNN [59]	74.7
POSTECH_DeconvNet_CRF_VOC [61]	74.8
BoxSup [60]	75.2
Context + CRF-RNN [76]	75.3
QO_4^{mres} [66]	75.5
DeepLab-CRF-Attention [17]	75.7
CentraleSuperBoundaries++ [18]	76.0
DeepLab-CRF-Attention-DT [63]	76.3
H-ReNet + DenseCRF [89]	76.8
LRR_4x_COCO [90]	76.8
DPN [62]	77.5
Adelaide_Context [40]	77.8
Oxford_TVG_HO_CRF [91]	77.9
Context CRF + Guidance CRF [92]	78.1
Adelaide_VeryDeep_FCN_VOC [93]	79.1
DeepLab-CRF (ResNet-101)	79.7

TABLE 5: Performance on PASCAL VOC 2012 *test* set. We have added some results from recent arXiv papers on top of the official leadearboard results.

3. DeepLab v3

DeepLab v2 vs. DeepLab v3 (Rethinking Atrous Convolution for Semantic Image Segmentation)

➢ **공통점**

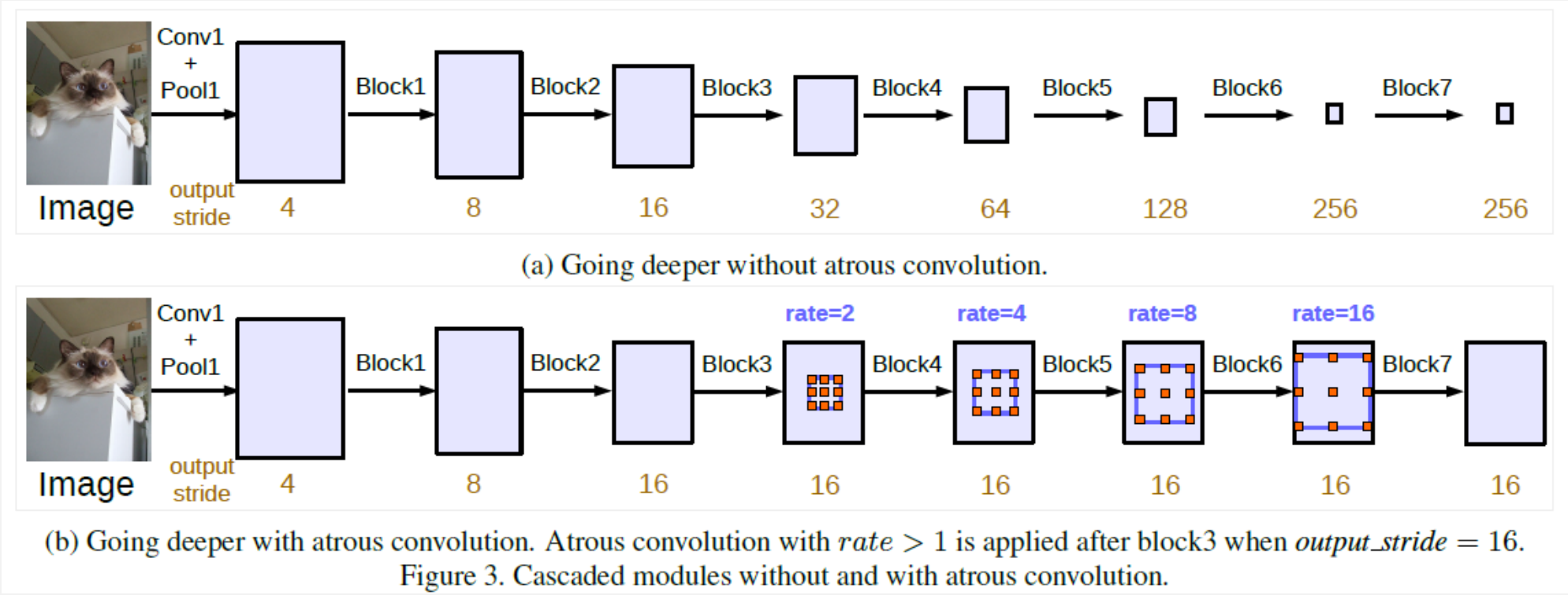
- Atrous Convolution
- Atrous Spatial Pyramid Pooling

➢ **차이점**

- Without DenseCRF
- Going Deeper with Atrous Convolution
- Multi-grid Method
- Bootstrapping Method

3. DeepLab v3

Going Deeper with Atrous Convolution



점점 깊어질 수록 세부 정보가 사라짐
→ Cascade ResNet blocks with Atrous Convolution

3. DeepLab v3

Multi-grid Method

- Cascade ResNet blocks 에서 Atrous Convolution을 어떻게 적용할 것인가?

