Semantic Segmentation

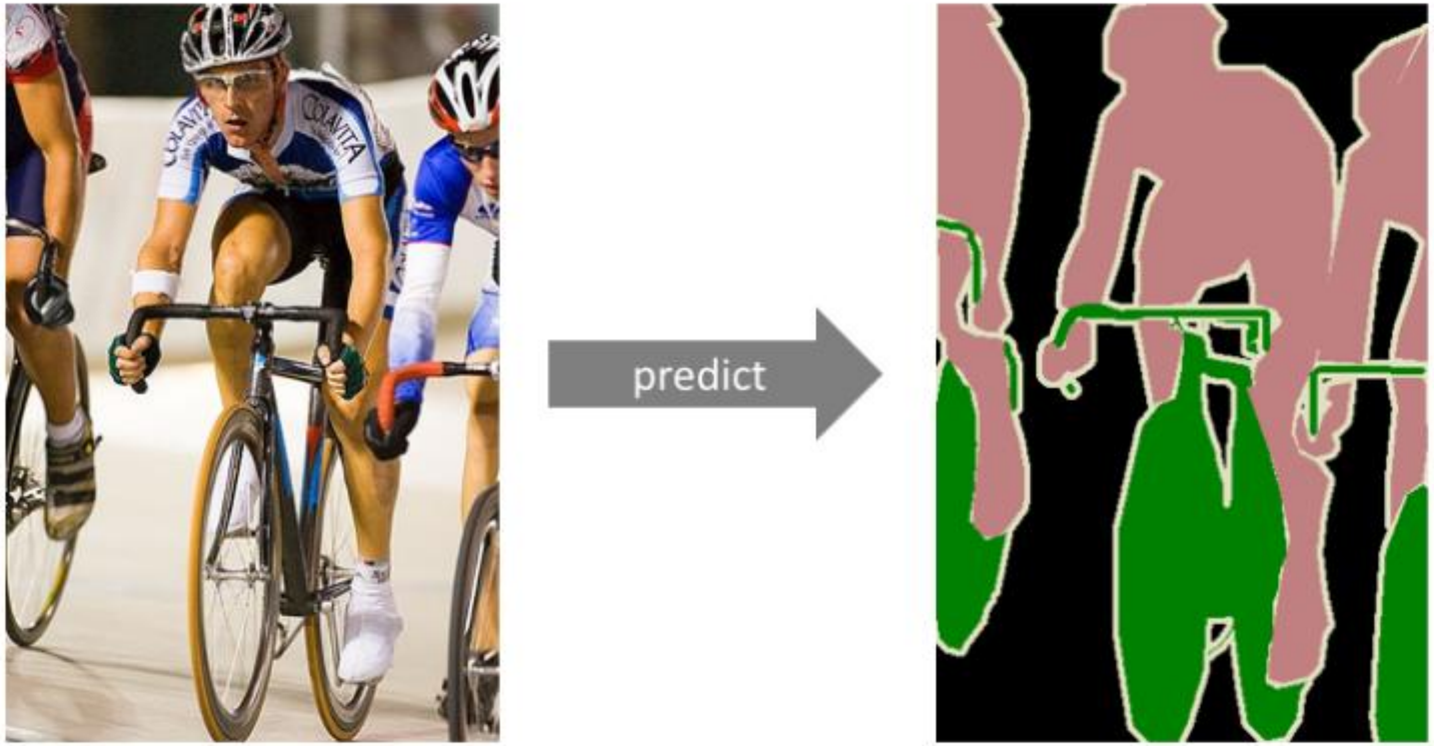
DeepLab Series Summary

2020. 05. 25. 월

김태우

CONTENTS

Semantic Segmentation은 Pixel level Classification 이다



Person
Bicycle
Background

주로 의료영상 분석, 자율주행 등 다양한 분야에 활용

그중에서 DeepLab 알고리즘은 Semantic Segmentation Task 에서 상위권에 많이 포진

TASK	DATASET	MODEL	METRIC NAME	METRIC VALUE	GLOBAL RANK	USES EXTRA TRAINING DATA	COMPARE
Lesion Segmentation	Anatomical Tracings of Lesions After Stroke (ATLAS)	DeepLab v3+	Dice	0.4609	# 5	✕	See all
Lesion Segmentation	Anatomical Tracings of Lesions After Stroke (ATLAS)	DeepLab v3+	IoU	0.3458	# 4	✕	See all
Lesion Segmentation	Anatomical Tracings of Lesions After Stroke (ATLAS)	DeepLab v3+	Precision	0.5831	# 5	✕	See all
Lesion Segmentation	Anatomical Tracings of Lesions After Stroke (ATLAS)	DeepLab v3+	Recall	0.4491	# 5	✕	See all
Semantic Segmentation	Cityscapes test	DeepLabv3+ (Xception-JFT)	Mean IoU (class)	82.1%	# 9	✓	See all
Semantic Segmentation	Cityscapes val	DeepLabv3+ (Dilated-Xception-71)	mIoU	79.6%	# 6	✕	See all
Image Classification	ImageNet	Modified Aligned Xception	Top 1 Accuracy	79.81%	# 45	✕	See all
Image Classification	ImageNet	Modified Aligned Xception	Top 5 Accuracy	94.83%	# 35	✕	See all
Semantic Segmentation	PASCAL VOC 2012 test	DeepLabv3+ (Xception-JFT)	Mean IoU	89.0%	# 1	✕	See all
Semantic Segmentation	PASCAL VOC 2012 val	DeepLabv3+ (Xception-JFT)	mIoU	84.56%	# 2	✓	See all

https://paperswithcode.com/paper/encoder-decoder-with-atrous-separable

DeepLab은 총 4번의 개정본이 출판

DeepLab V1

Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. ICLR 2015.

DeepLab V2

DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. TPAMI 2017.

DeepLab V3

Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv 2017.

DeepLab V3+

Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. arXiv 2018.

