BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation-ECCV2018

KIST
송명하

Content

Introduction

출처 : https://medium.com/hyunjulie/1%ED%8E%B8-semantic-segmentation-%EC%B2%AB%EA%B1%B8%EC%9D%8C-4180367ec9cb

Introduction

출처 : https://medium.com/hyunjulie/1%ED%8E%B8-semantic-segmentation-%EC%B2%AB%EA%B1%B8%EC%9D%8C-4180367ec9cb

Introduction

출처 : https://medium.com/hyunjulie/1%ED%8E%B8-semantic-segmentation-%EC%B2%AB%EA%B1%B8%EC%9D%8C-4180367ec9cb

Introduction

출처 : https://medium.com/hyunjulie/1%ED%8E%B8-semantic-segmentation-%EC%B2%AB%EA%B1%B8%EC%9D%8C-4180367ec9cb

Introduction

Real time ?

Introduction

Real time ? 실시간

Introduction

Real time Semantic Segmentation

Introduction

Real time Semantic Segmentation

속도가 빠르면 빠를수록 성능이 안좋음

Introduction

Real-time Semantic Segmentation 가속화하는 3가지 방법

1. Input size고정(Crop, resize)

2. Resize대신 채널 가지치기방법

3. 마지막 단에 down sampling 많이 하기

Introduction

Real-time Semantic Segmentation 가속화하는 3가지 방법

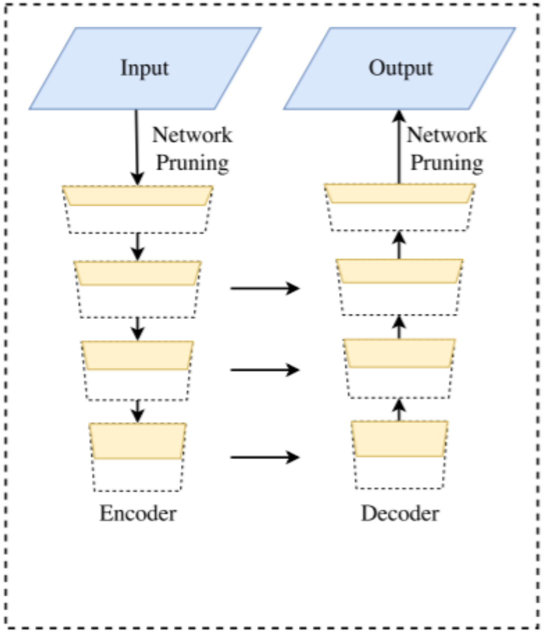
1. Input size고정(Crop, resize)
   - Spatial 정보 잃음

2. Resize대신 채널 가지치기방법
   - 공간 능력을 약하게 만듦

3. 마지막 단에 down sampling 많이 하기
   - Receptive Field가 충분히 커지지 못해서 discriminative ability가 poor

Introduction

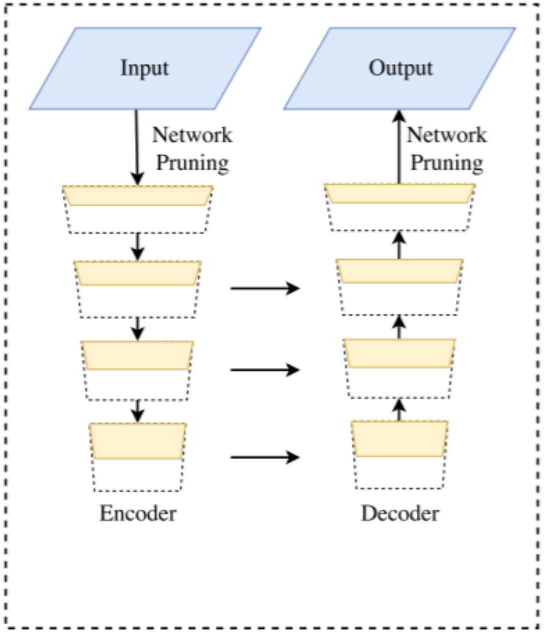
공간정보를 잃지않고 잘 사용하는 방법 -> U-Net구조가 널리 사용됨



계층적 Feature를 backbone에서 섞으며 spatial resolution을 점차 증가시키고, detail을 살림

Introduction

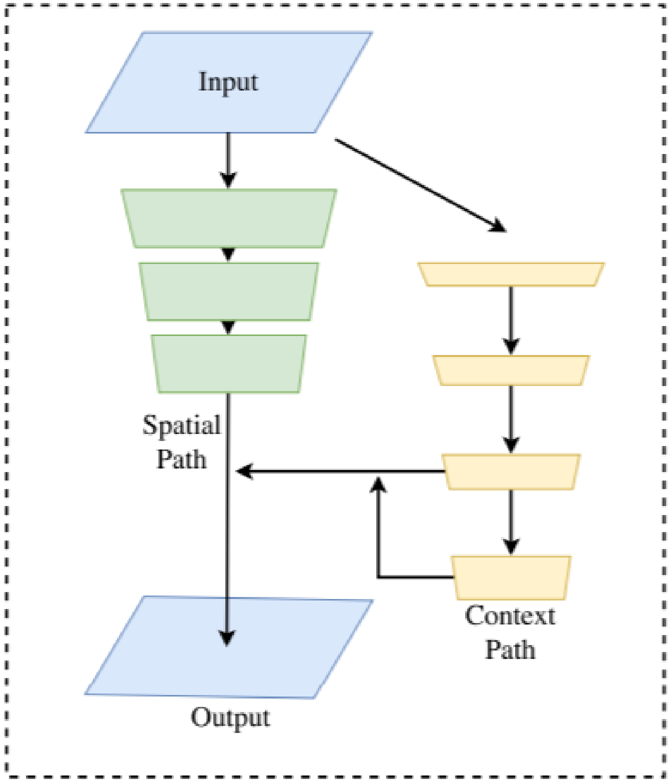
공간정보를 잃지않고 잘 사용하는 방법 -> U-Net구조가 널리 사용됨



1. 속도 저하
   -> 초기 high resolution feature 때문에
2. 대부분의 spatial information이 pruning 또는 cropping에 잃은 것들 쉽게 복원이 안됨

Introduction

BiseNet with two parts :



1. Spatial Path
   3개의 conv 스택 1/8 feature map을 얻기 위해

2. Context path
   global average pooling layer, Xception을 추가
   -> Backbone network의 receptive field
   maximum.

3. Feature Fusion Module, Attention refinement
   Module 만들음

Introduction

Contribution

1. Decoupling the function of spatial information preservation and receptive filed offering into two paths. Specifically, we propose a Bilateral Segmentation Network (BiSeNet) with Spatial Path and Context path

2. We design two specific modules, Feature Fusion Module(FFM) and Attention Refinement Module(ARM), to further improve the accuracy with acceptable cost.

3. We achieve impressive results on the benchmarks of Cityscapes, CamVid and coco-stuff. More Specifically, we obtain the results of 68.4% in the Cityscapes test dataset with the speed of 105FPS

Related work

1. Spatial information preserving

2. U-Shape method

3. Context information

4. Real time segmentation

Related work

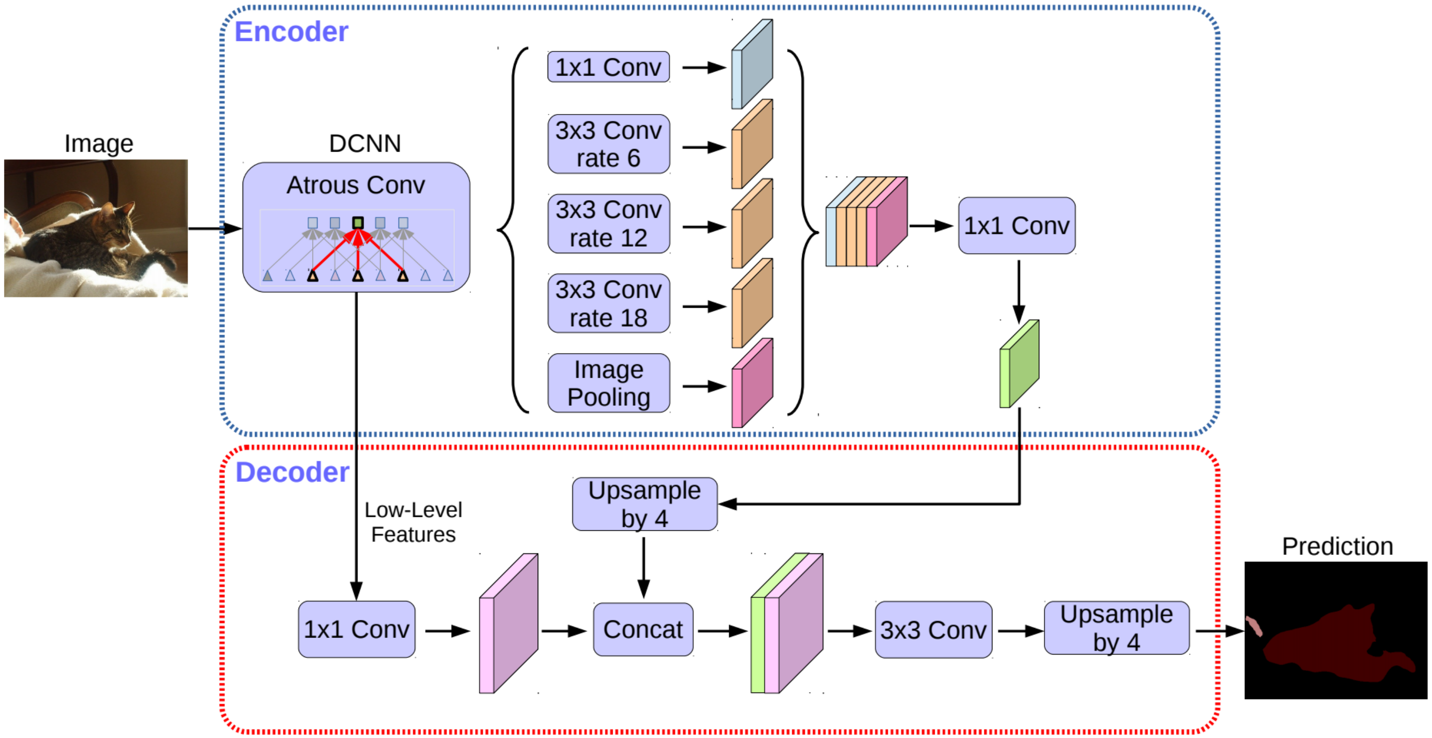
1. Spatial information preserving
   - DUC, PSPNet, DeepLab v2, Deeplab v3

   - Global Convolution Network

Related work

1. Spatial information preserving

- DUC, PSPNet, DeepLab v2, Deeplab v3 use the dilated convolution to preserve the spatial size of the feature map.

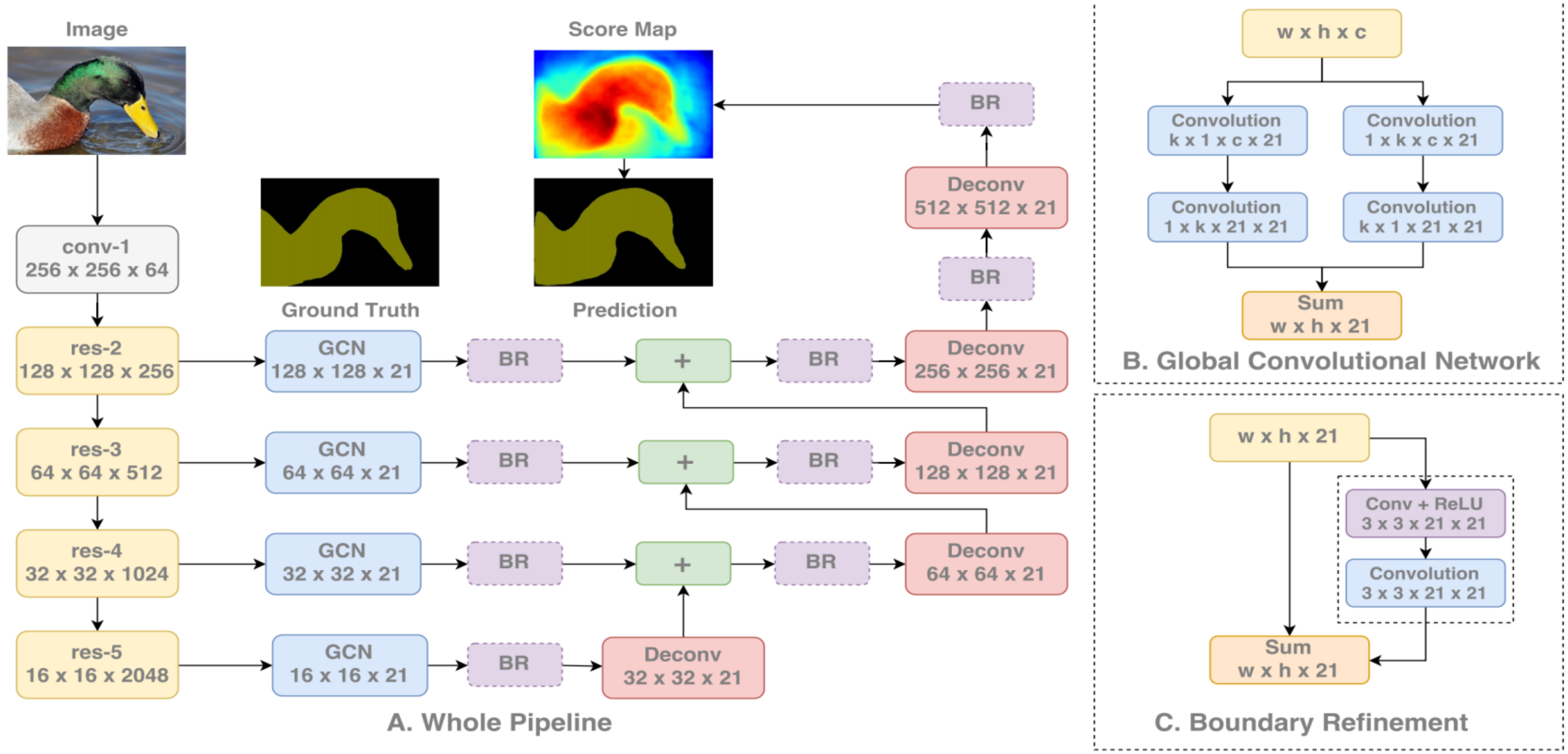


Encoder–Decoder with Atrous Separable Convolution for Semantic Image Segmentation

Related work

1. Spatial information preserving

- Global Convolution Network utilizes the "Large kernel" to enlarge the receptive filed**(Large Kernel)**