  → Grid를 설정해서 적용

Multi-Grid	block4	block5	block6	block7
(1, 1, 1)	68.39	73.21	75.34	75.76
(1, 2, 1)	70.23	75.67	76.09	**76.66**
(1, 2, 3)	73.14	75.78	75.96	76.11
(1, 2, 4)	73.45	75.74	75.85	76.02
(2, 2, 2)	71.45	74.30	74.70	74.62

Table 3. Employing multi-grid method for ResNet-101 with different number of cascaded blocks at *output_stride* = 16. The best model performance is shown in bold.

3. DeepLab v3

DeepLab v3 with ASPP

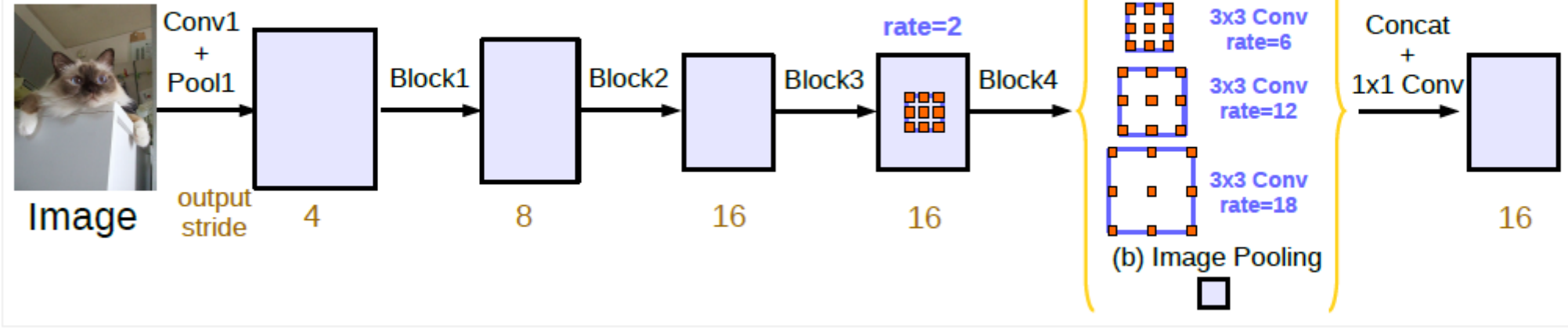
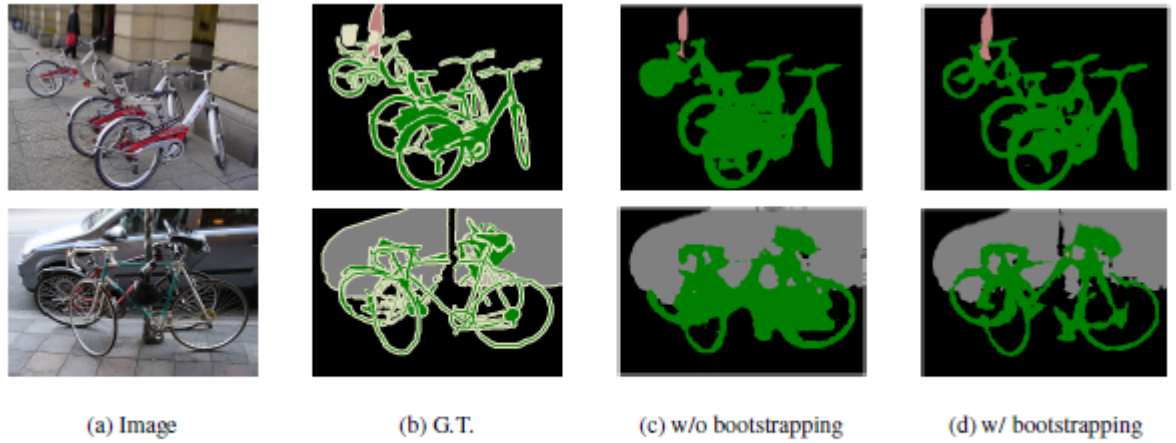


Figure 5. Parallel modules with atrous convolution (ASPP), augmented with image-level features.