각 버전에서의 핵심 알고리즘

DeepLab V1

Atrous Convolution

DeepLab V2

Multi-scale context 적용을 위한 Atrous Spatial Pyramid Pooling (ASPP) + (CRF)

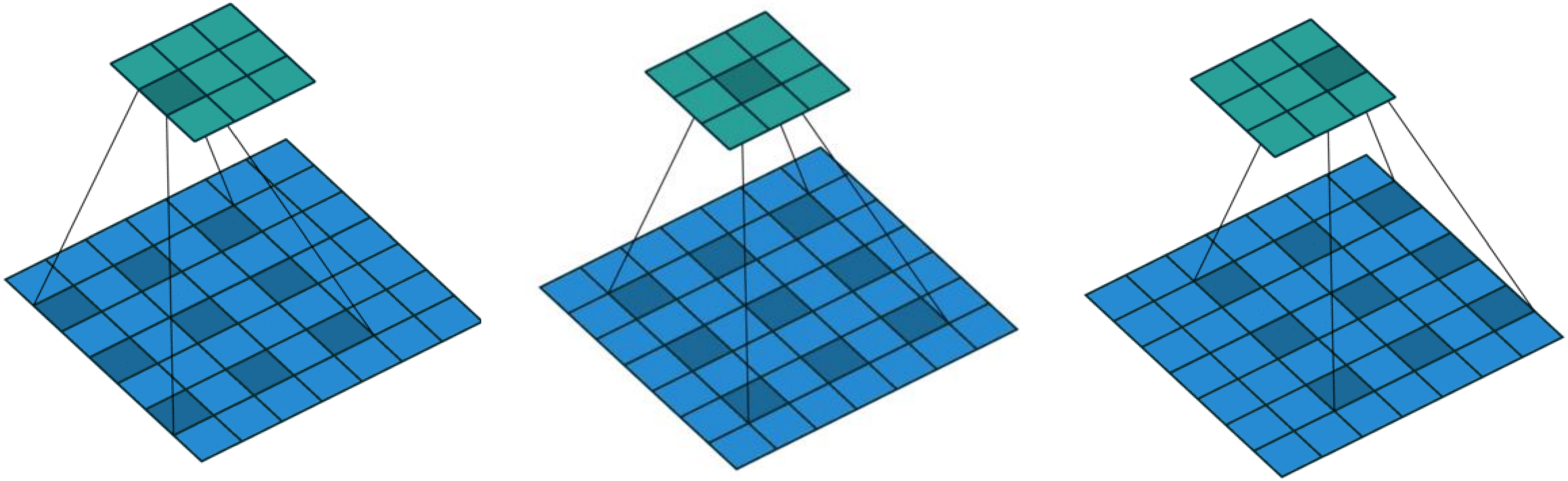
DeepLab V3

ResNet 구조에 Atrous convolution을 활용해 좀 더 dense 한 feature map 얻기

DeepLab V3+

Separable Convolution 과 Atrous convolution 을 결합한 Atrous separable Convolution 의 활용 제안

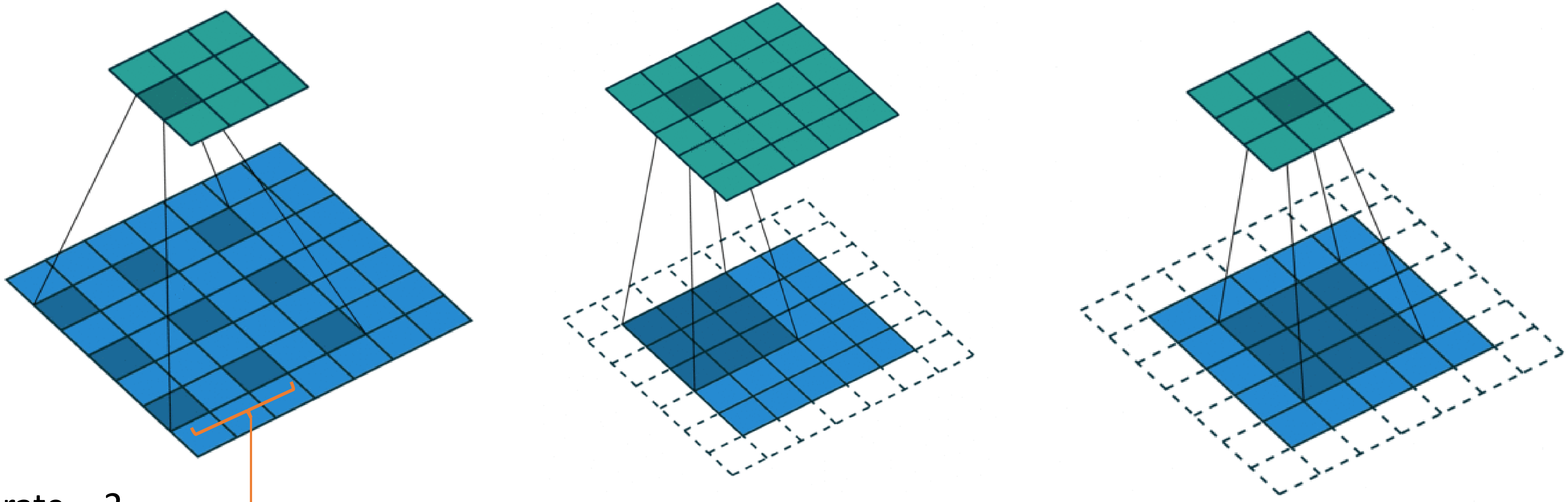
Atrous(Dilated) Convolution



파란색이 input , 초록색이 output

출처: https://3months.tistory.com/213 [Deep Play]