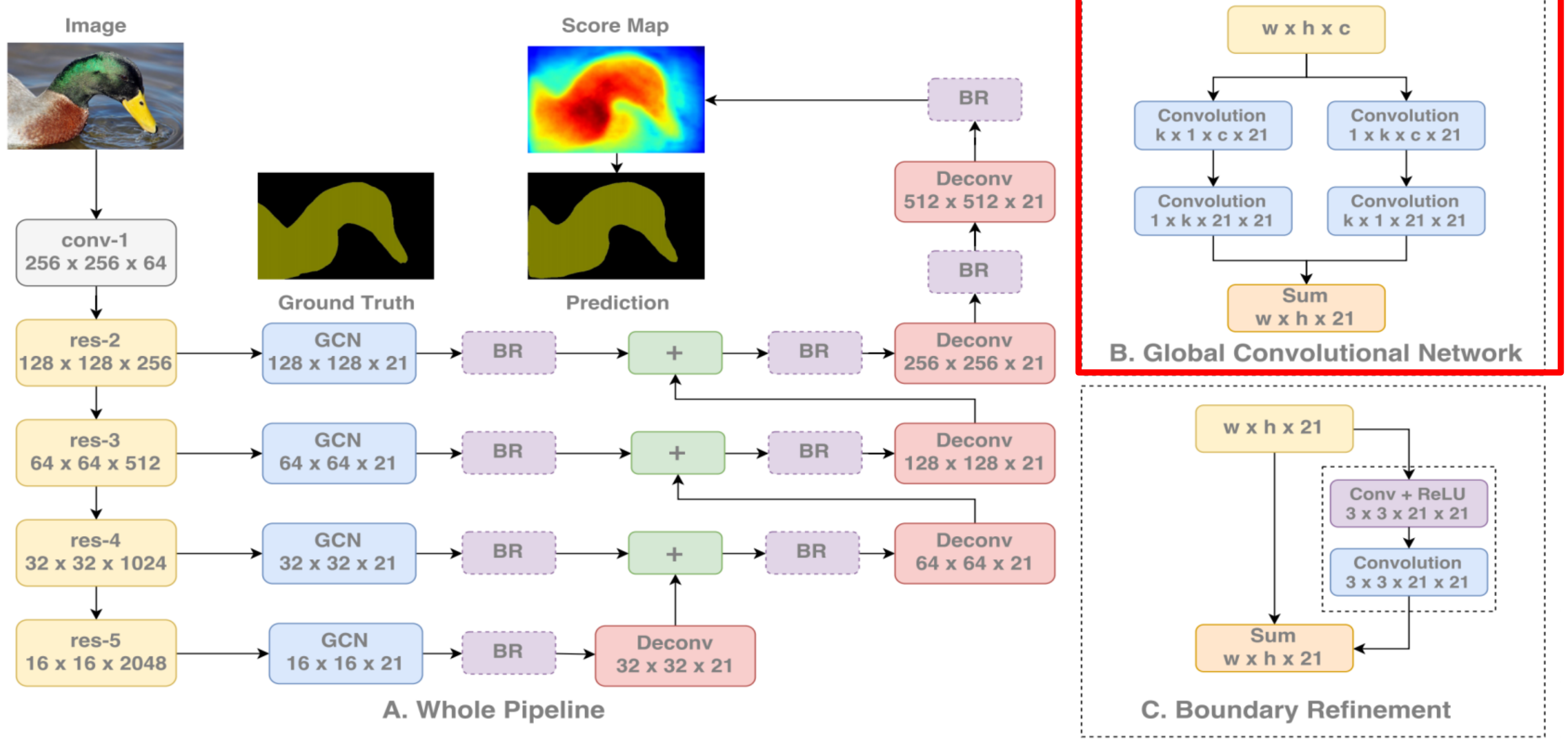


Large Kernel Matters -- Improve Semantic Segmentation by Global Convolutional Network

Related work

1. Spatial information preserving

- Global Convolution Network utilizes the “Large kernel” to enlarge the receptive filed**(Large Kernel)**

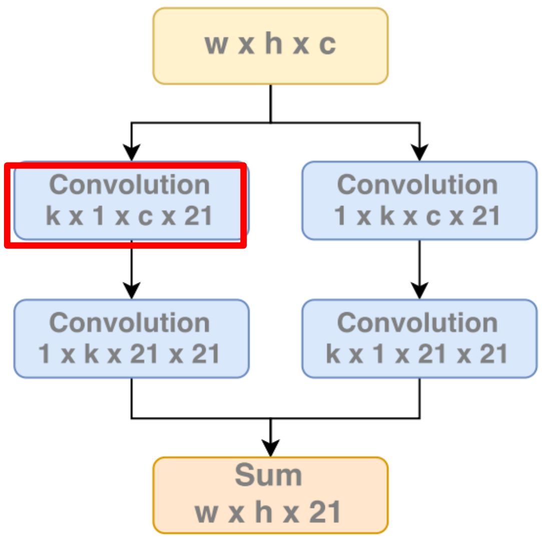


Large Kernel Matters -- Improve Semantic Segmentation by Global Convolutional Network

Related work

1. Spatial information preserving

- Global Convolution Network utilizes the "Large kernel" to enlarge the receptive filed**(Large Kernel)**



B. Global Convolutional Network

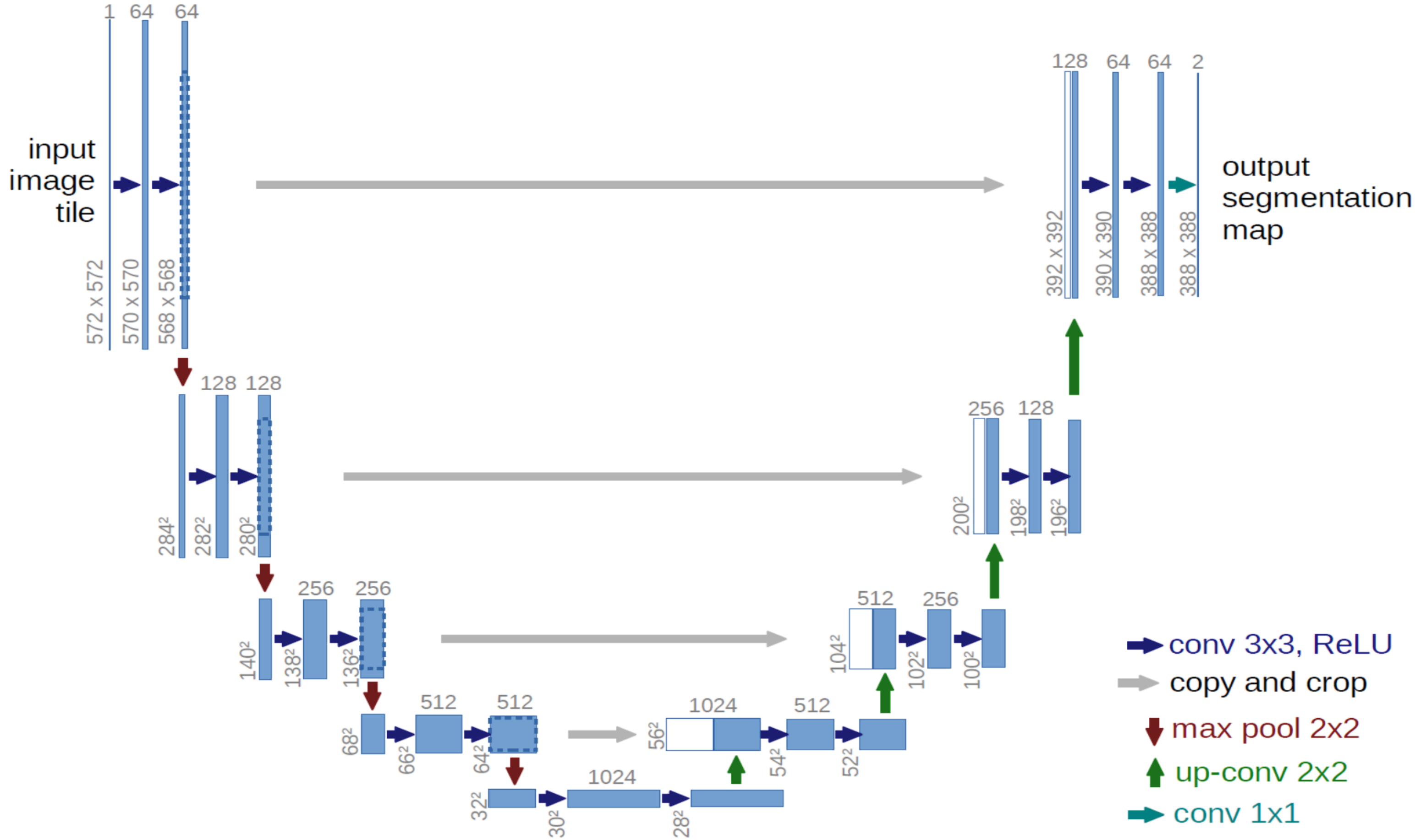
```
self.conv_l1 = nn.Conv2d(c, out_c, kernel_size=(k[0],1), padding =((int(k[0]-1)/2),0))
self.conv_l2 = nn.Conv2d(out_c, out_c, kernel_size=(1,k[0]), padding =(0,int((k[0]-1)/2)))
self.conv_r1 = nn.Conv2d(c, out_c, kernel_size=(1,k[1]), padding =(0,int((k[1]-1)/2)))
self.conv_r2 = nn.Conv2d(out_c, out_c, kernel_size=(k[1],1), padding =(int((k[1]-1)/2),0))
```

https://github.com/SConsul/Global_Convolutional_Network

Large Kernel Matters -- Improve Semantic Segmentation by Global Convolutional Network

Related work

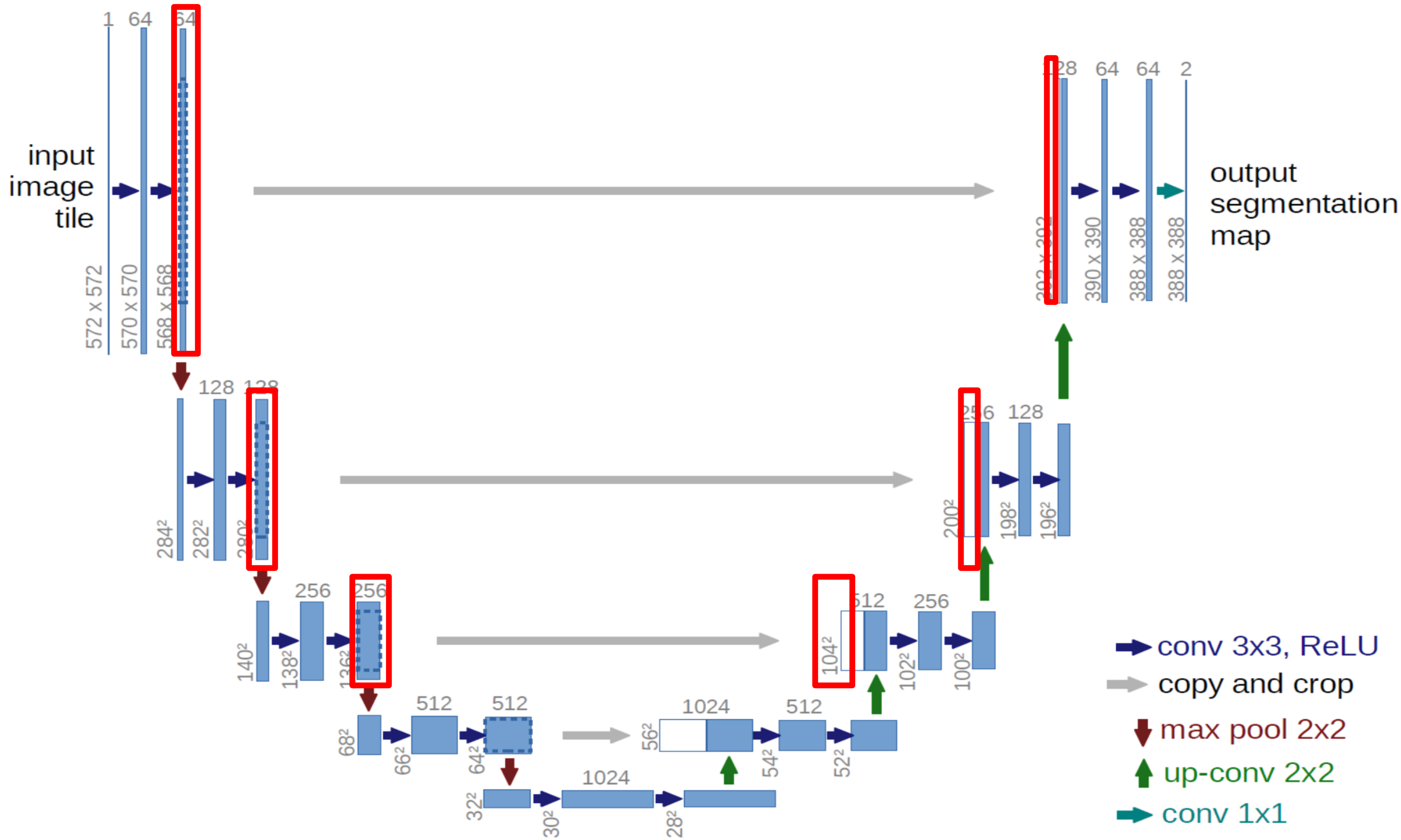
2. U-Shape method



https://tuatini.me/practical-image-segmentation-with-unet/

Related work

2. U-Shape method



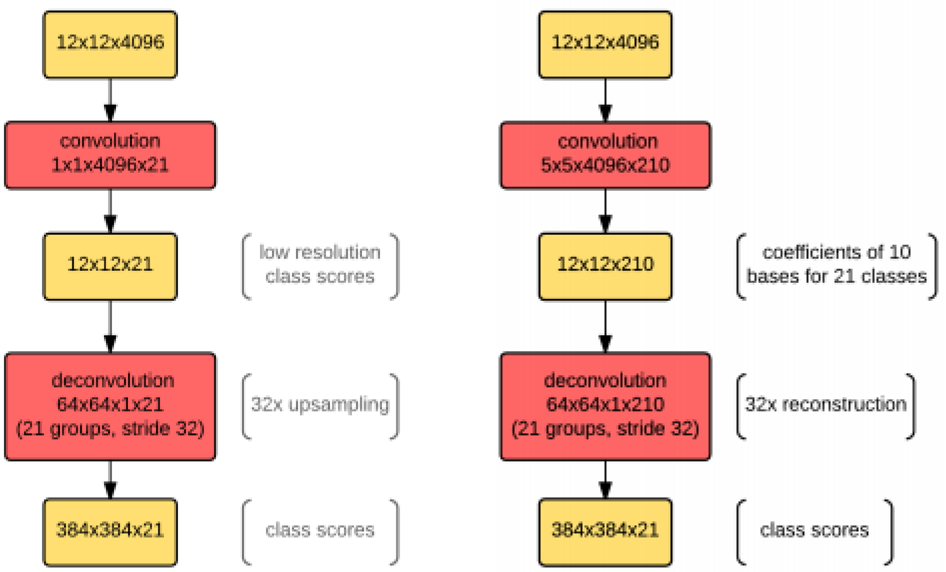
https://tuatini.me/practical-image-segmentation-with-unet/

Related work

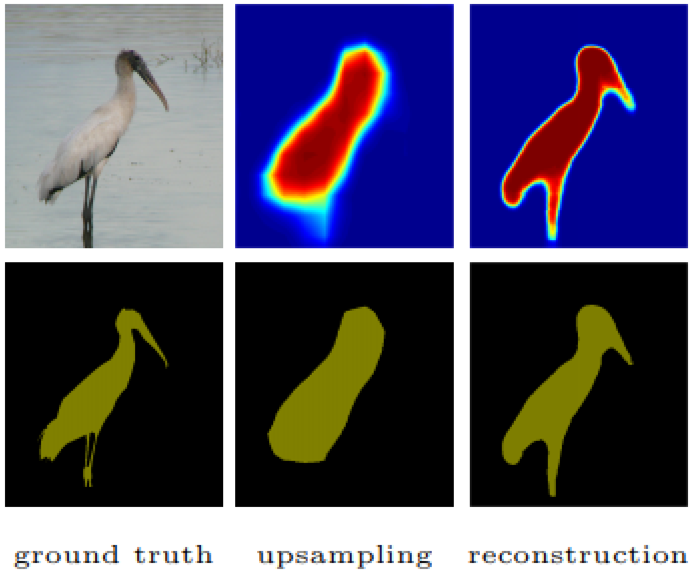
2. U-Shape method
   - LRR (Laplacian Pyramid Reconstruction) Network.
   - DFN (Discriminative Feature Network)