3. DeepLab v3

Bootstrapping Method

Sec. 4.1 for details). Besides, instead of performing pixel hard example mining as [85, 70], we resort to bootstrapping on hard images. In particular, we duplicate the images that contain hard classes (namely bicycle, chair, table, potted-plant, and sofa) in the training set. As shown in Fig. 7, the simple bootstrapping method is effective for segmenting the bicycle class. In the end, our 'DeepLabv3' achieves the performance of 85.7% on the test set without any DenseCRF post-processing, as shown in Tab. 7.



(a) Image (b) G.T. (c) w/o bootstrapping (d) w/ bootstrapping

Figure 7. Bootstrapping on hard images improves segmentation accuracy for rare and finely annotated classes such as bicycle.

3. DeepLab v3

Method	mIOU
Adelaide_VeryDeep_FCN_VOC [85]	79.1
LRR_4x_ResNet-CRF [25]	79.3
DeepLabv2-CRF [11]	79.7
CentraleSupelec Deep G-CRF [8]	80.2
HikSeg_COCO [80]	81.4
SegModel [75]	81.8
Deep Layer Cascade (LC) [52]	82.7
TuSimple [84]	83.1
Large_Kernel_Matters [68]	83.6
Multipath-RefineNet [54]	84.2
ResNet-38_MS_COCO [86]	84.9
PSPNet [95]	85.4
IDW-CNN [83]	86.3
CASIA_IVA_SDN [23]	86.6
DIS [61]	86.8
DeepLabv3	85.7
DeepLabv3-JFT	86.9

Table 7. Performance on PASCAL VOC 2012 *test* set.

4. DeepLab v3+

DeepLab v3 vs. DeepLab v3+ (Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation)

➢ **공통점**

- Atrous Convolution
- Atrous Spatial Pyramid Pooling

➢ **차이점**

- Encoder-Decoder Architecture
- Backbone Network (ResNet 101 / Xception)

4. DeepLab v3+

Encoder-Decoder Architecture

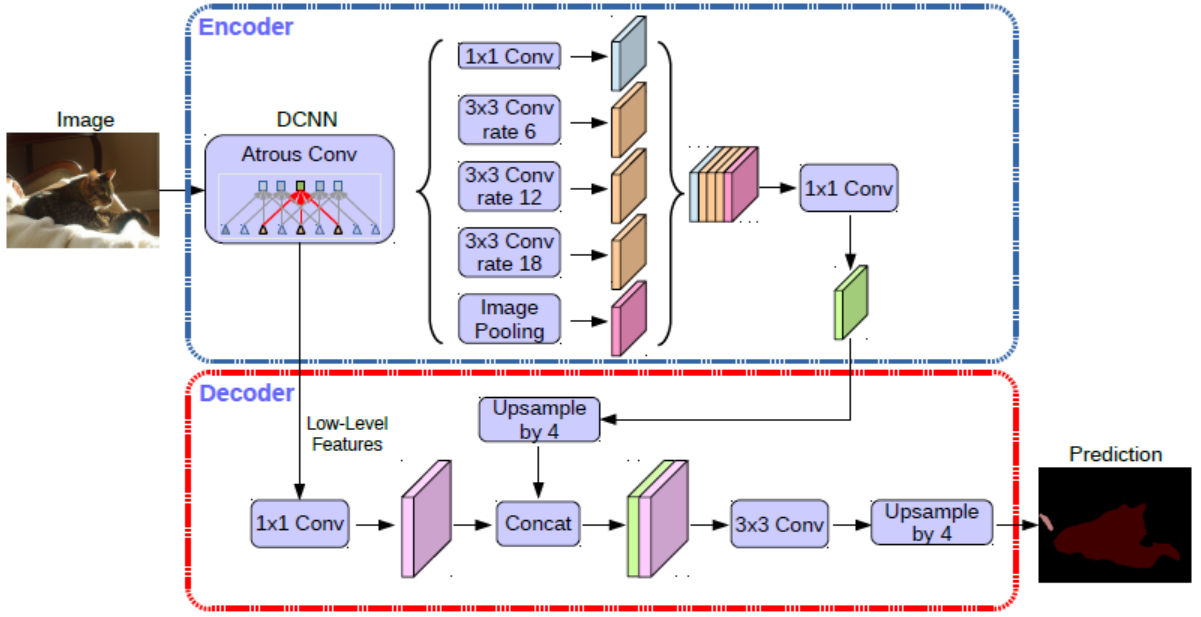


Fig. 2. Our proposed DeepLabv3+ extends DeepLabv3 by employing a encoder-decoder structure. The encoder module encodes multi-scale contextual information by applying atrous convolution at multiple scales, while the simple yet effective decoder module refines the segmentation results along object boundaries.