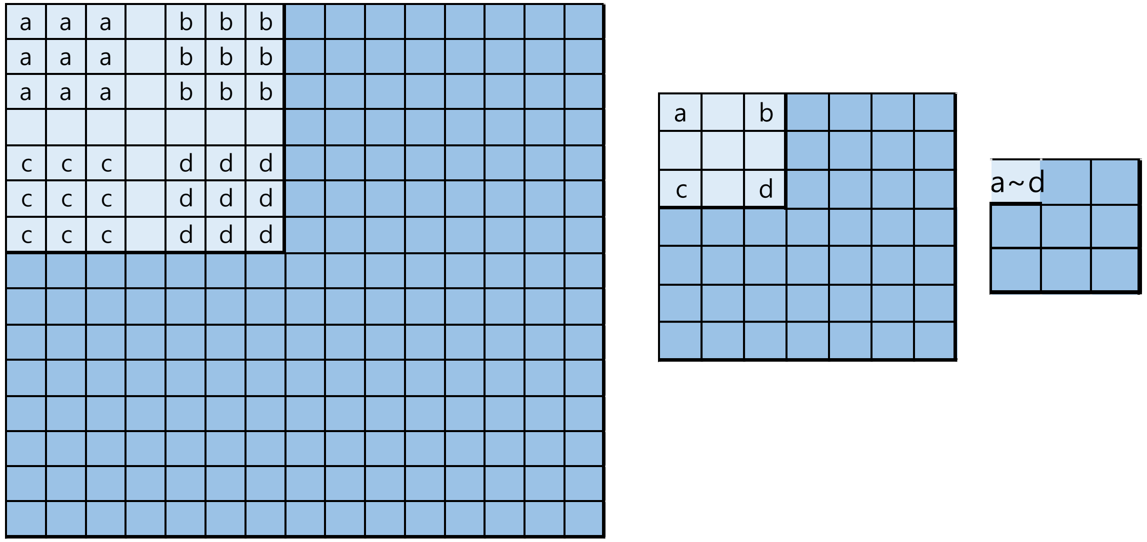
Atrous(Dilated) Convolution

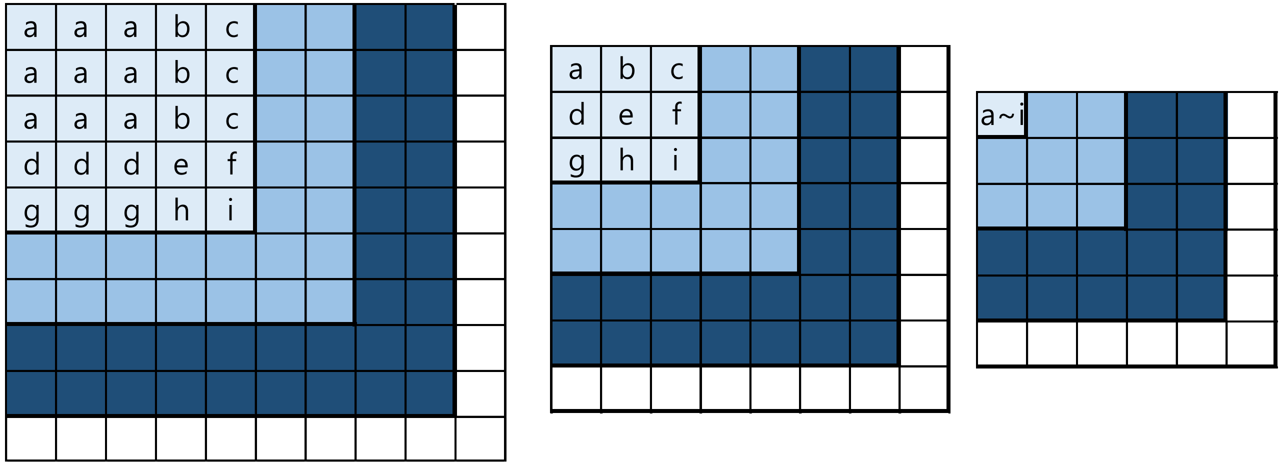


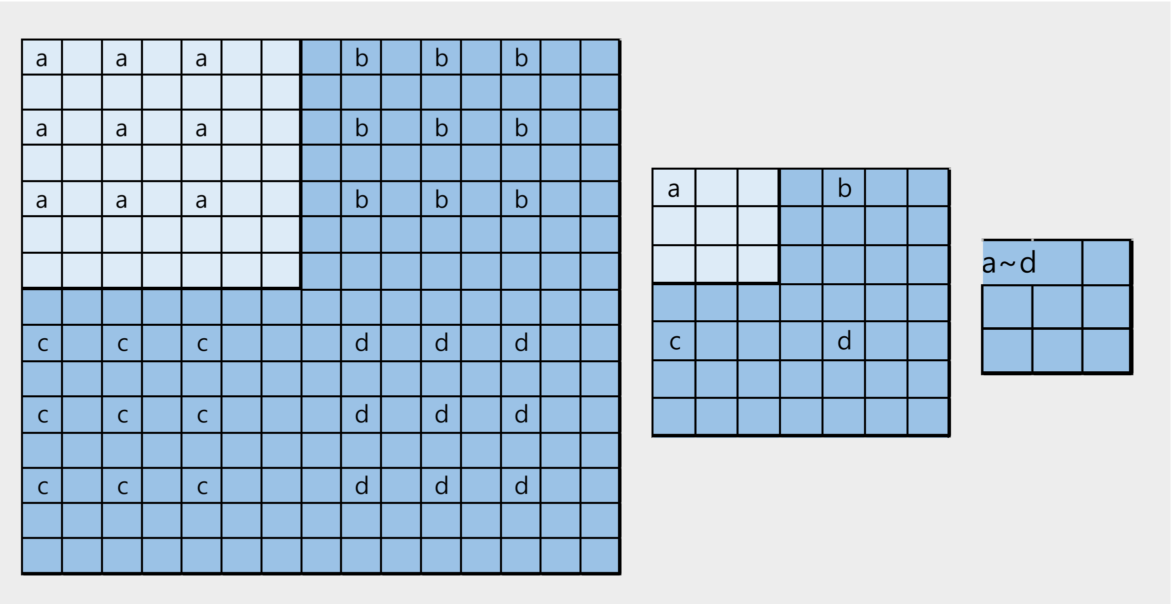
dilation rate = 2

출처: https://3months.tistory.com/213 [Deep Play]

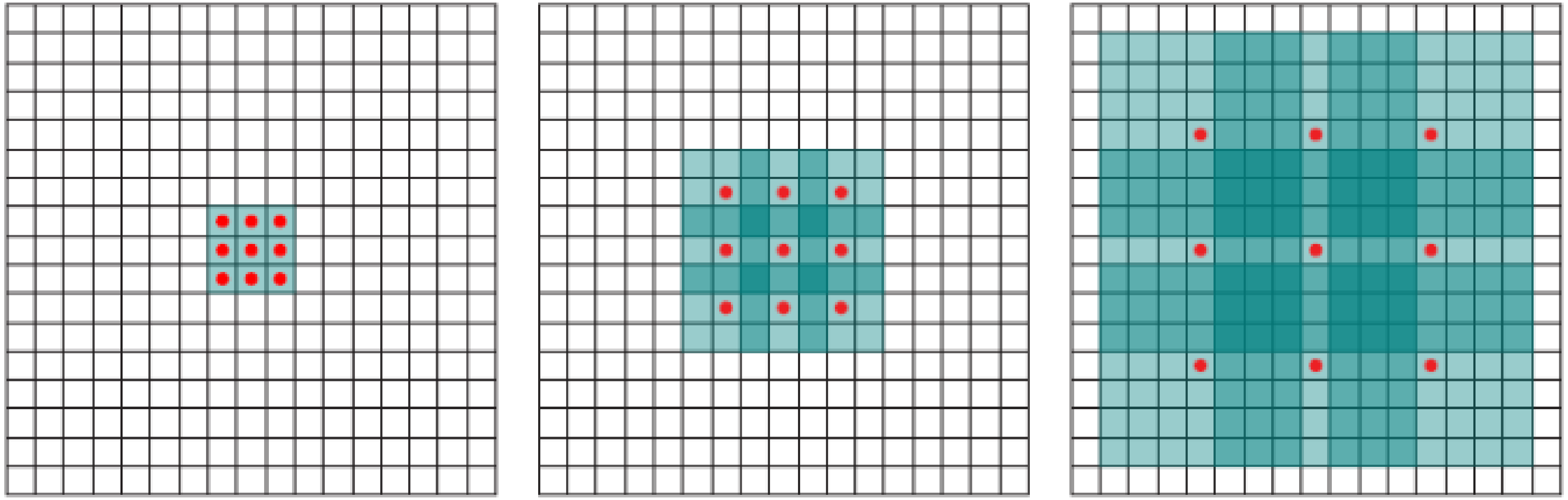
Atrous(Dilated) Convolution







Atrous(Dilated) Convolution



l=1 (left), l=2 (Middle), l=4 (Right)