Related work

2. U-Shape method
- LRR (Laplacian Pyramid Reconstruction) Network.



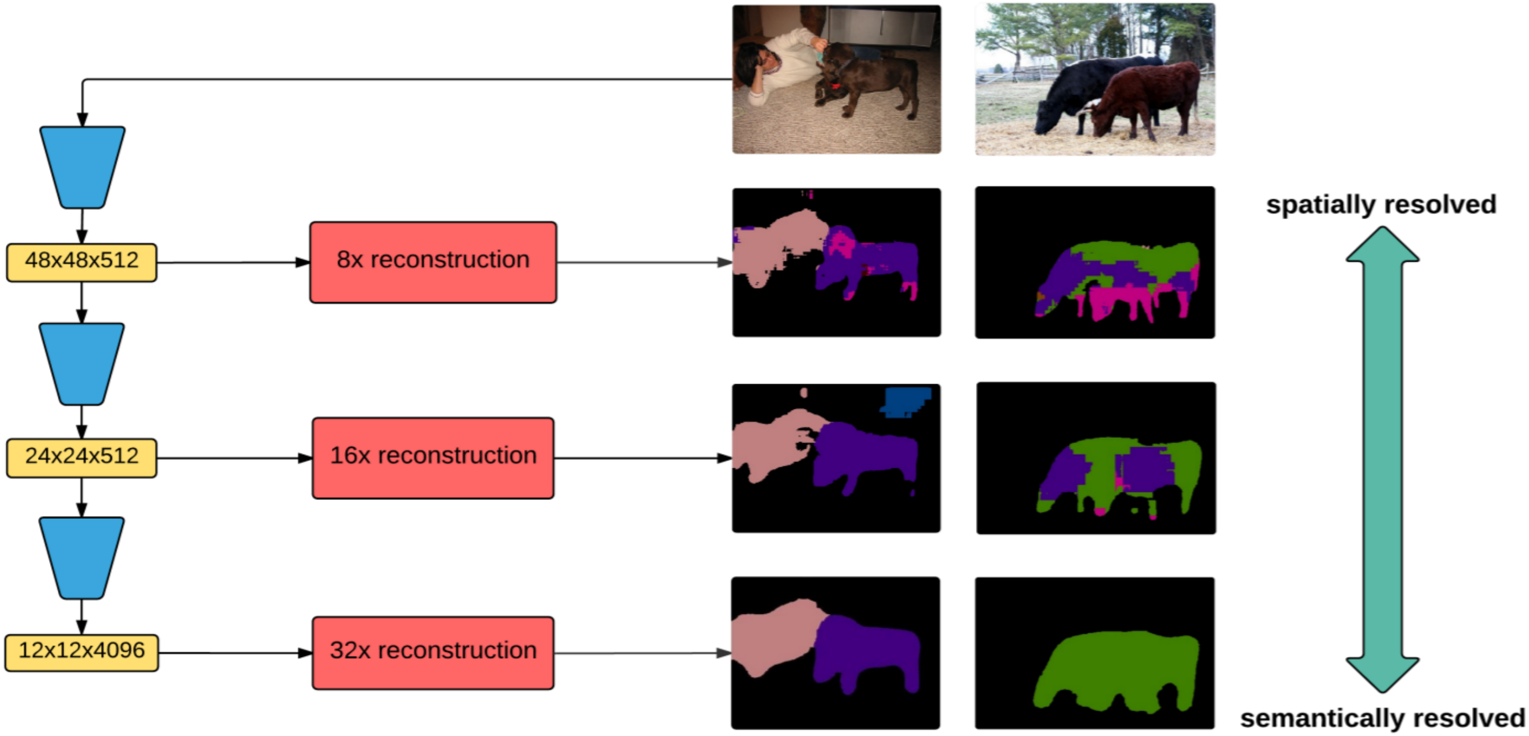
(a)



(b)

Related work

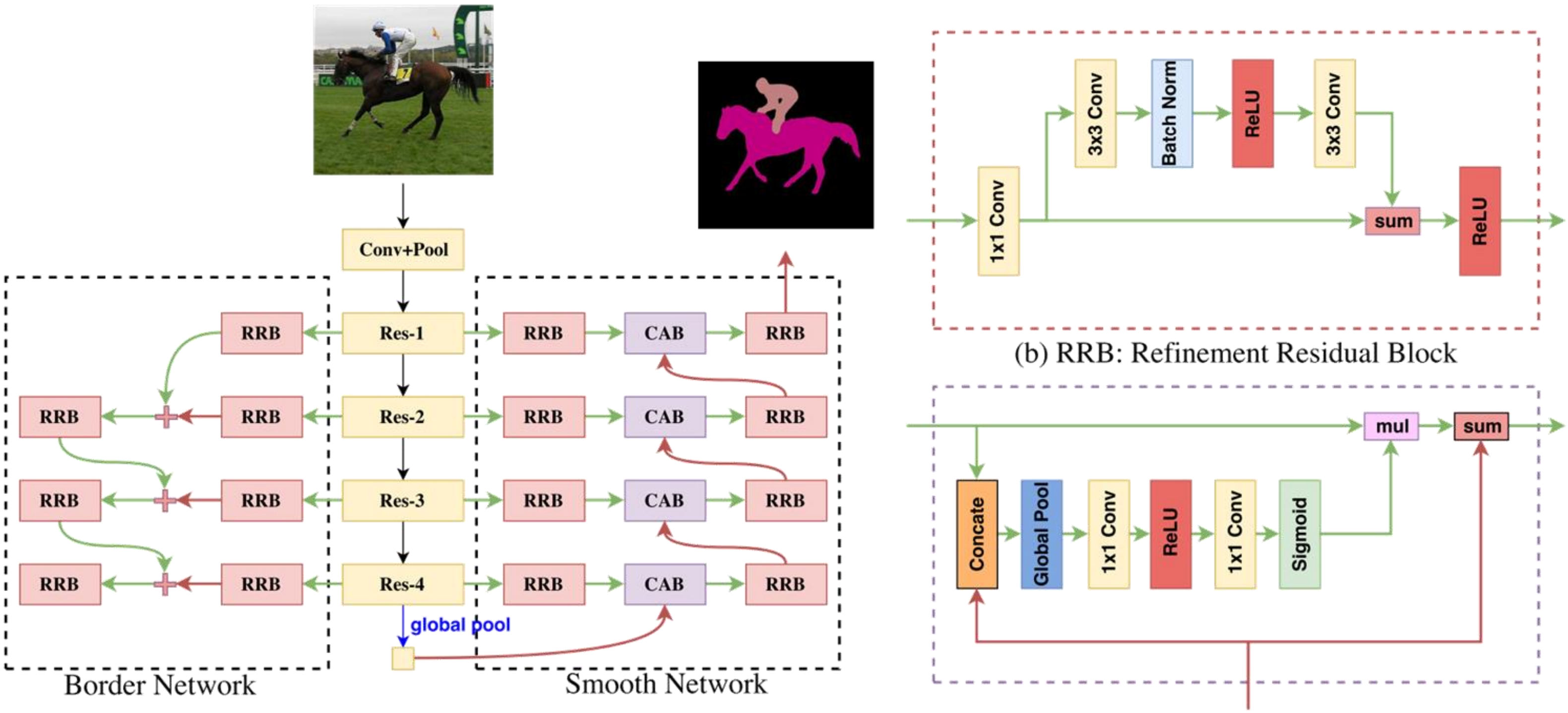
2. U-Shape method
- LRR (Laplacian Pyramid Reconstruction) Network.



Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation

Related work

2. U-Shape method
- DFN (Discriminative Feature Network)



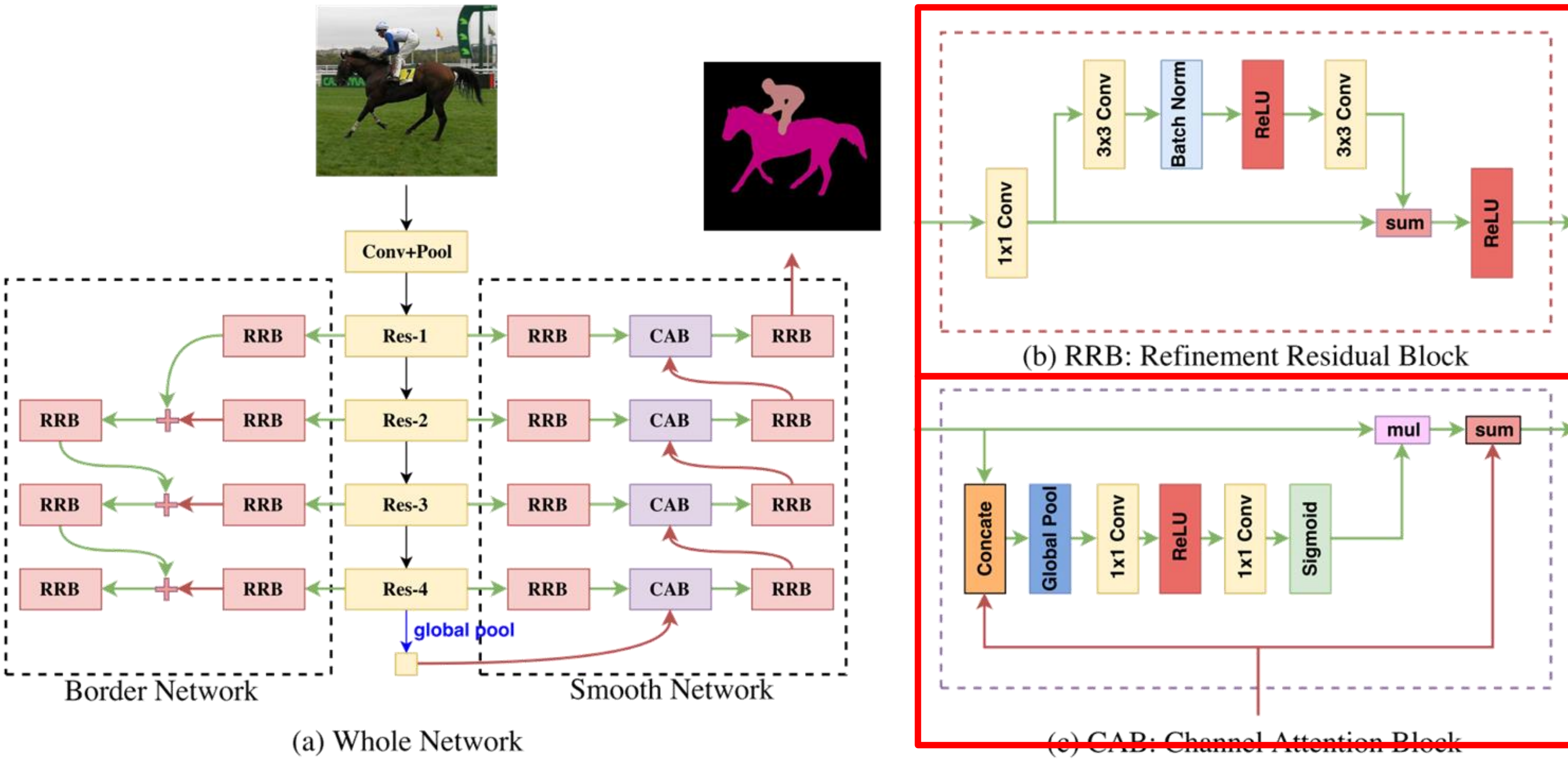
(a) Whole Network (c) CAB: Channel Attention Block

Learning a Discriminative Feature Network for Semantic Segmentation

designs a channel attention block to achieve the feature selection

Related work

2. U-Shape method
- DFN (Discriminative Feature Network)



Learning a Discriminative Feature Network for Semantic Segmentation

designs a channel attention block to achieve the feature selection

Related work

2. U-Shape method

However, in the U-shape structure, some lost spatial information cannot be easily recovered

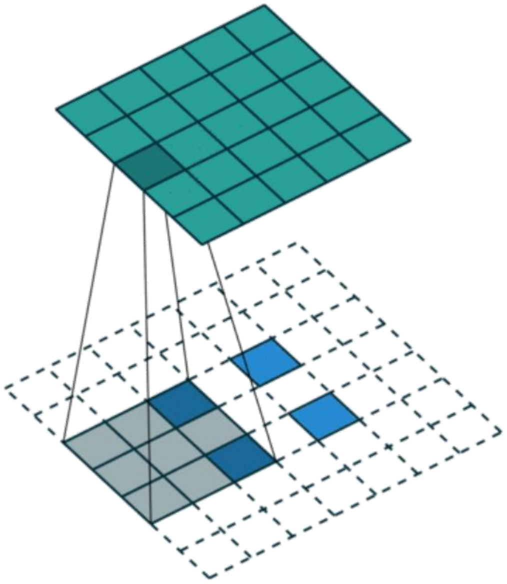
Related work

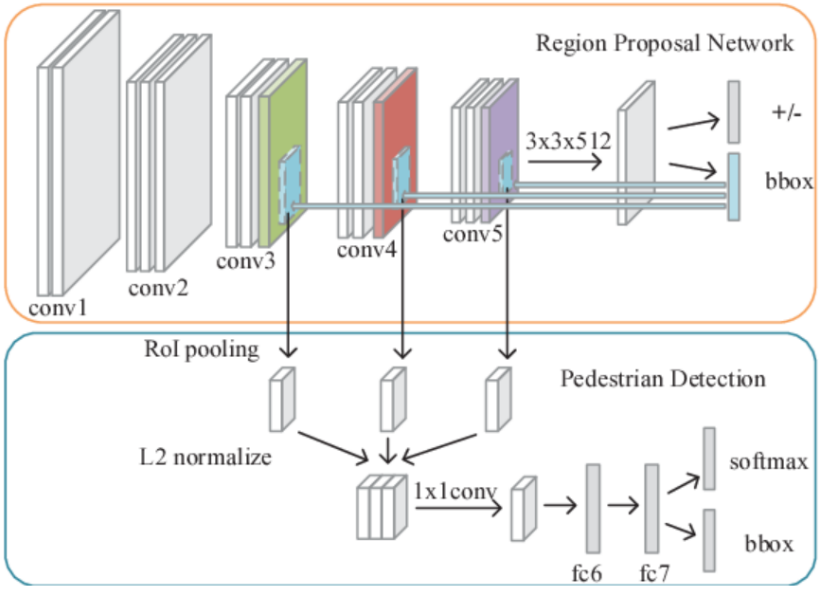
3. Context information

주된 방법 enlarge the receptive field or fuse different context information

Related work

3. Context information



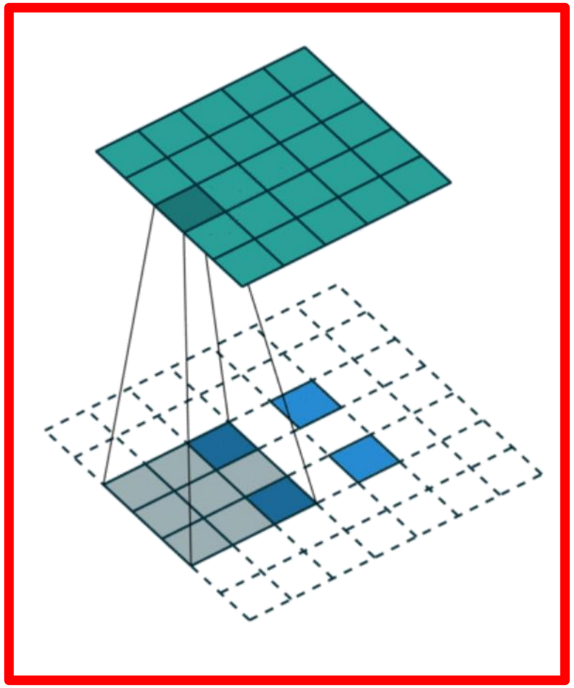


https://zzsza.github.io/data/2018/02/23/introduction-convolution/

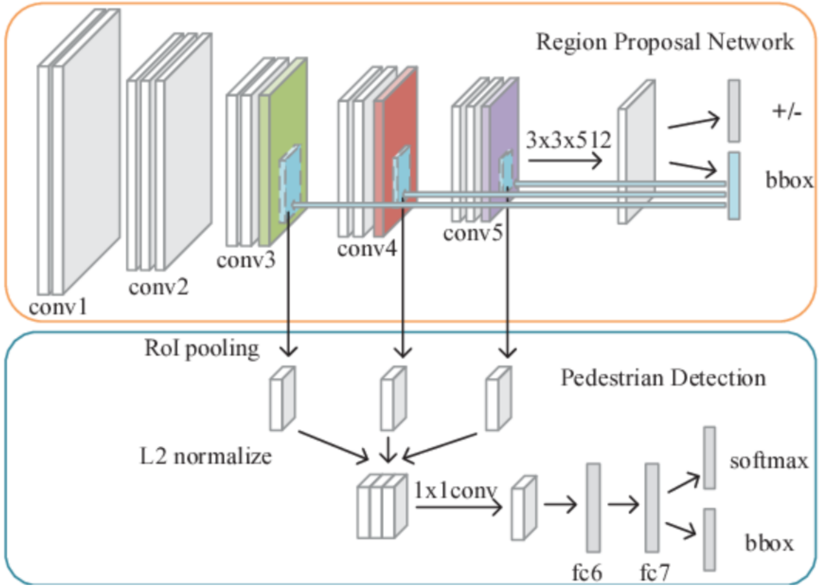
https://www.semanticscholar.org/paper/Pedestrian-detection-via-multi-scale-feature-fusion-Guo-Yin/07a6468d70dd62ce63a90b1b67651729f9c3037a/figure/0