4. DeepLab v3+

Backbone Network (ResNet 101 / Xception)

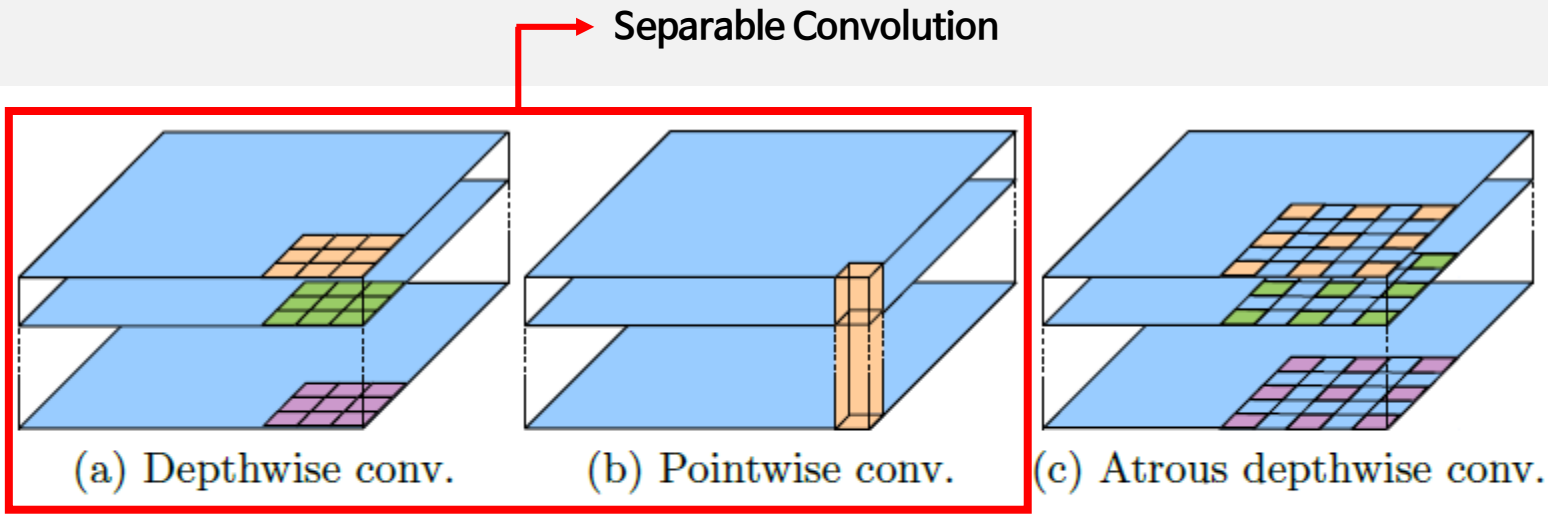


Fig. 3. 3×3 Depthwise separable convolution decomposes a standard convolution into (a) a depthwise convolution (applying a single filter for each input channel) and (b) a pointwise convolution (combining the outputs from depthwise convolution across channels). In this work, we explore *atrous separable convolution* where atrous convolution is adopted in the depthwise convolution, as shown in (c) with $rate = 2$.

4. DeepLab v3+

Backbone Network (ResNet 101 / Xception)