전체적인 특징을 잡아내기 위해서는 **receptive field**는 높으면 높을 수록 좋다

Atrous(Dilated) Convolution

전체적인 특징을 잡아내기 위해서는 **receptive field**는 높으면 높을 수록 좋다

필터의 크기를 크게하면 연산량 + 오버피팅 +

기존 CNN에서는 conv + pooling 으로 해결

But, pooling 시 기존 정보 손실

따라서 Atrous, Dilated Conv 로 적은 연산량으로

보다 큰 receptive field 를 가져가자!

Deeplab v2 - ASPP

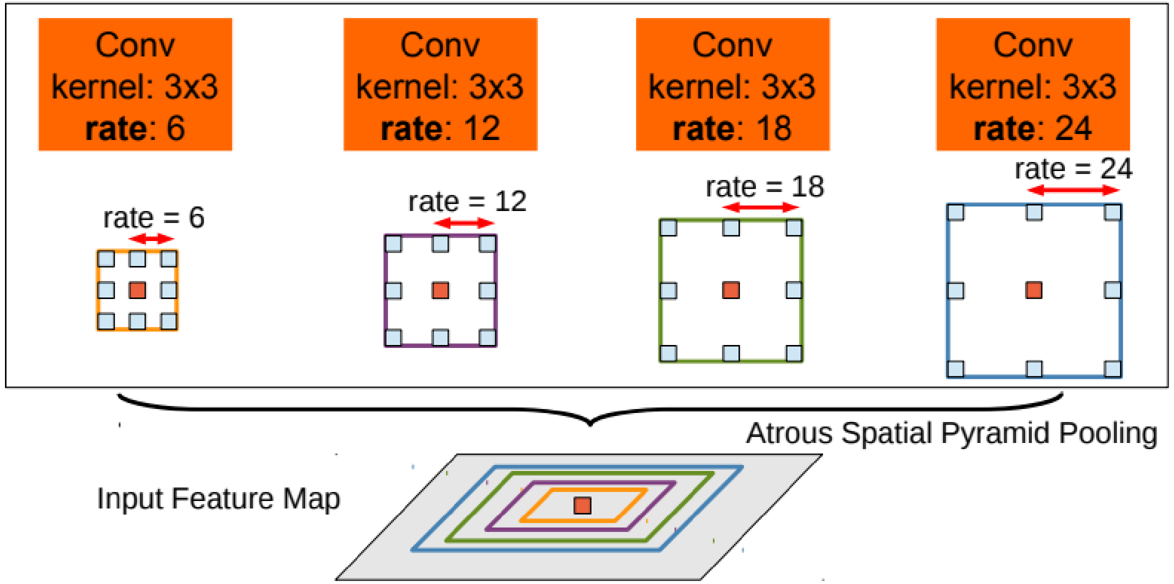
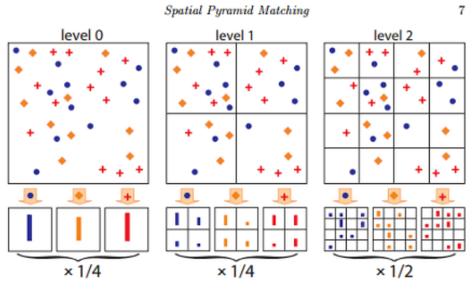


Fig. 4: Atrous Spatial Pyramid Pooling (ASPP). To classify the center pixel (orange), ASPP exploits multi-scale features by employing multiple parallel filters with different rates. The effective Field-Of-Views are shown in different colors.



[Spatial Pyramid Matching]

spatial pyramid pooling 기법과 같이

여러 개의 rate가 다른 atrous convolution을 병렬로 적용,

다시 합쳐(concat)주는 **ASPP** 기법

Deeplab v2 - ASPP

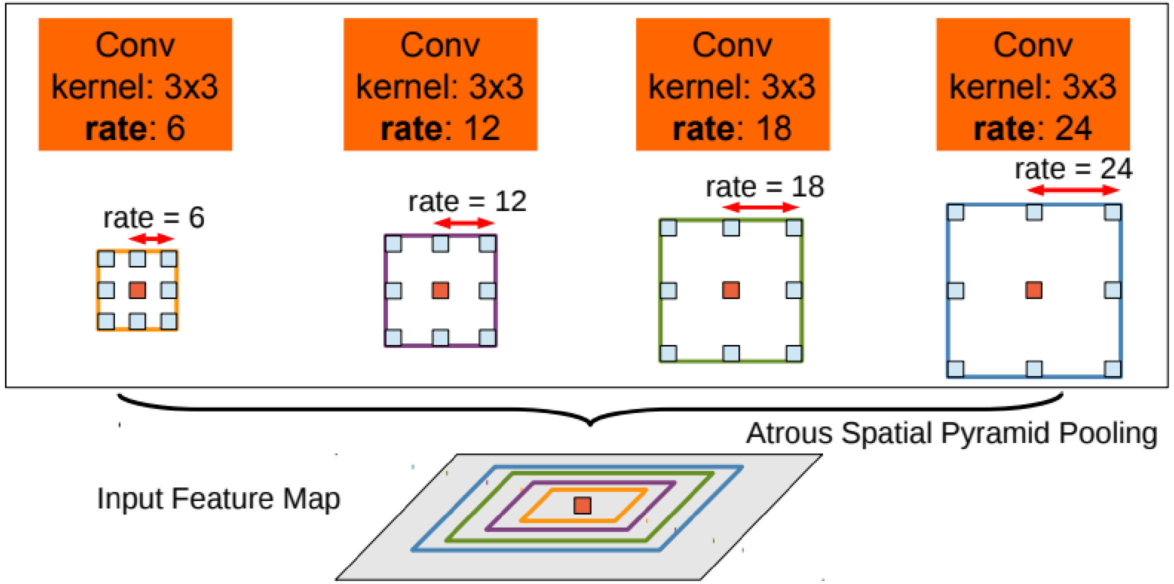
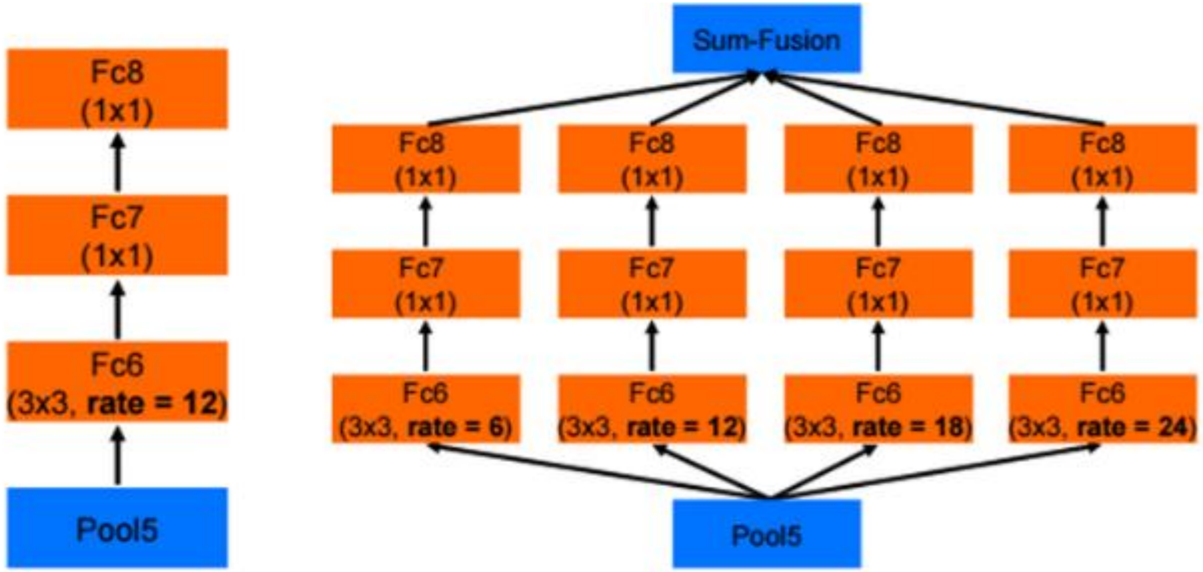