Related work

3. Context information



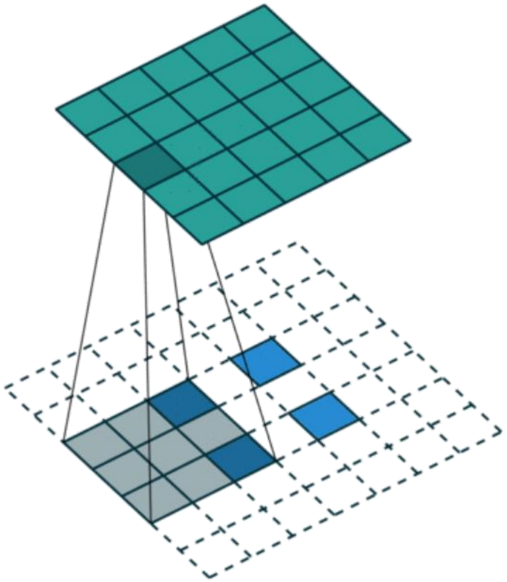
https://zzsza.github.io/data/2018/02/23/introduction-convolution/



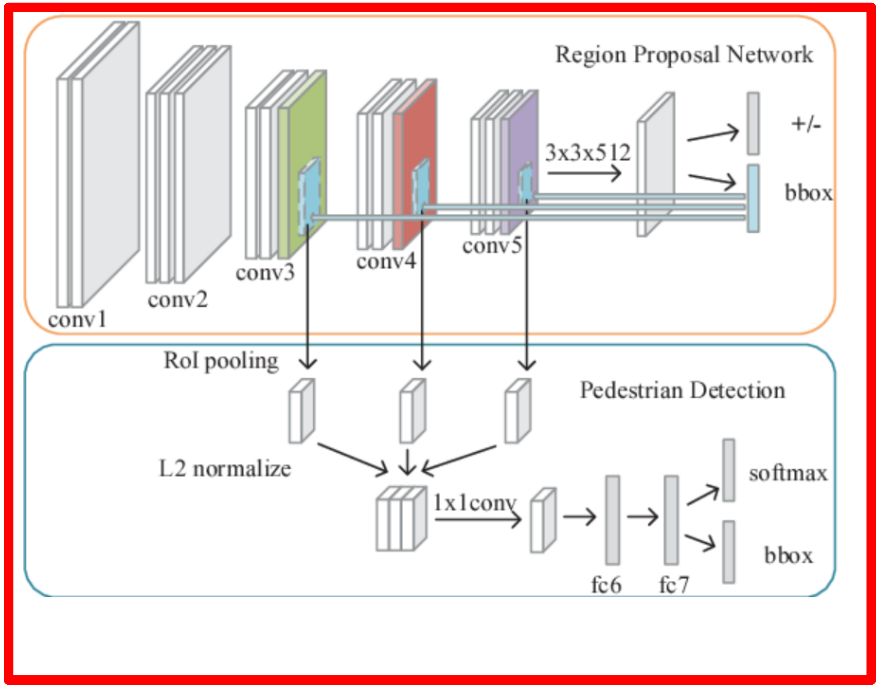
https://www.semanticscholar.org/paper/Pedestrian-detection-via-multi-scale-feature-fusion-Guo-Yin/07a6468d70dd62ce63a90b1b67651729f9c3037a/figure/0

Related work

3. Context information



https://zzsza.github.io/data/2018/02/23/introduction-convolution/



https://www.semanticscholar.org/paper/Pedestrian-detection-via-multi-scale-feature-fusion-Guo-Yin/07a6468d70dd62ce63a90b1b67651729f9c3037a/figure/0

Related work

3. Context information

- Rethinking Atrous Convolution for Semantic Image Segmentation ASPP(Atrous Spatial Pyramid Pooling) module
- PSPNet(Pyramid Scene Parsing Network)

Related work

3. Context information

- Rethinking Atrous Convolution for Semantic Image Segmentation ASPP(Atrous Spatial Pyramid Pooling) module

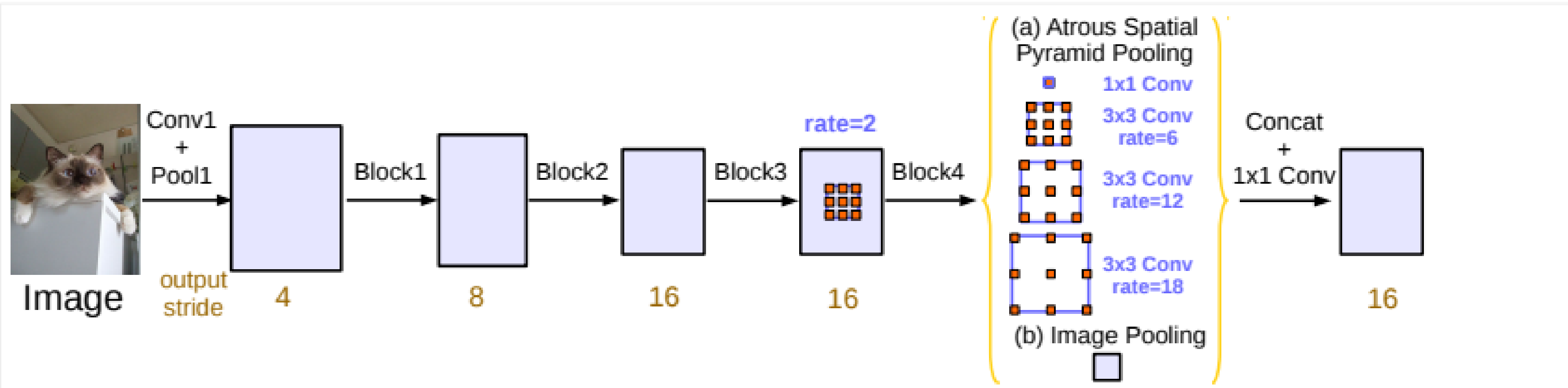
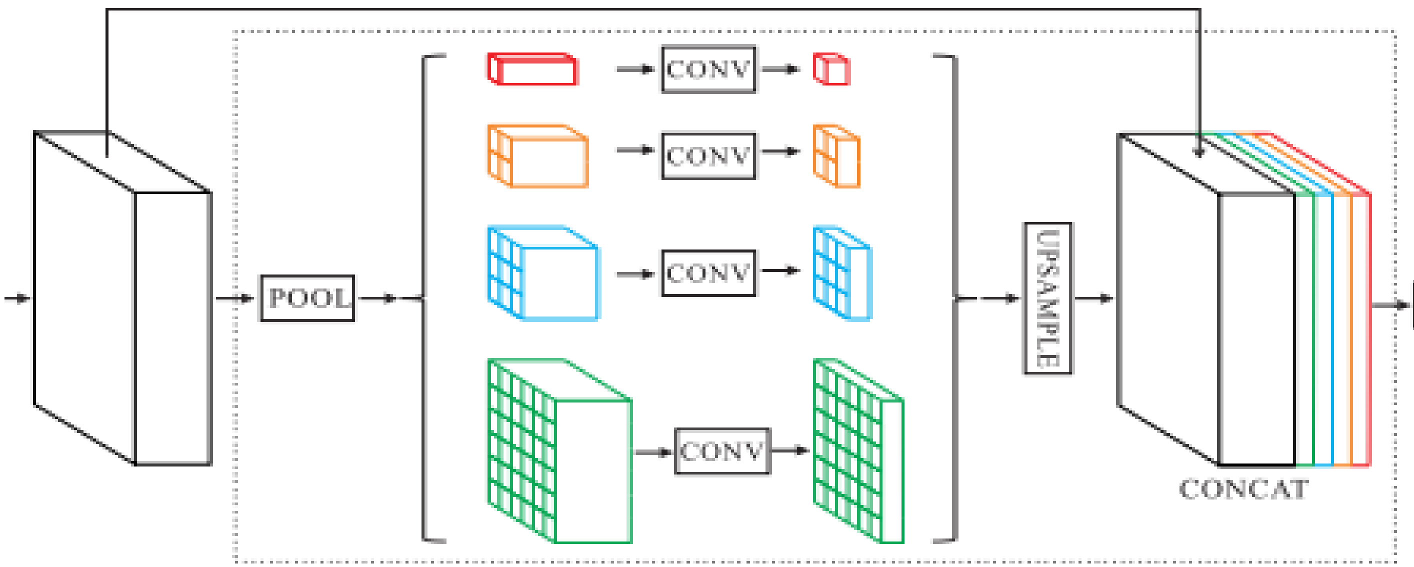


Figure 5. Parallel modules with atrous convolution (ASPP), augmented with image-level features.

Related work

3. Context information

- PSPNet(Pyramid Scene Parsing Network)



applies a PSP module which contains several different scales of average pooling layers.

Related work

4. Real-time segmentation

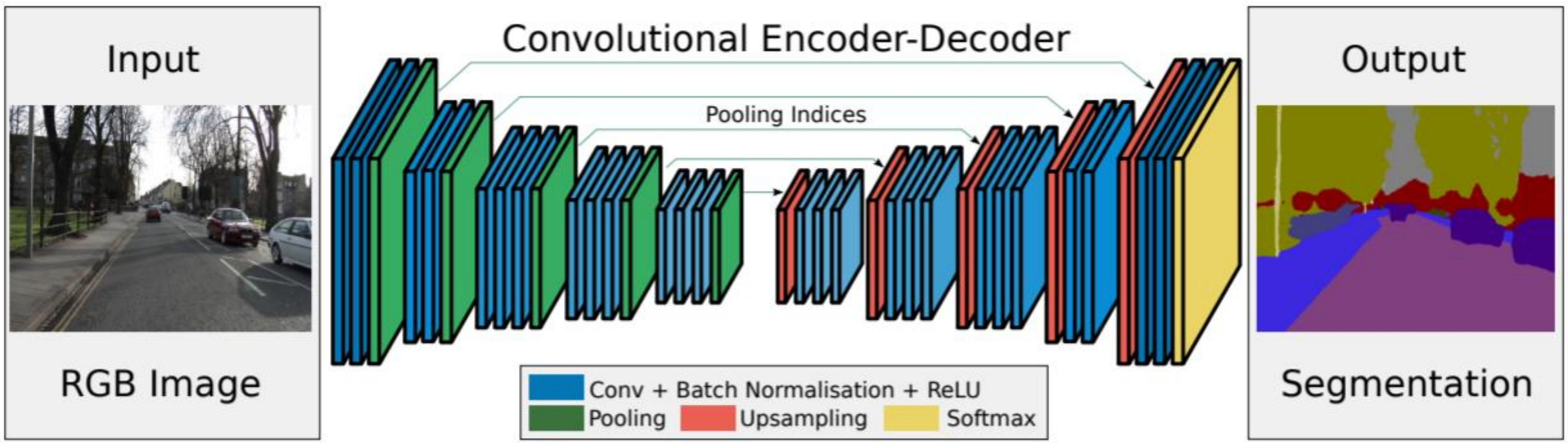
- **Segnet**

- **E-Net**

- **ICNet**

Related work

4. Real-time segmentation

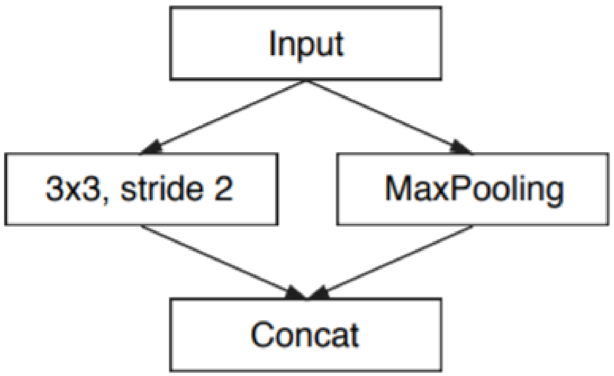


SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

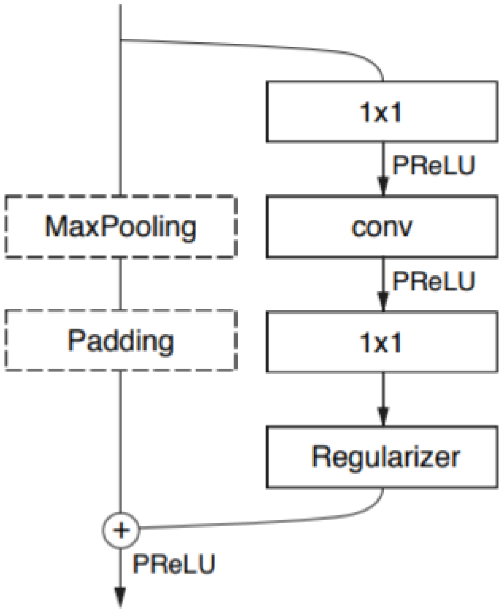
Segnet utilizes a small network structure and the skip-connected method to achieve a fast speed

Related work

4. Real-time segmentation



(a)



ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation

E-Net designs a lightweight network from scratch and delivers an extremely high speed.