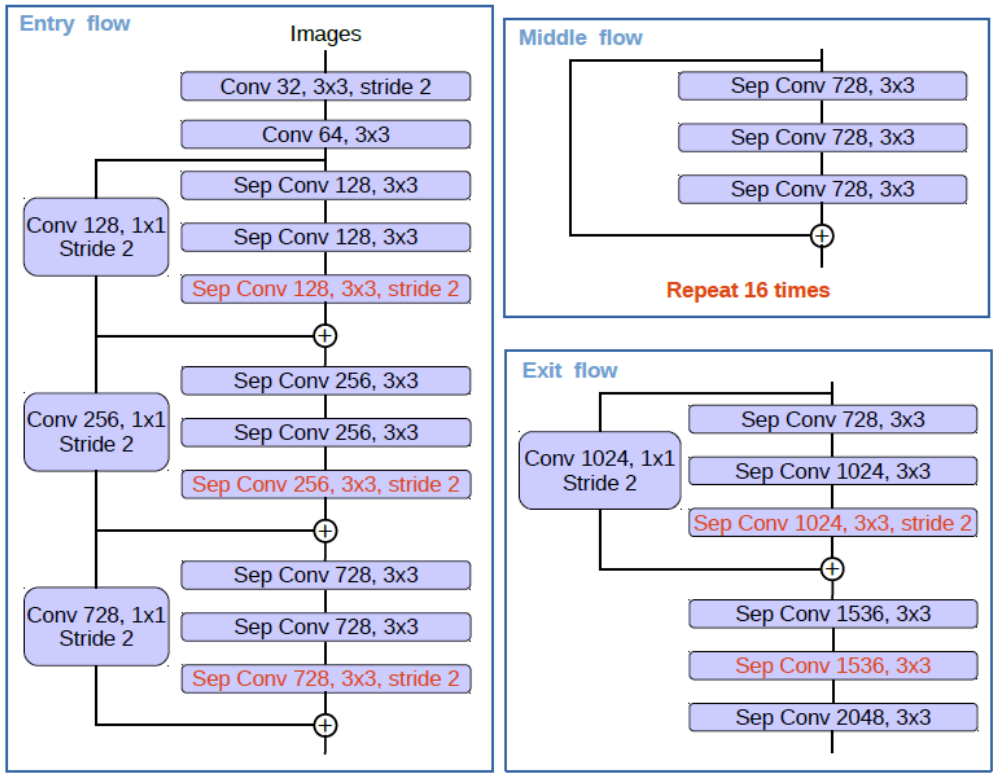
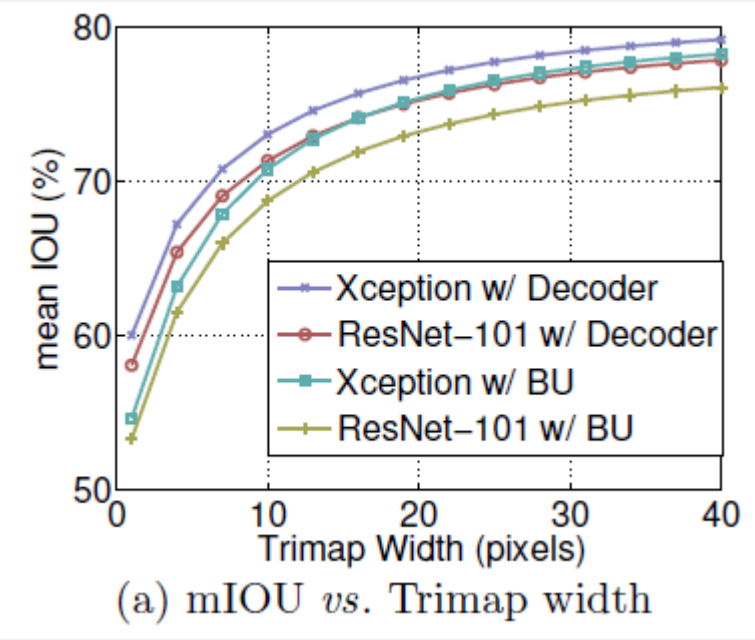


Fig. 4. We modify the Xception as follows: (1) more layers (same as MSRA's modification except the changes in Entry flow), (2) all the max pooling operations are replaced by depthwise separable convolutions with striding, and (3) extra batch normalization and ReLU are added after each 3×3 depthwise convolution, similar to MobileNet.



(a) mIOU *vs.* Trimap width

4. DeepLab v3+

Method	mIOU
Deep Layer Cascade (LC) [82]	82.7
TuSimple [77]	83.1
Large_Kernel_Matters [60]	83.6
Multipath-RefineNet [58]	84.2
ResNet-38_MS_COCO [83]	84.9
PSPNet [24]	85.4
IDW-CNN [84]	86.3
CASIA_IVA_SDN [63]	86.6
DIS [85]	86.8
DeepLabv3 [23]	85.7
DeepLabv3-JFT [23]	86.9
DeepLabv3+ (Xception)	87.8
DeepLabv3+ (Xception-JFT)	89.0

Table 6. PASCAL VOC 2012 *test* set results with top-performing models.

Reference

- DeepLab v1

  L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A L. Yuille, “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs”, in ICLR, 2015.

- DeepLab v2

  L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs”, arXiv:1606.00915, 2016.

- DeepLab v3

  L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation”, arXiv:1706.05587, 2017.

- DeepLab v3+

  L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation”, arXiv:1802.026113, 2018.

감사합니다