Fig. 4: Atrous Spatial Pyramid Pooling (ASPP). To classify the center pixel (orange), ASPP exploits multi-scale features by employing multiple parallel filters with different rates. The effective Field-Of-Views are shown in different colors.

```
if stride == 1:
    depth_padding = 'same'
else:
    kernel_size_effective = kernel_size + (kernel_size - 1) * (rate - 1)
    pad_total = kernel_size_effective - 1
    pad_beg = pad_total // 2
    pad_end = pad_total - pad_beg
    x = ZeroPadding2D((pad_beg, pad_end))(x)
    depth_padding = 'valid'
```

https://github.com/bonlime/keras-deeplab-v3-plus/blob/0c0bbcd21bbdcbc2cf7cb1e8acab836c43bbf26d/.ipynb_checkpoints/model-Copy1-checkpoint.py#L111

Deeplab v2 - ASPP



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

Method	before CRF	after CRF
LargeFOV	65.76	69.84
ASPP-S	66.98	69.73
ASPP-L	68.96	71.57

ASPP는 r 을 12로 고정시킨 것보다 1.7% +

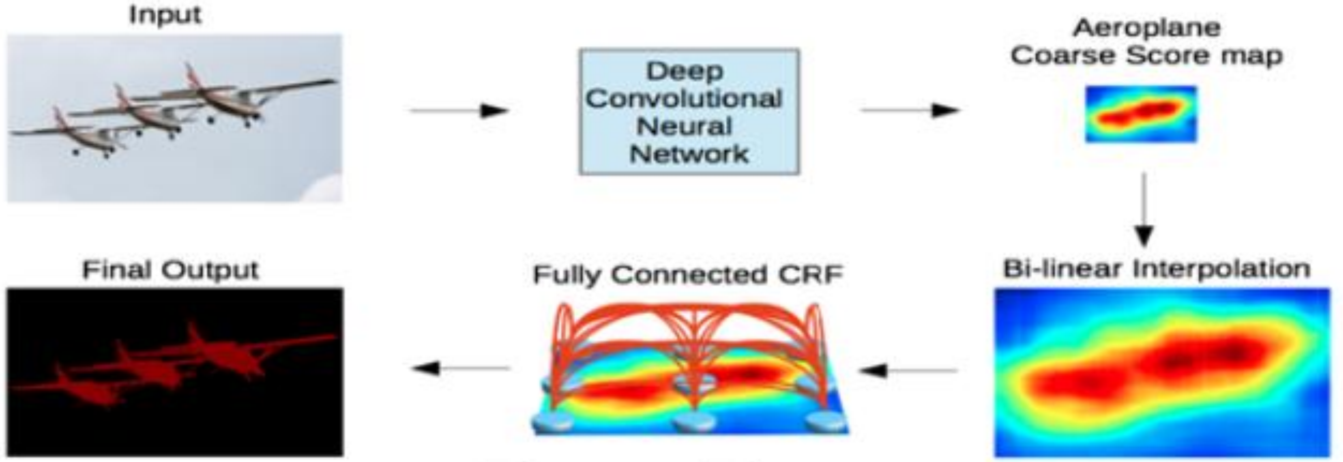
ASPP-S는 r = {2,4,8,12}

ASPP-L는 r = {6,12,18,24}

이때 실험 Backbone은 VGG-16 (DeepLab v2 의 best backbone 은 Resnet101)

추가로 CRF 가 있지만

Deeplab v2 - ASPP



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

What is **CRF**?