Related work

4. Real-time segmentation

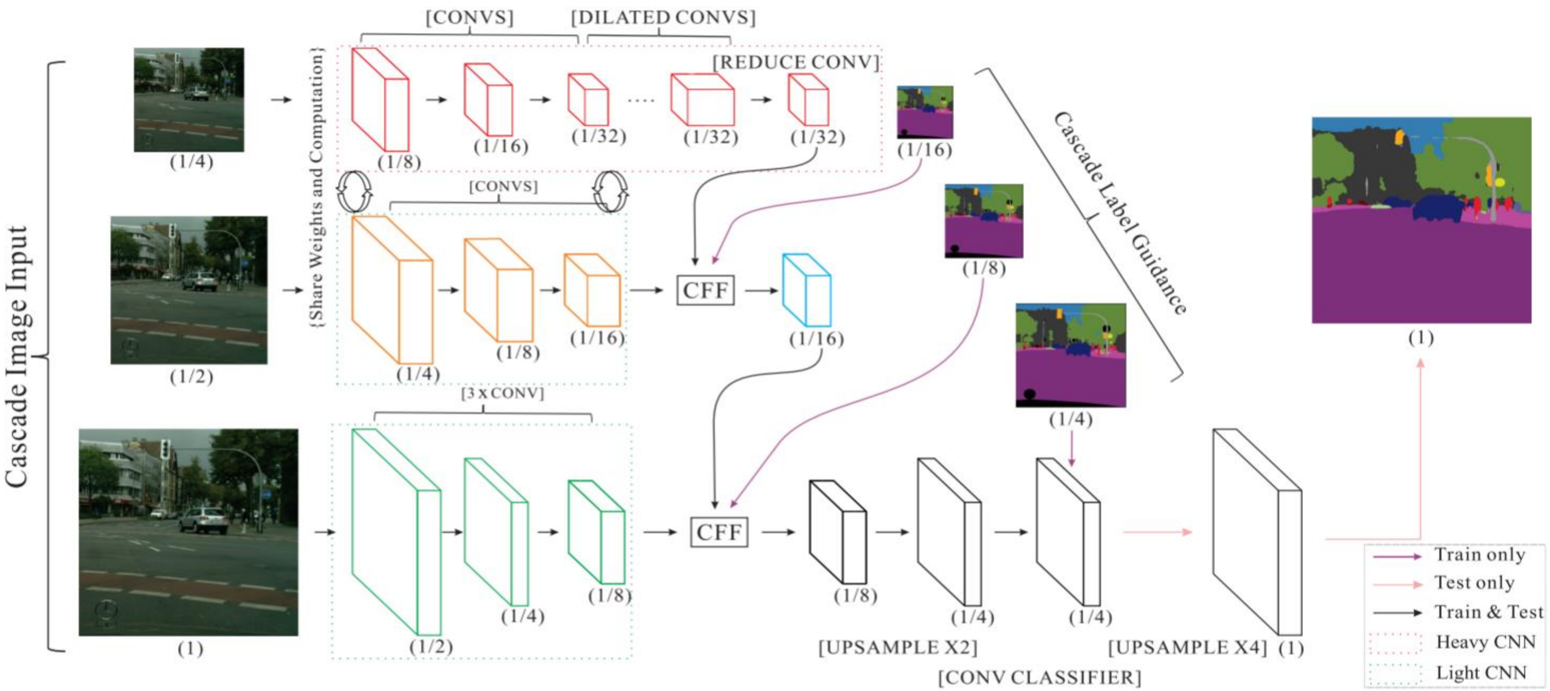
	GFLOPs	Parameters	Model size (fp16)
SegNet	286.03	29.46M	56.2 MB
ENet	3.83	0.37M	0.7 MB

ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation

E-Net designs a lightweight network from scratch and delivers an extremely high speed.

Related work

4. Real-time segmentation



ICNet for Real-Time Semantic Segmentation on High-Resolution Images

ICNet uses the image cascade to speed up the semantic segmentation method.

Related work

4. Real-time segmentation

Bisenet은 employs a lightweight model to provide sufficient receptive filed.

Furthermore, we set a shallow but wide network to capture adequate spatial information

Bilateral Segmentation Network

Semantic Segmentation 방법론

1) Spatial information preserving

2) Sufficient receptive field

Bilateral Segmentation Network

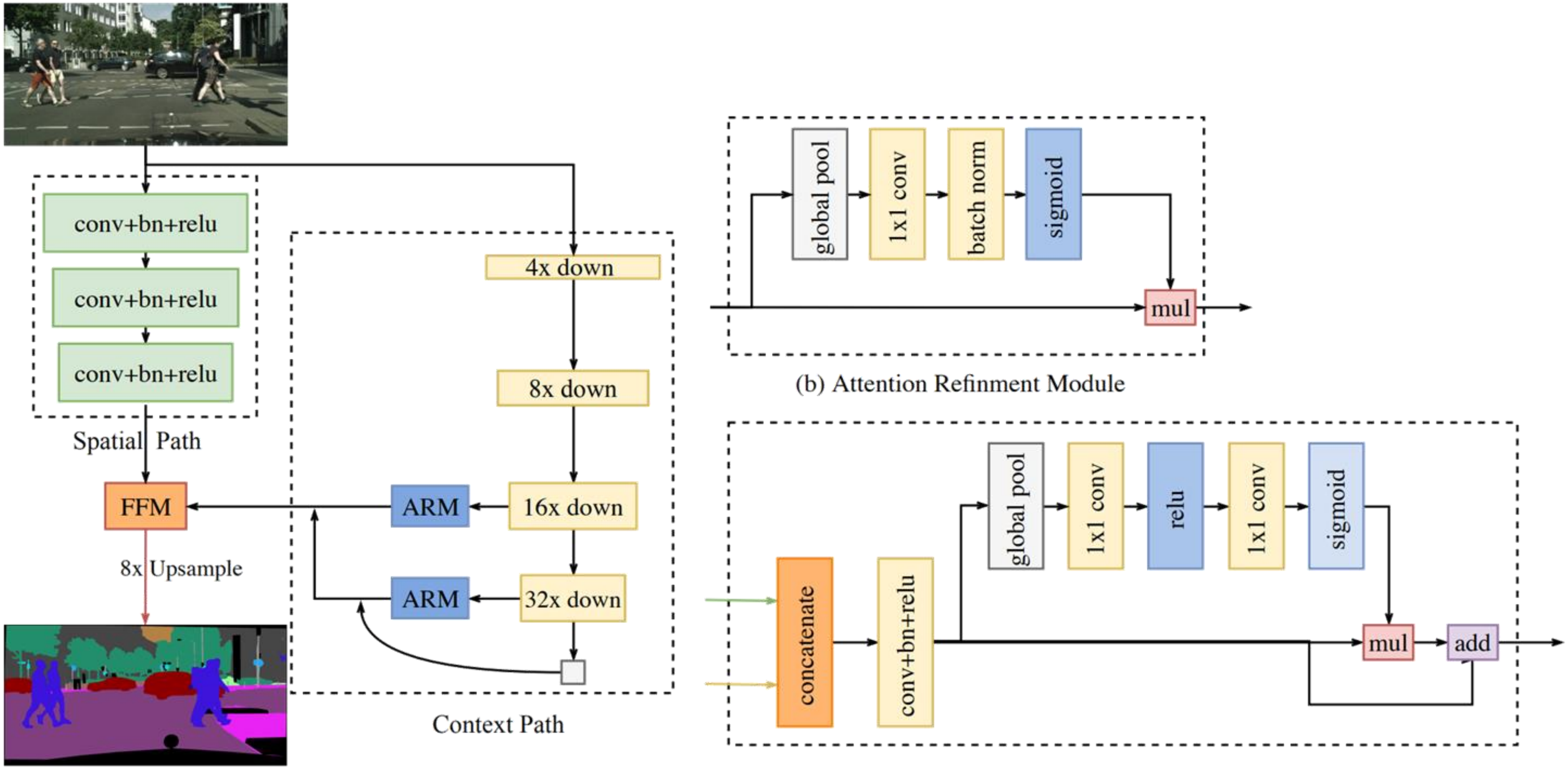
Semantic Segmentation 방법론

1) Spatial information preserving

2) Sufficient receptive field

→ But 두가지 동시에 달성하기는 너무 힘들다.

Bilateral Segmentation Network

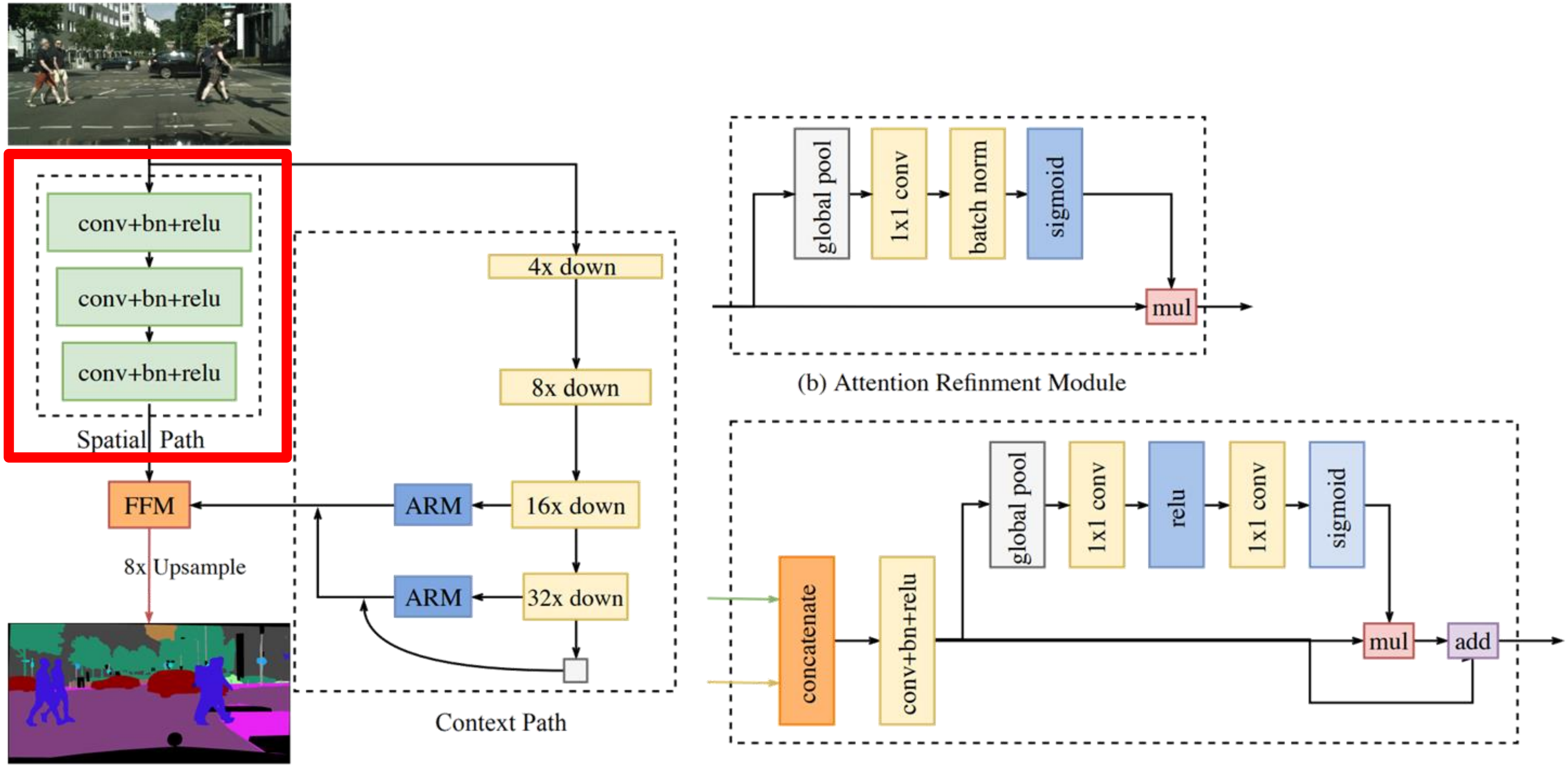


(a) Network Architecture

(b) Attention Refinment Module

(c) Feature Fusion Module

Bilateral Segmentation Network

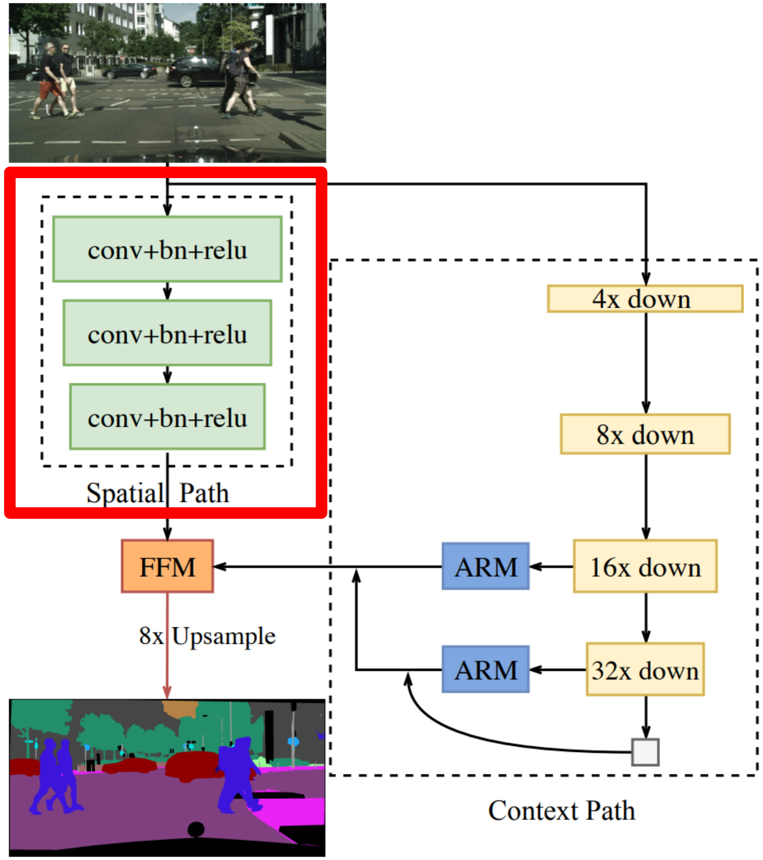


(a) Network Architecture

(b) Attention Refinment Module

(c) Feature Fusion Module

Bilateral Segmentation Network

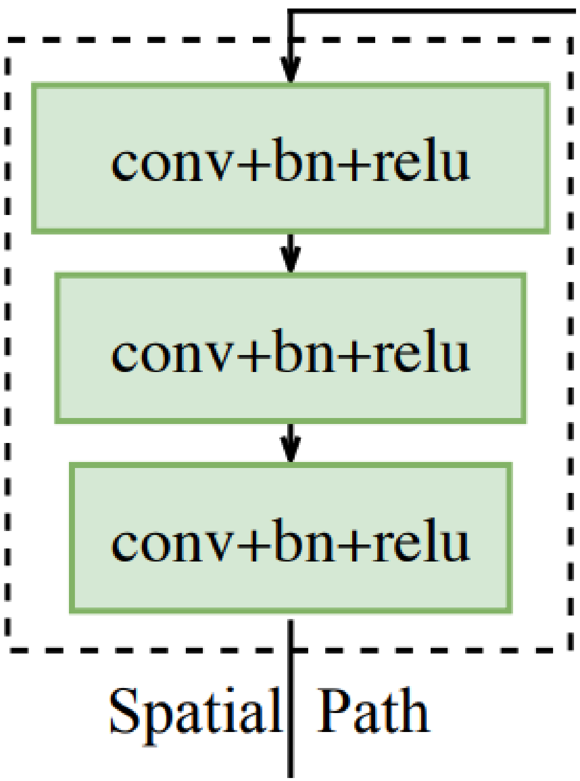


(a) Network Architecture

Designed to encode maximum spatial information

Bilateral Segmentation Network

1. Spatial Path



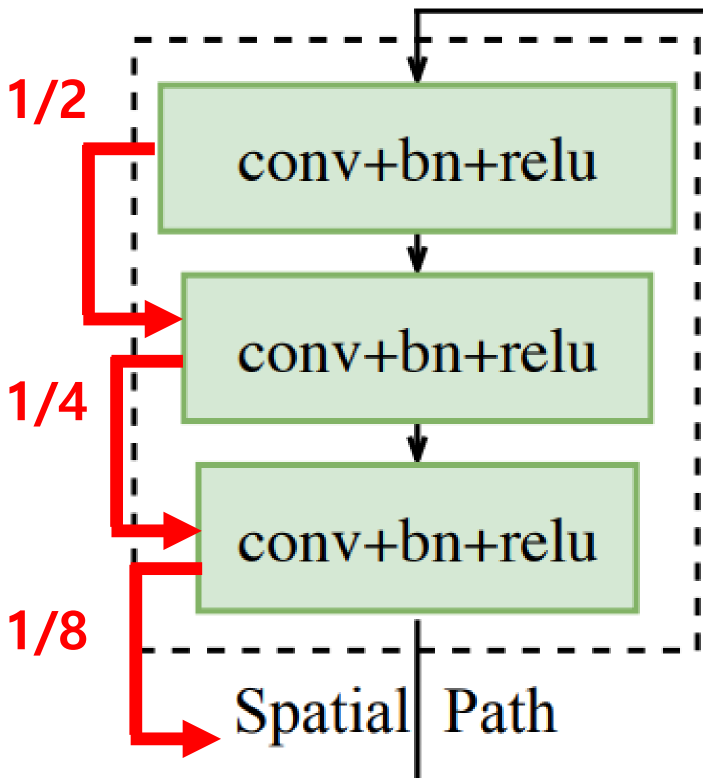
1) Spatial Path to preserve the spatial size of the original input image and encode affluent spatial information

2) this path extracts the output feature maps that is 1/8 of the original image

3) It encodes rich spatial information due to the large spatial size of feature maps.

Bilateral Segmentation Network

1. Spatial Path



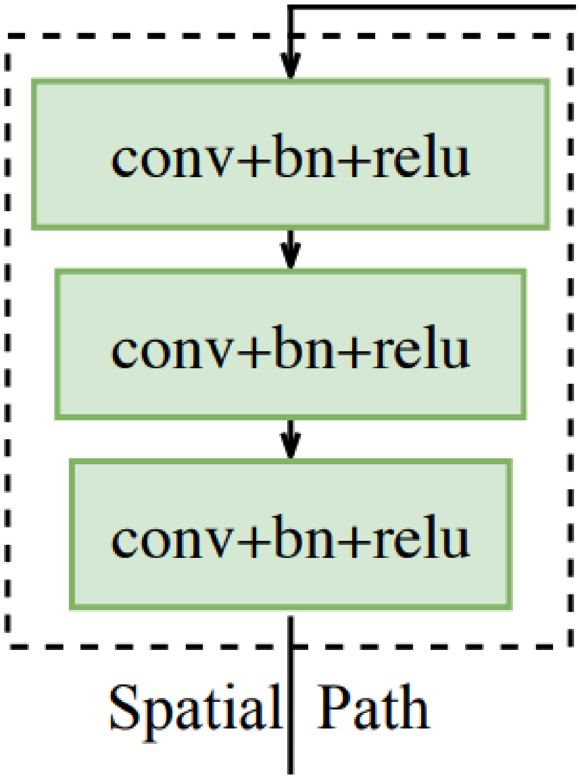
1) Spatial Path to preserve the spatial size of the original input image and encode affluent spatial information

2) this path extracts the output feature maps that is 1/8 of the original image

3) It encodes rich spatial information due to the large spatial size of feature maps.

Bilateral Segmentation Network

1. Spatial Path

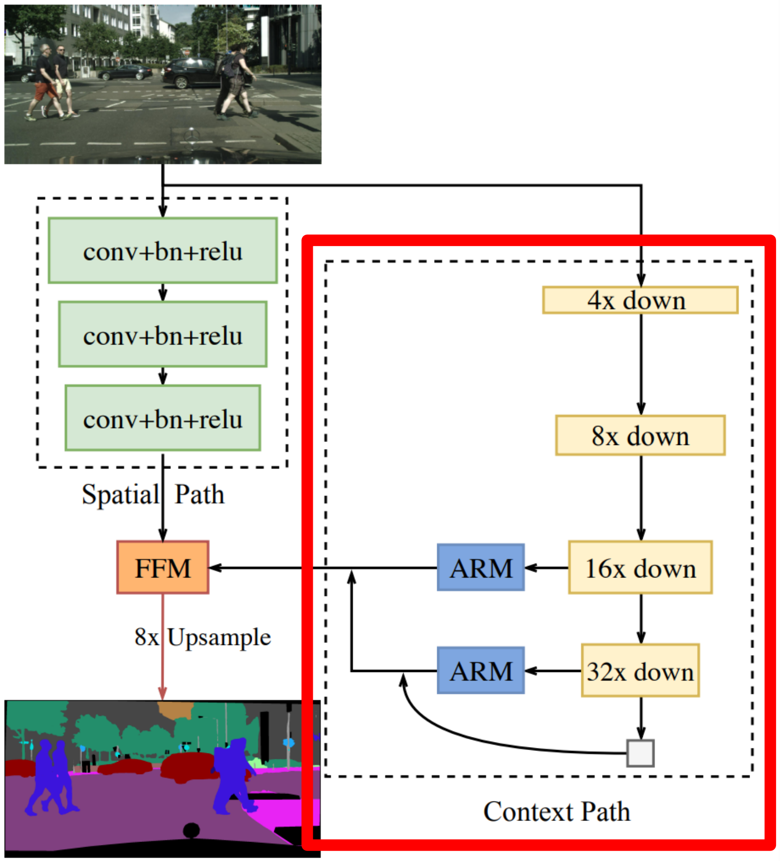


1) Spatial Path to preserve the spatial size of the original input image and encode affluent spatial information

2) this path extracts the output feature maps that is 1/8 of the original image

3) It encodes rich spatial information due to the large spatial size of feature maps.

Bilateral Segmentation Network



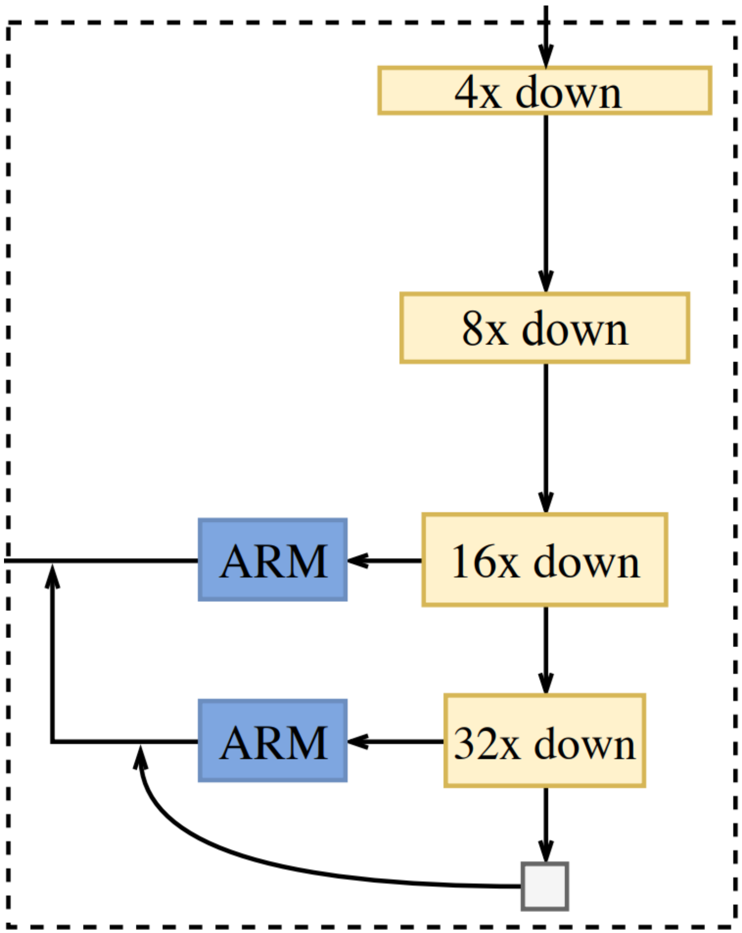
(a) Network Architecture

1) Designed to encode sufficient receptive field

2) Design in consideration of speed and computation

Bilateral Segmentation Network

2. Context Path



Context Path

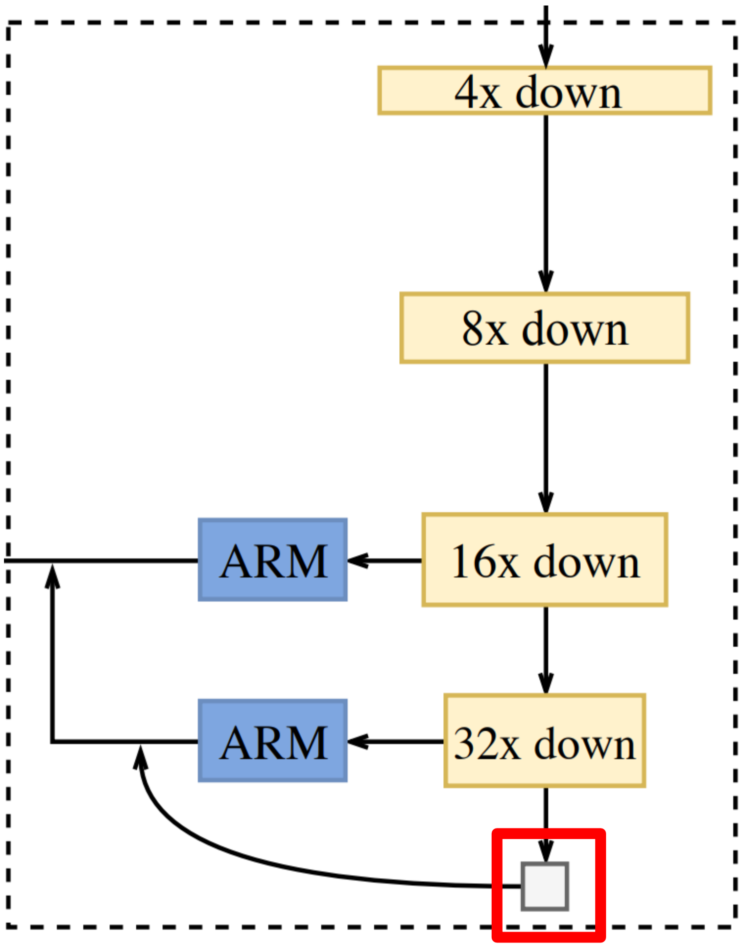
1) the lightweight model, like Xception, can down sample the feature map fast to obtain large receptive filed, which encodes high level semantic context information

2) Then we add a global average pooling on the tail of the lightweight model, which can provide the maximum receptive field with global context information

3) Finally, we combine the up-sampled output feature of global pooling and the features of the lightweight model.

Bilateral Segmentation Network

2. Context Path



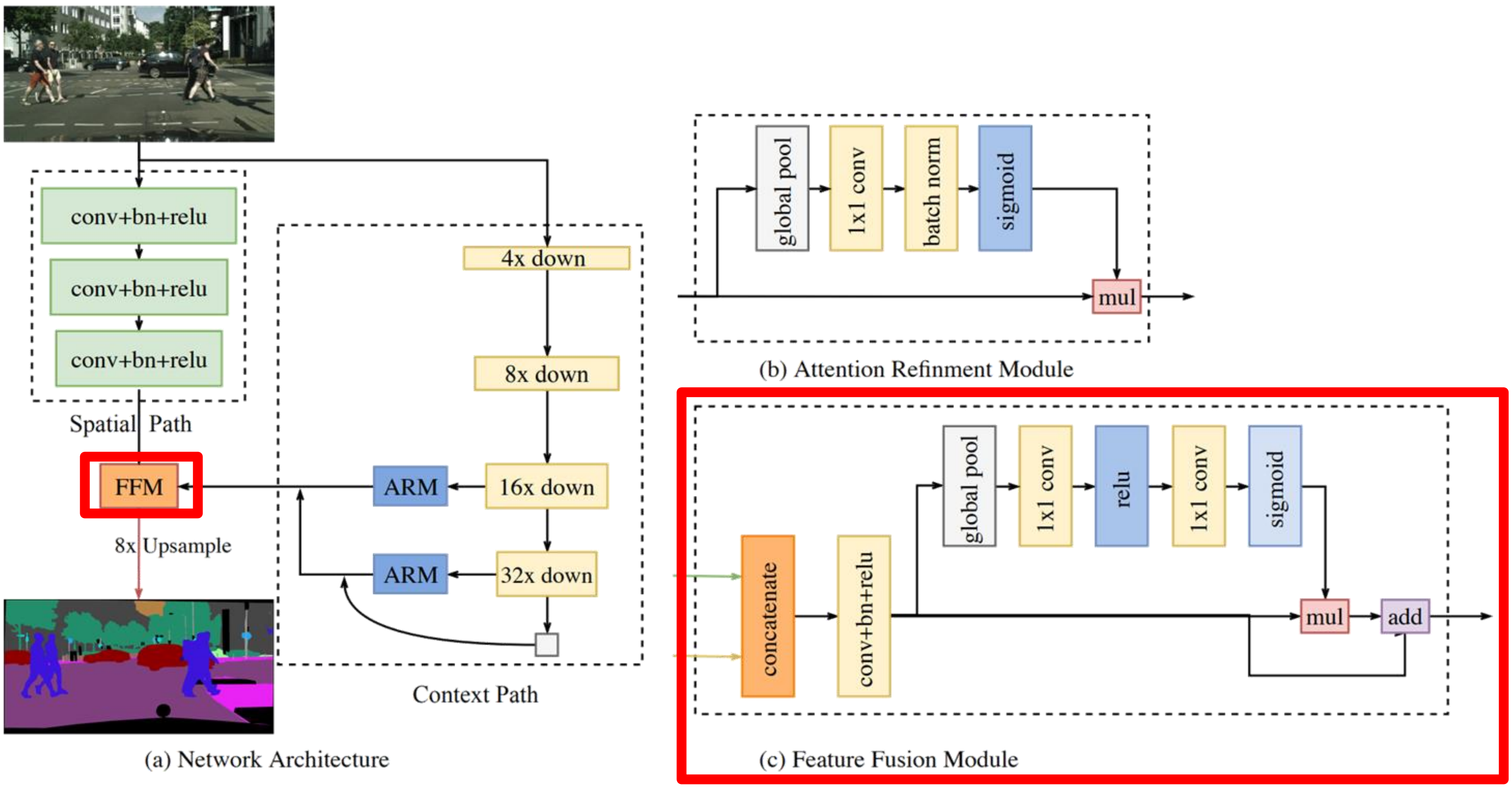
Context Path

1) the lightweight model, like Xception, can down sample the feature map fast to obtain large receptive filed, which encodes high level semantic context information

2) Then we add a global average pooling on the tail of the lightweight model, which can provide the maximum receptive field with global context information

3) Finally, we combine the up-sampled output feature of global pooling and the features of the lightweight model.