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

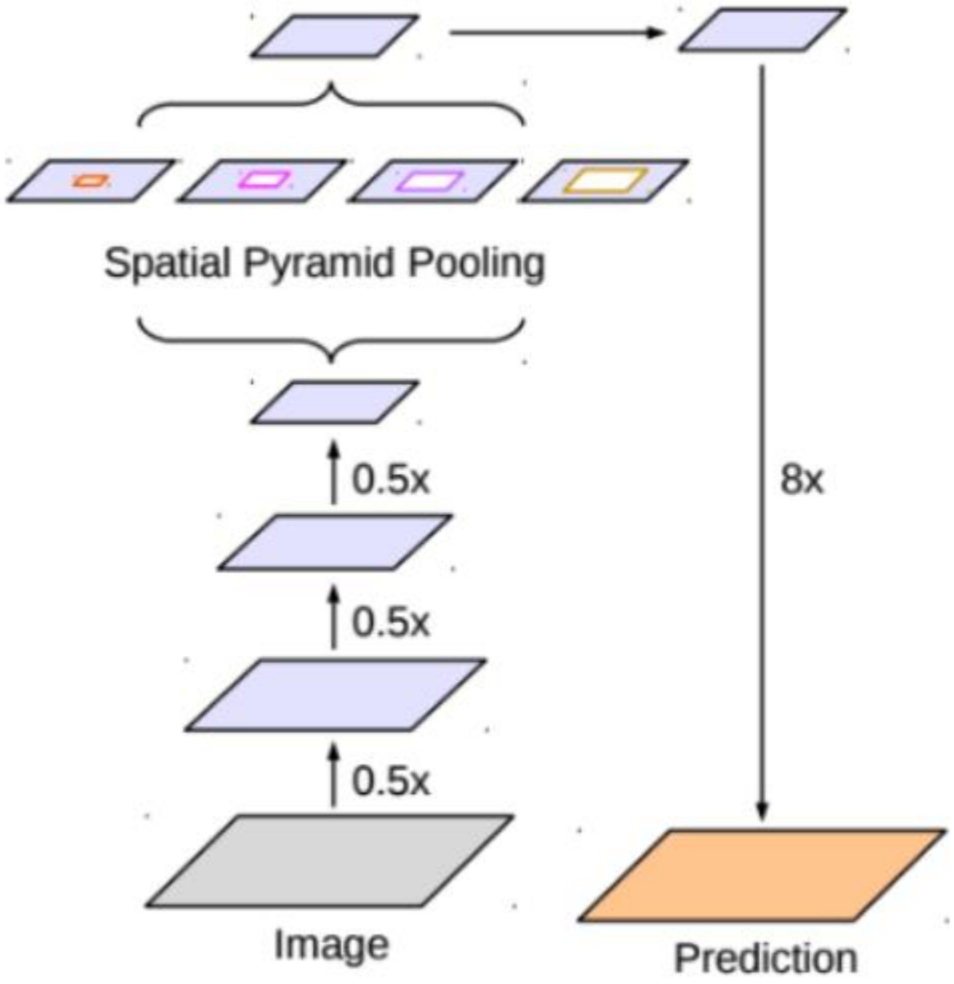
From DCNN label probabilities — Gaussian, pairwise

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

Differences in position and intensity — Just position

Mean Field Approximation with **Gaussian Convolutions**

Deeplab v3

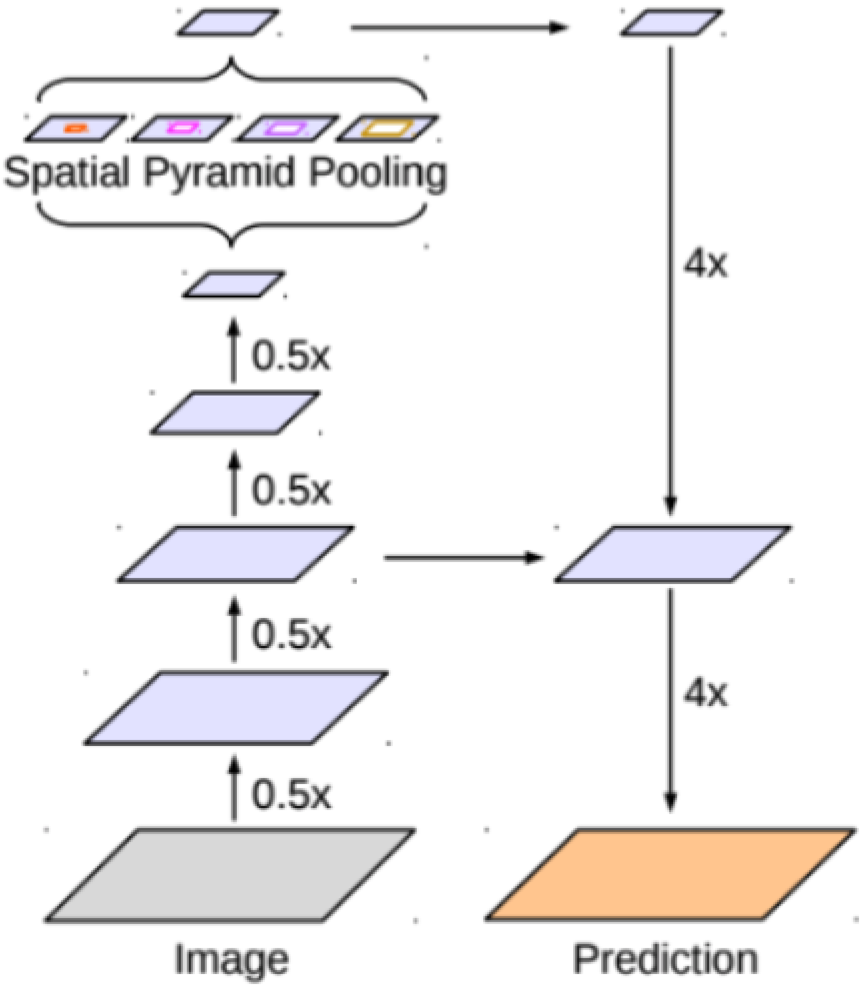


DeepLab V3 구조

V3에서는 기존 **ResNet** 구조에 atrous convolution을 활용해 좀 더 dense한 feature map을 얻는 방법을 여러가지 실험을 통해 제안

+ v2 의 CRF remove!

Deeplab v3 + : Depth wise conv

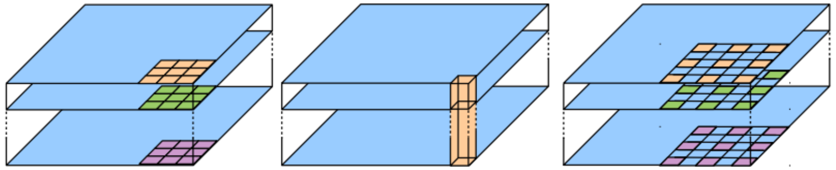


DeepLab V3+ 구조

Encoder: ResNet with atrous convolution → Xception (Inception with separable convolution, **Depth-wise** convolution)

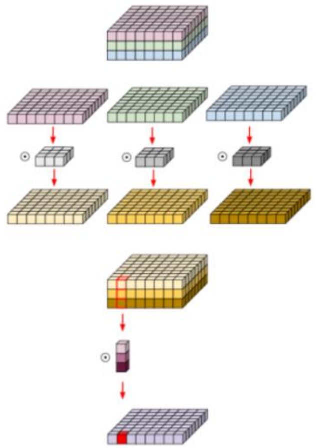
ASPP → **ASSPP** (Atrous **Separable** Spatial Pyramid Pooling)

Decoder: Bilinear upsampling → Simplified **U-Net style** decoder

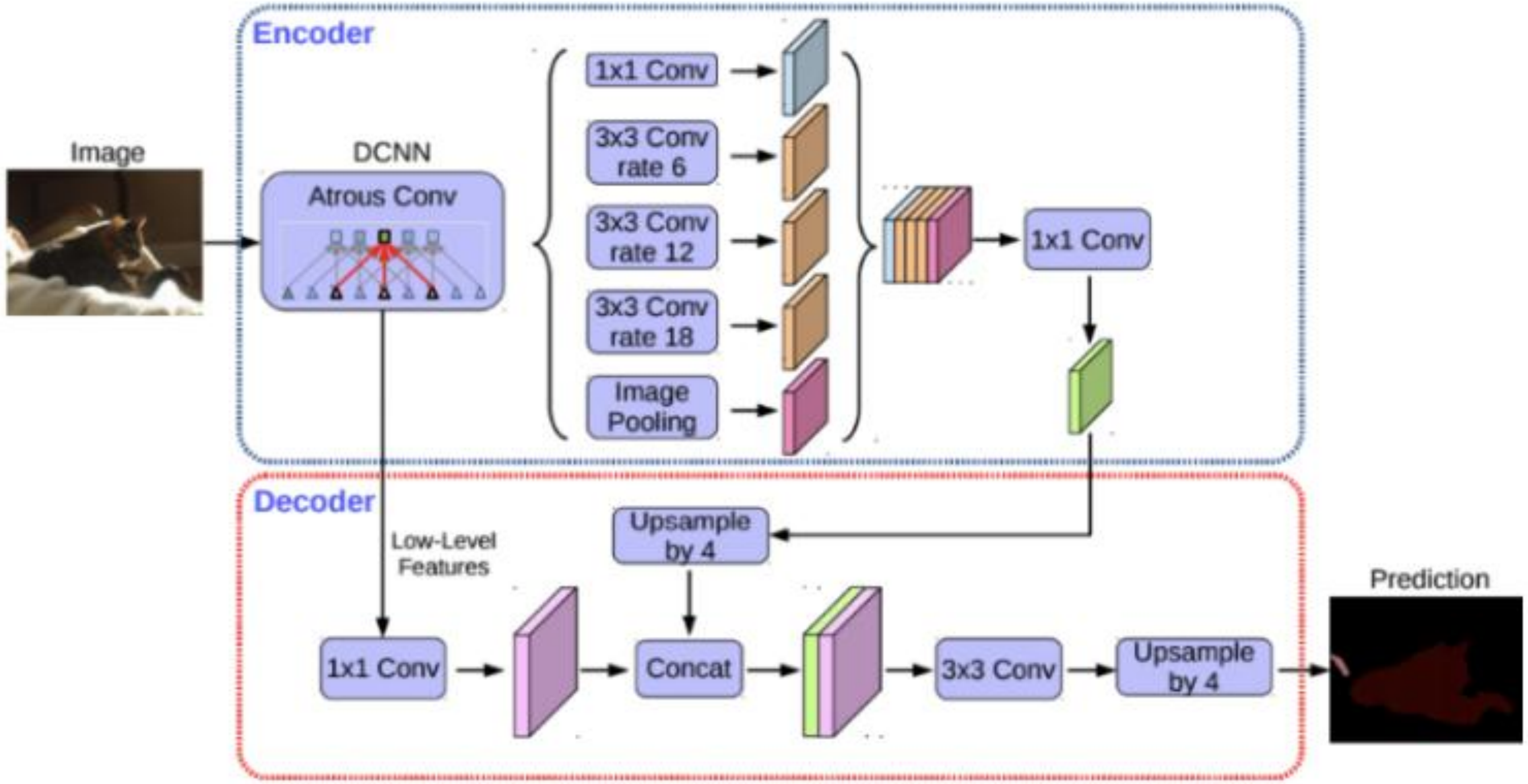


(a) Depthwise conv. (b) Pointwise conv. (c) Atrous depthwise conv.

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.



Deeplab v3 + Architecture



DeepLab V3+ 구조 디테일

Deeplab v3 + Result

Encoder train OS	Encoder eval OS	Decoder	MS	Flip	SC	COCO	JFT	mIOU	Multiply-Adds
16	16							79.17%	68.00B
16	16		✓					80.57%	601.74B
16	16		✓	✓				80.79%	1203.34B
16	8							79.64%	240.85B
16	8		✓					81.15%	2149.91B
16	8		✓	✓				81.34%	4299.68B
16	16	✓						79.93%	89.76B
16	16	✓	✓					81.38%	790.12B
16	16	✓	✓	✓				81.44%	1580.10B
16	8	✓						80.22%	262.59B
16	8	✓	✓					81.60%	2338.15B
16	8	✓	✓	✓				81.63%	4676.16B
16	16	✓			✓			79.79%	54.17B
16	16	✓	✓	✓	✓			81.21%	928.81B
16	8	✓			✓			80.02%	177.10B
16	8	✓	✓	✓	✓			81.39%	3055.35B
16	16	✓			✓	✓		82.20%	54.17B
16	16	✓	✓	✓	✓	✓		83.34%	928.81B
16	8	✓			✓	✓		82.45%	177.10B
16	8	✓	✓	✓	✓	✓		83.58%	3055.35B
16	16	✓			✓	✓	✓	83.03%	54.17B
16	16	✓	✓	✓	✓	✓	✓	84.22%	928.81B
16	8	✓			✓	✓	✓	83.39%	177.10B
16	8	✓	✓	✓	✓	✓	✓	84.56%	3055.35B

Table 5. Inference strategy on the PASCAL VOC 2012 *val* set when using modified Xception as feature extractor. **train OS**: The *output stride* used during training. **eval OS**: The *output stride* used during evaluation. **Decoder**: Employing the proposed decoder structure. **MS**: Multi-scale inputs during evaluation. **Flip**: Adding left-right flipped inputs. **SC**: Adopting depthwise separable convolution for both ASPP and decoder modules. **COCO**: Models pretrained on MS-COCO dataset. **JFT**: Models pretrained on JFT dataset.