Bilateral Segmentation Network

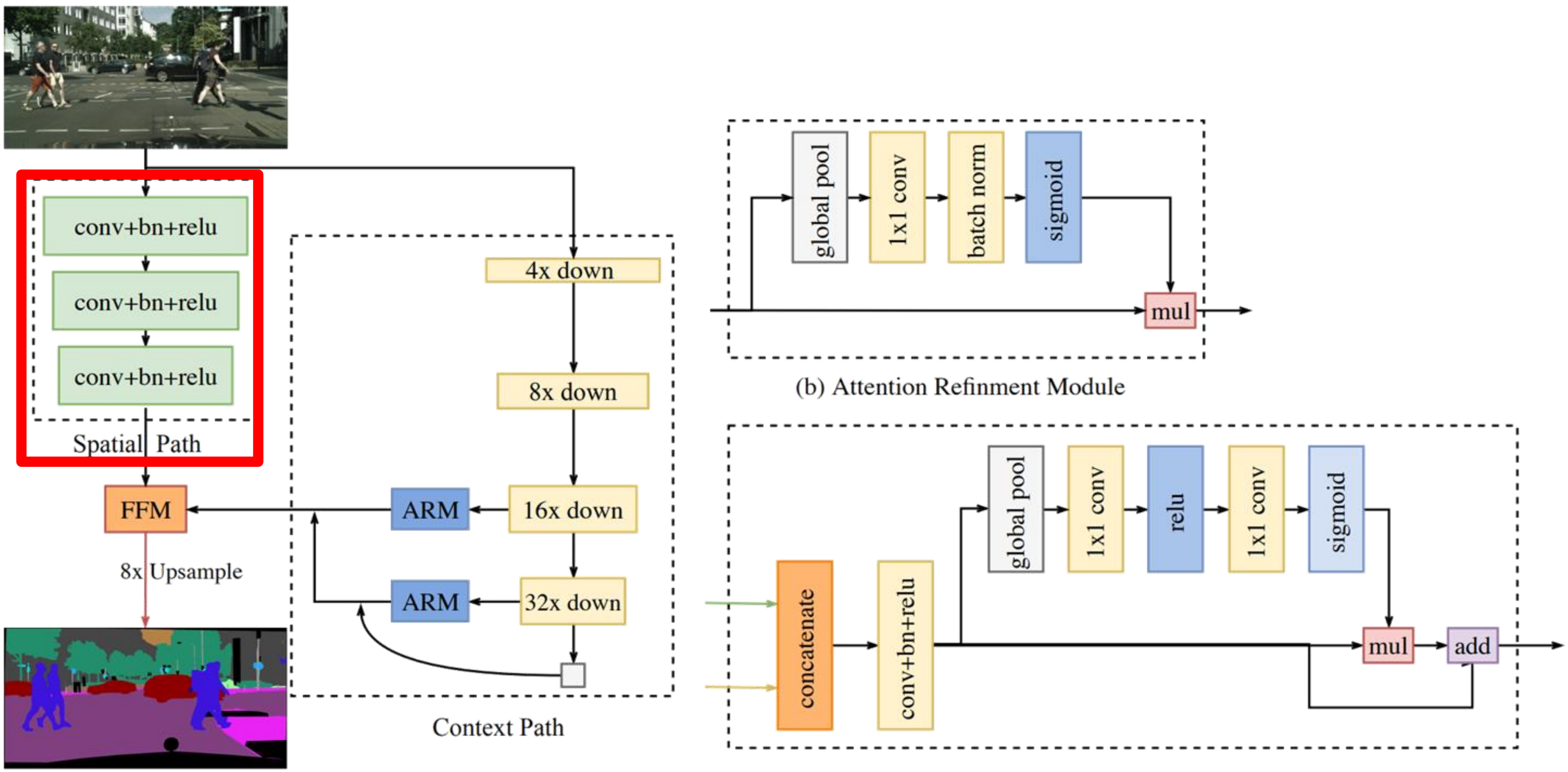


(a) Network Architecture

(b) Attention Refinment Module

(c) Feature Fusion Module

Bilateral Segmentation Network

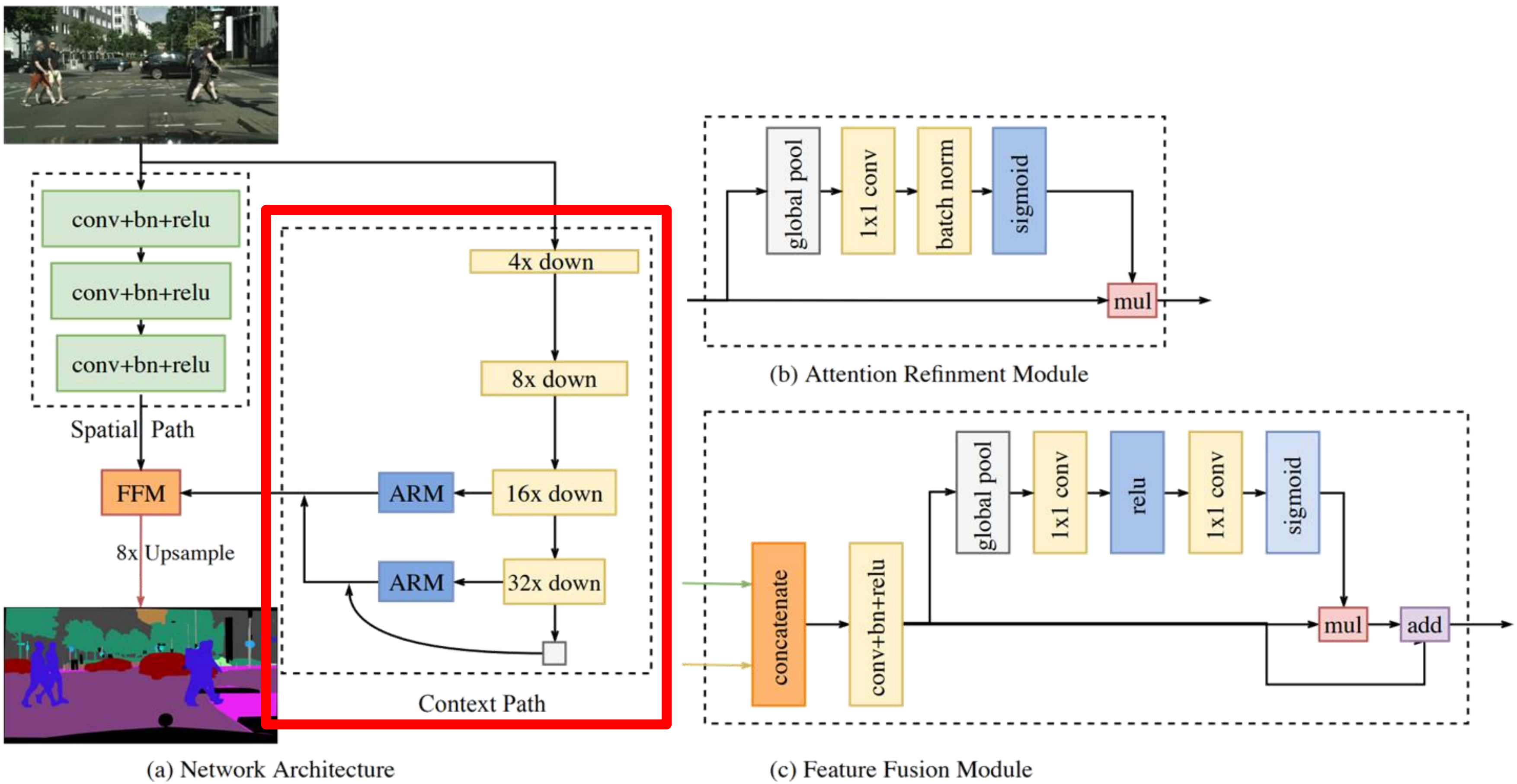


(a) Network Architecture

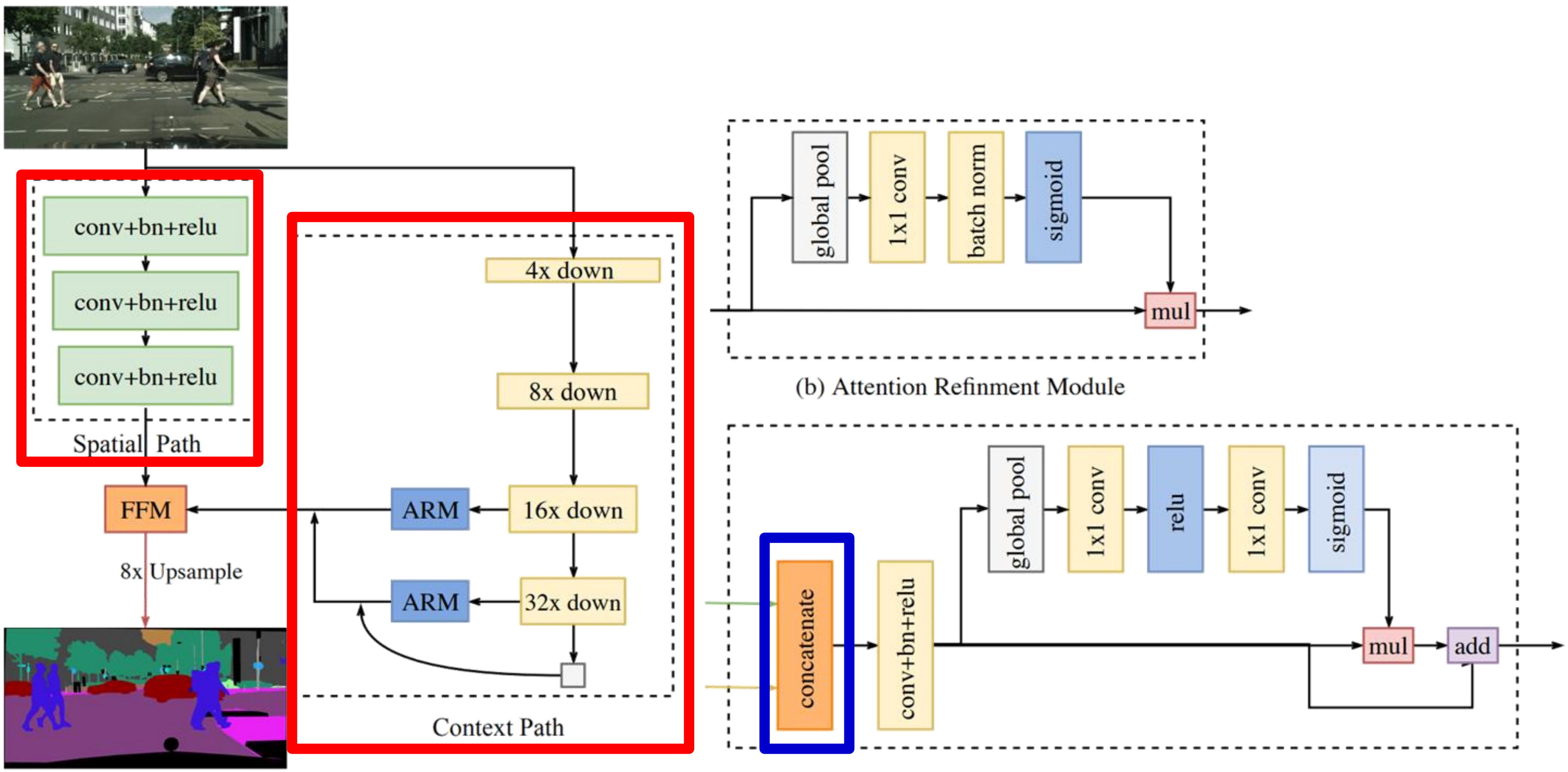
(b) Attention Refinment Module

(c) Feature Fusion Module

Bilateral Segmentation Network



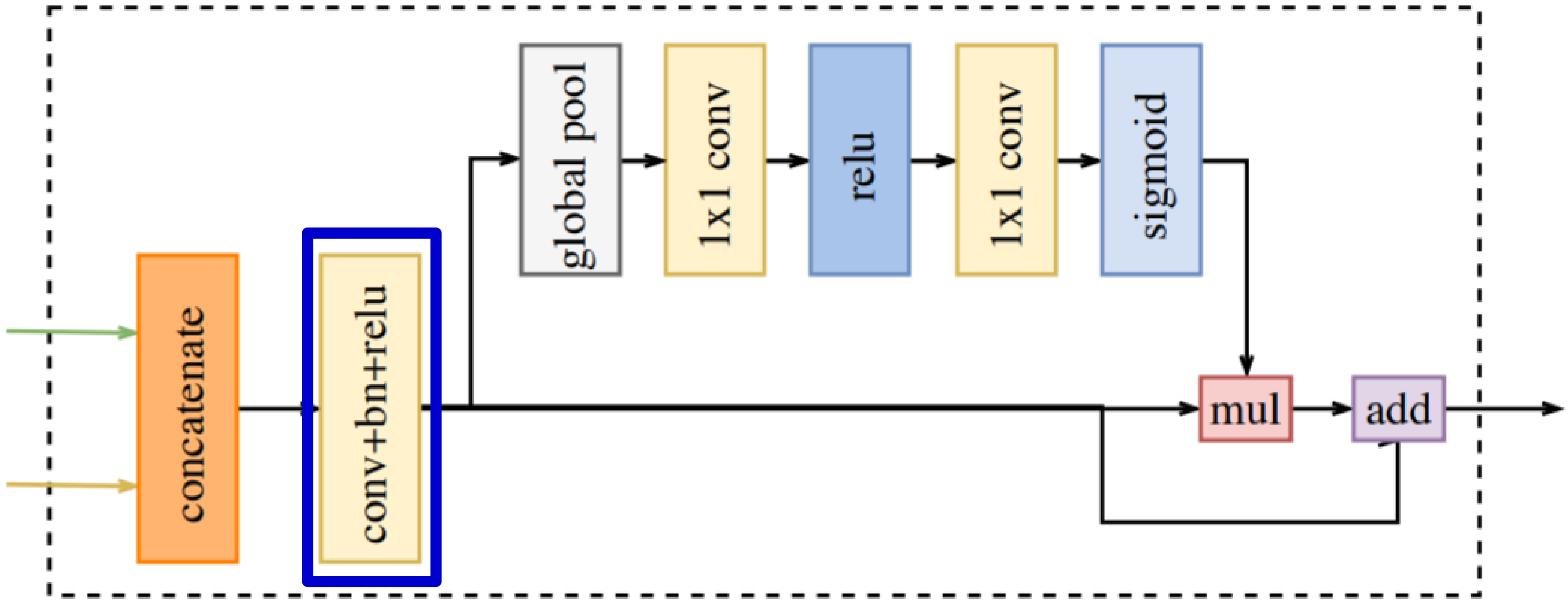
Bilateral Segmentation Network



(a) Network Architecture

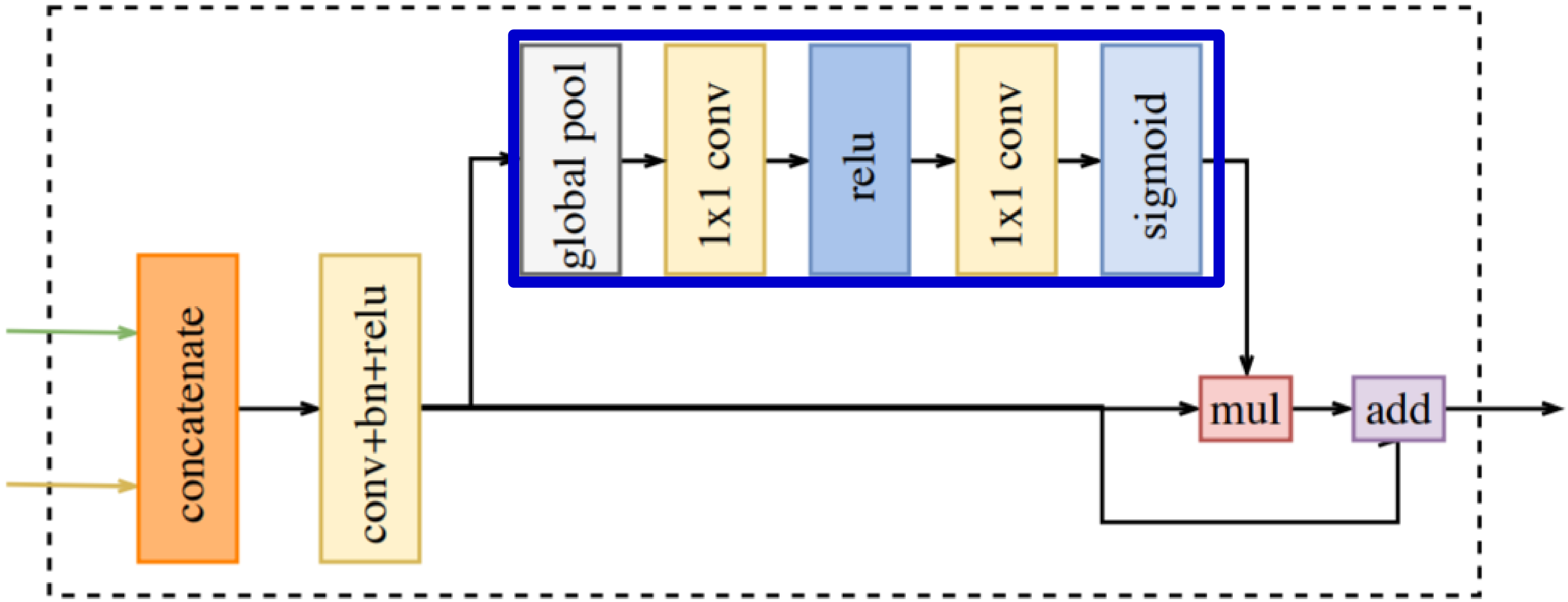
(c) Feature Fusion Module

Bilateral Segmentation Network



(c) Feature Fusion Module

Bilateral Segmentation Network



(c) Feature Fusion Module

This weight vector can re-weight the features, which amounts to feature selection and combination.