Encoder: ResNet with atrous **mIOU 2% 향상**

convolution → **Xception** (Inception with separable

convolution, **Depth-wise** convolution)

연산량 감소

ASPP → **ASSPP** (Atrous **Separable** Spatial Pyramid Pooling)

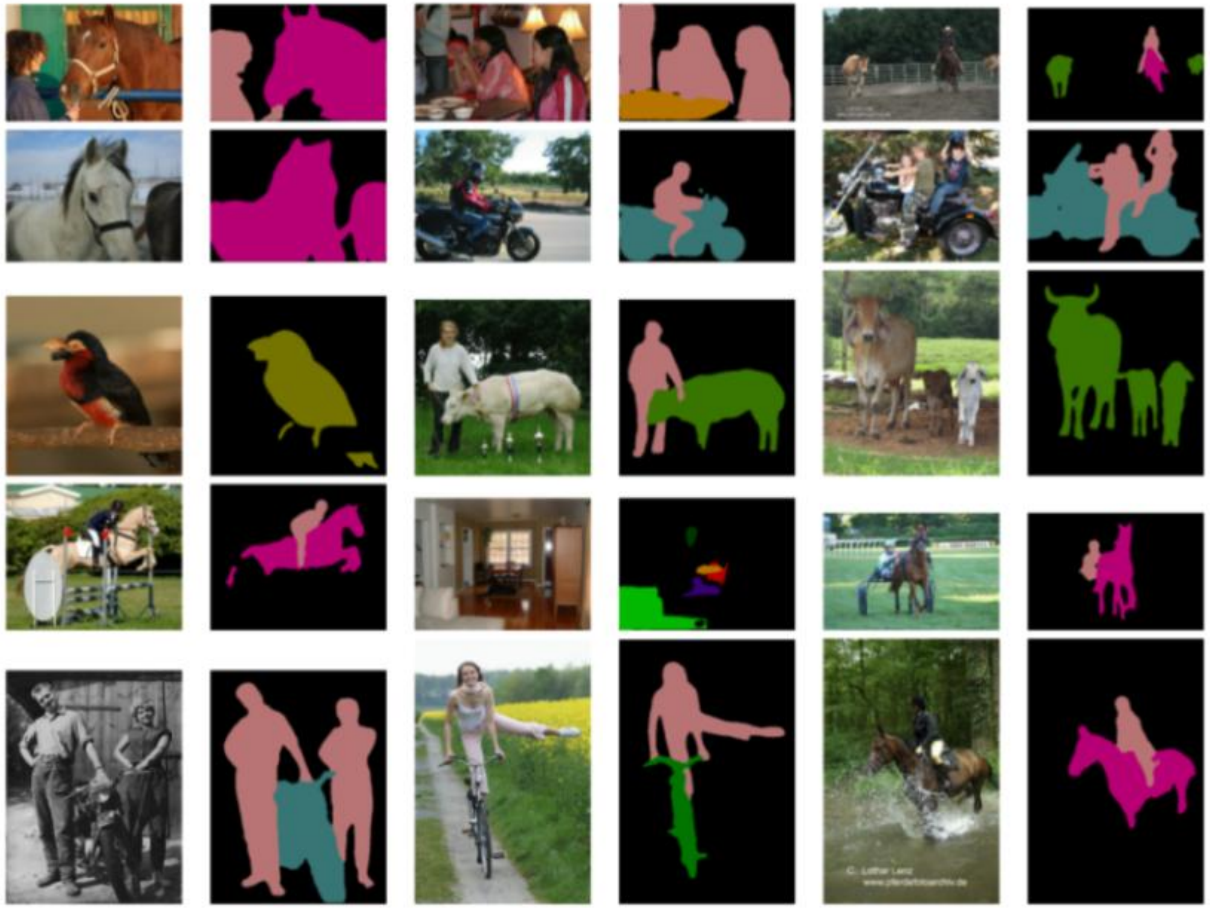
Decoder: Bilinear upsampling → Simplified **U-Net style**

decoder **mIOU 1.64% 향상**

Deeplab v3 + Result

Encoder train OS	Encoder eval OS	Decoder	MS	Flip	SC	COCO	JFT	mIOU	Multiply-Adds
16	16							79.17%	68.00B
16	16		✓					80.57%	601.74B
16	16		✓	✓				80.79%	1203.34B
16	8							79.64%	240.85B
16	8		✓					81.15%	2149.91B
16	8		✓	✓				81.34%	4299.68B
16	16	✓						79.93%	89.76B
16	16	✓	✓					81.38%	790.12B
16	16	✓	✓	✓				81.44%	1580.10B
16	8	✓						80.22%	262.59B
16	8	✓	✓					81.60%	2338.15B
16	8	✓	✓	✓				81.63%	4676.16B
16	16	✓			✓			79.79%	54.17B
16	16	✓	✓	✓	✓			81.21%	928.81B
16	8	✓			✓			80.02%	177.10B
16	8	✓	✓	✓	✓			81.39%	3055.35B
16	16	✓			✓	✓		82.20%	54.17B
16	16	✓	✓	✓	✓	✓		83.34%	928.81B
16	8	✓			✓	✓		82.45%	177.10B
16	8	✓	✓	✓	✓	✓		83.58%	3055.35B
16	16	✓			✓	✓	✓	83.03%	54.17B
16	16	✓	✓	✓	✓	✓	✓	84.22%	928.81B
16	8	✓			✓	✓	✓	83.39%	177.10B
16	8	✓	✓	✓	✓	✓	✓	84.56%	3055.35B

Table 5. Inference strategy on the PASCAL VOC 2012 *val* set when using modified *Xception* as feature extractor. **train OS**: The *output st* used during training. **eval OS**: The *output stride* used during evaluation. **Decoder**: Employing the proposed decoder structure. Multi-scale inputs during evaluation. **Flip**: Adding left-right flipped inputs. **SC**: Adopting depthwise separable convolution for both A and decoder modules. **COCO**: Models pretrained on MS-COCO dataset. **JFT**: Models pretrained on JFT dataset.



Pascal VOC 2012 validation set에서의 visualization 결과

Summary

DeepLab V1

Atrous Convolution

DeepLab V2

Multi-scale context 적용을 위한 Atrous Spatial Pyramid Pooling (ASPP) + (CRF)

DeepLab V3

ResNet 구조에 Atrous convolution을 활용해 좀 더 dense 한 feature map 얻기

DeepLab V3+

Separable Convolution 과 Atrous convolution 을 결합한 Atrous separable Convolution 의 활용 제안

참고 자료

V1 : https://arxiv.org/abs/1412.7062

V2 : https://arxiv.org/abs/1606.00915

V3 : https://arxiv.org/abs/1706.05587

V3 + : https://arxiv.org/abs/1802.02611

전체 발표 구조 : https://blog.lunit.io/2018/07/02/deeplab-v3-encoder-decoder-with-atrous-separable-convolution-for-semantic-image-segmentation/

1, 2 설명 : https://m.blog.naver.com/PostView.nhn?blogId=laonple&logNo=221000648527&proxyReferer=https%3A%2F%2Fwww.google.com%2F

참고 정보 슬라이드 : https://www.slideshare.net/WhiKwon/fcn-to-deeplabv3