Bilateral Segmentation Network

3. Network architecture

$$\boxed{loss = \frac{1}{N}\sum_{i} L_i = \frac{1}{N}\sum_{i} -log\left(\frac{e^{p_i}}{\sum_j e^{p_j}}\right)}$$

$$L(X;W) = l_p(X;W) + \alpha \sum_{i=2}^{K} l_i(X_i;W)$$

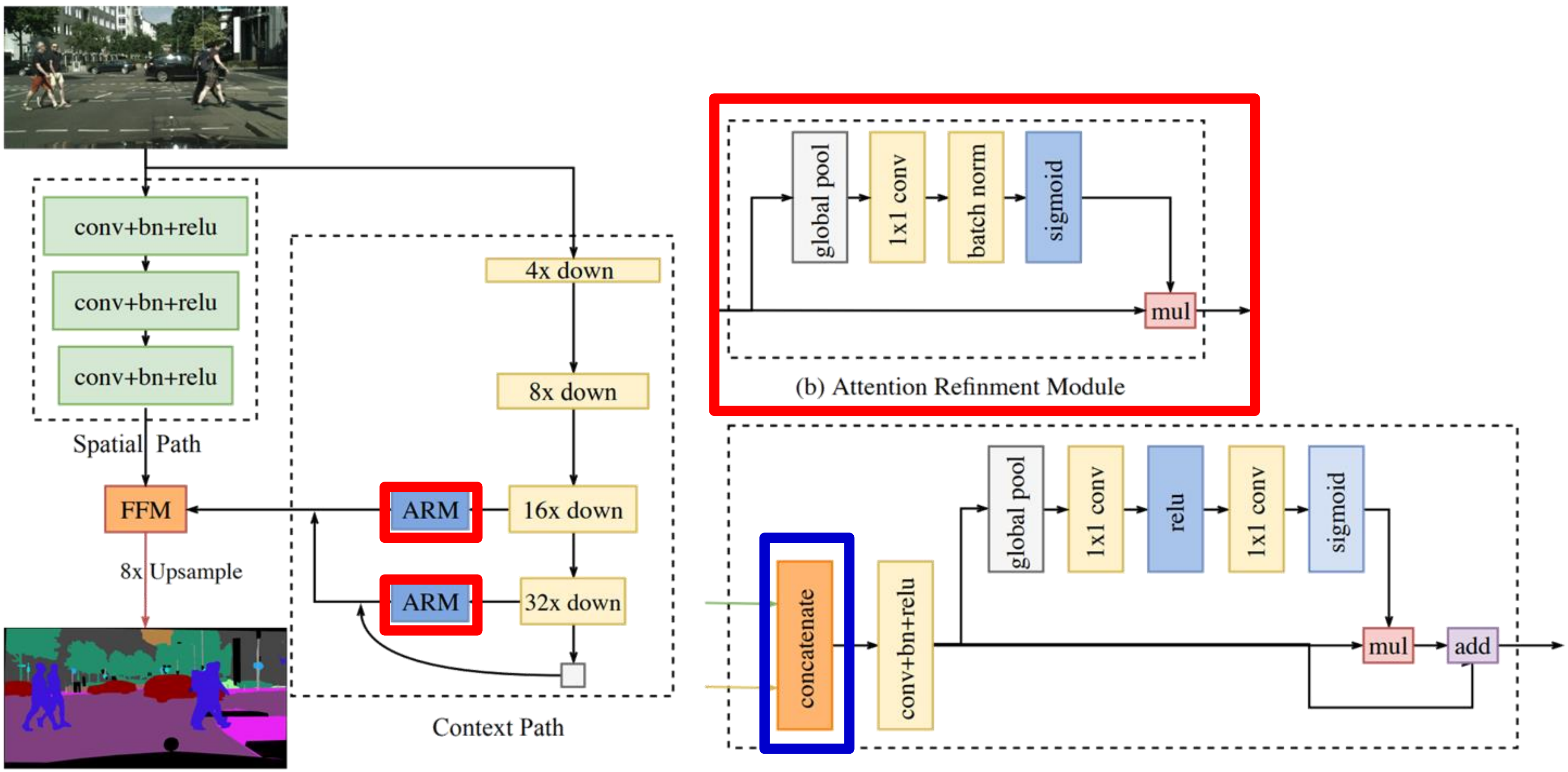
Bilateral Segmentation Network

3. Network architecture

$$loss = \frac{1}{N}\sum_{i} L_i = \frac{1}{N}\sum_{i} -log\left(\frac{e^{p_i}}{\sum_j e^{p_j}}\right)$$

$$\boxed{L(X;W) = l_p(X;W) + \alpha \sum_{i=2}^{K} l_i(X_i;W)}$$

Bilateral Segmentation Network



(a) Network Architecture

(b) Attention Refinment Module

(c) Feature Fusion Module

Bilateral Segmentation Network

Experimental Result

- Modified Xception model, Xception39, into the real-time semantic segmentation task.

- Dataset

1) Cityscapes

2) CamVid

3) COCO Stuff

Experimental Result

Total 5,000, 2,975 for training, 500 for validation data and 1,525 test data resolution 2,048x1,024, 19 classes

Experimental Result

2) CamVid

Total 701, 367 for training, 101 for validation and 233 for testing. Resolution 960x720 and 11 semantic category

Experimental Result

3) COCO-Stuff

Total 164,000, 118,000 for training, 5,000 for validation, 20,000 for test and 20,000 for test-challenge. It covers 91 stuff classes and 1 class 'unlabeled'

Bilateral Segmentation Network

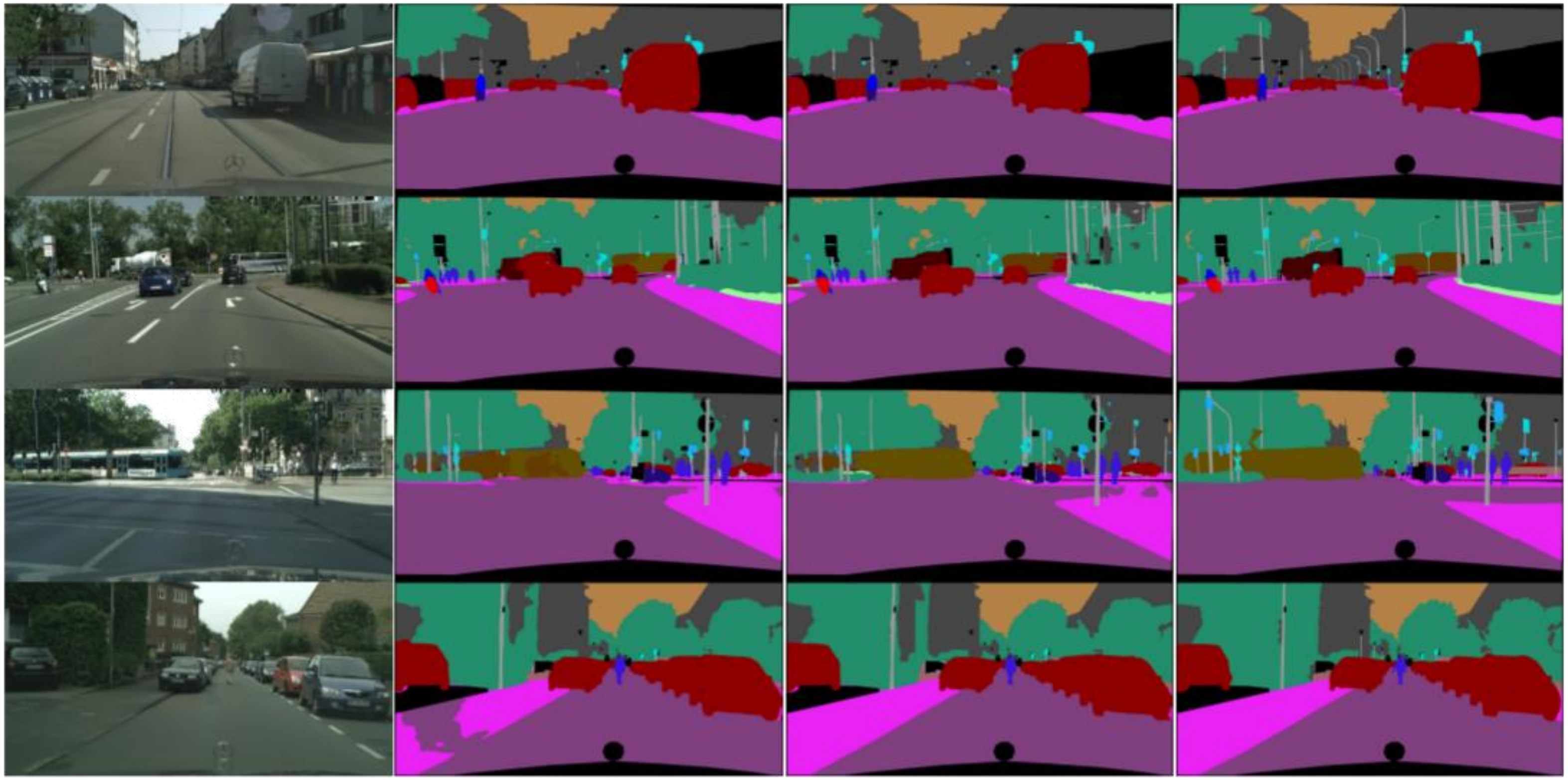
Experimental Result – Implementation protocol

- **Mini-batch stochastic gradient descent(SGD) with batch size 16, momentum 0.9 weight decay 1e-4 in training. We apply the 'poly' learning rate strategy in which the initial rate is multiplied by (1-iter/max_iter)power each iteration with power 0.9. the initial learning rate is 2.5e-2.**

- **Data augmentation : we employ the mean subtraction, random horizontal flip and random scale on the input images to augment the dataset in training process. The scales contains {0.75~2.0). Finally, we randomly crop the image into fix size for training**

Bilateral Segmentation Network

Experimental Result – Ablation Study



(a) Image (b) U-Shape (c) BiSeNet (d) GT

Bilateral Segmentation Network

Experimental Result – Ablation Study

Method	Mean IOU(%)
CP	66.01
CP+SP(Sum)	66.82
CP+SP(FFM)	67.42
CP+SP(FFM)+GP	68.42
CP+SP(FFM)+ARM	68.72
CP+SP(FFM)+GP+ARM	71.40

Bilateral Segmentation Network

Experimental Result – Ablation Study(Spatial Path)

Method	Mean IOU(%)
CP	66.01
CP+SP(Sum)	66.82
CP+SP(FFM)	67.42
CP+SP(FFM)+GP	68.42
CP+SP(FFM)+ARM	68.72
CP+SP(FFM)+GP+ARM	71.40

The Spatial Path encodes abundant details of spatial information

Bilateral Segmentation Network

Experimental Result – Ablation Study(ARM)

Method	Mean IOU(%)
CP	66.01
CP+SP(Sum)	66.82
CP+SP(FFM)	67.42
CP+SP(FFM)+GP	68.42
CP+SP(FFM)+$\boxed{\text{ARM}}$	$\boxed{68.72}$
CP+SP(FFM)+GP+ARM	71.40

For the original feature, it is easy to capture the global context information without the complex up-sample operation.

Bilateral Segmentation Network

Experimental Result – Ablation Study(FFM)

Method	Mean IOU(%)
CP	66.01
CP+SP(Sum)	66.82
CP+SP(FFM)	67.42
CP+SP(FFM)+GP	68.42
CP+SP(FFM)+ARM	68.72
CP+SP(FFM)+GP+ARM	71.40

Bilateral Segmentation Network

Experimental Result – Ablation Study(GP)

Method	Mean IOU(%)
CP	66.01
CP+SP(Sum)	66.82
CP+SP(FFM)	67.42
CP+SP(FFM)+GP	68.42
CP+SP(FFM)+ARM	68.72
CP+SP(FFM)+GP+ARM	71.40

Bilateral Segmentation Network

Experimental Result – Speed and Accuracy Analysis

Method	BaseModel	GFLOPS	Parameters
SegNet [1]	VGG16 [29]	286.0	29.5M
ENet [25]	From scratch	3.8	0.4M
Ours	Xception39	2.9	5.8M
Ours	Res18	10.8	49.0M

Bilateral Segmentation Network

Experimental Result – Speed and Accuracy Analysis

Method	NVIDIA Titan X						NVIDIA Titan XP					
	640×360		1280×720		1920×1080		640×360		1280×720		1920×1080	
	ms	fps	ms	fps	ms	fps	ms	fps	ms	fps	ms	fps
SegNet [1]	69	14.6	289	3.5	637	1.6	-	-	-	-	-	-
ENet [25]	7	135.4	21	46.8	46	21.6	-	-	-	-	-	-
Ours[1]	**5**	**203.5**	**12**	**82.3**	**24**	**41.4**	**4**	**285.2**	**8**	**124.1**	**18**	**57.3**
Ours[2]	8	129.4	21	47.9	43	23	5	205.7	13	78.8	29	34.4

Bilateral Segmentation Network

Experimental Result – Speed and Accuracy Analysis

Method	BaseModel	Mean IOU(%)		FPS
		val	*test*	
SegNet [1]	VGG16	-	56.1	-
ENet [25]	From scratch	-	58.3	-
SQ [30]	SqueezeNet [14]	-	59.8	-
ICNet [39]	PSPNet50 [40]	67.7	69.5	30.3
DLC [17]	Inception-ResNet-v2	-	71.1	-
Two-column Net [34]	Res50	74.6	72.9	14.7
Ours	Xception39	69.0	68.4	**105.8**
Ours	Res18	**74.8**	**74.7**	65.5

Conclusions

- BiSeNet is proposed in this paper to improve the speed and accuracy of real-time semantic segmentation simultaneously

- Our proposed BiSeNet contains two paths : Spatial Path(SP) and Context Path(CP).

- SP is designed to preserve the spatial information from original images.

- CP utilizes the lightweight model and GP to obtain sizeable receptive filed rapidly

- With the affluent spatial details and large receptive field, we achieve the result of 68.4% mean IOU on Cityscapes test dataset at 105FPS

End