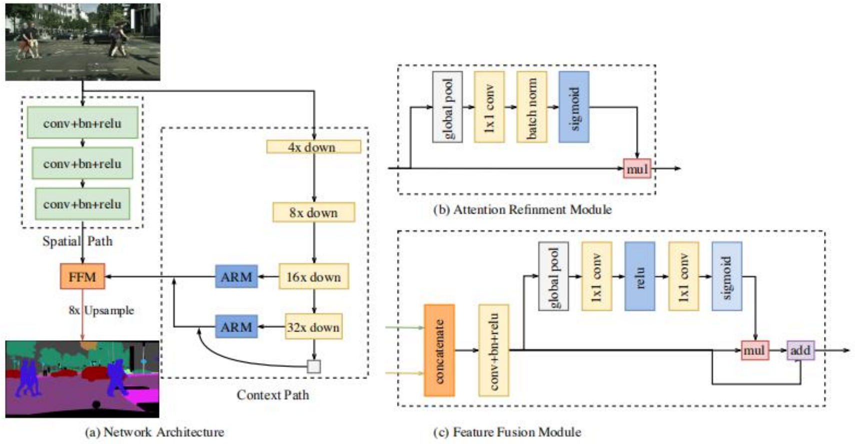
Rethinking BiSeNet For Real-time Semantic Segmentation

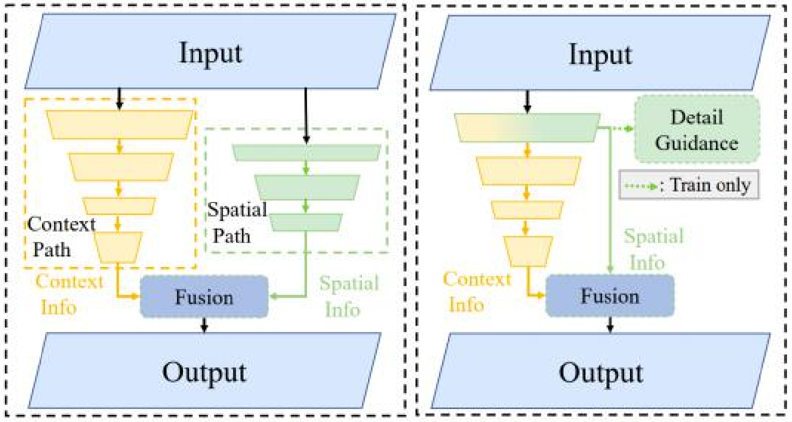
——from CVPR2021 by Meituan

Presenter: Gong, Qiqi

@BJTU

Reviewing Bilateral Segmentation Network(**BiSeNet**)



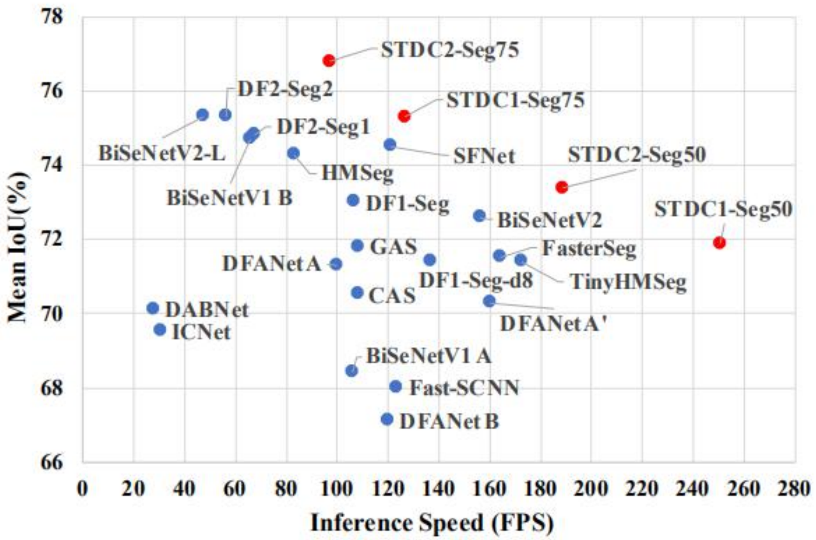


(a) BiSeNet (b) Ours

- Real-time Segmentation
- Two paths:
  - Spatial Path
    - Keep resolution
    - Three conv. blocks
  - Context Path
    - Provide sufficient receptive fields
    - Backbones from pretrained classification network
- Contribution
  - Overcoming corns from restricted input size (multi-scale objects)

Motivation & Contribution

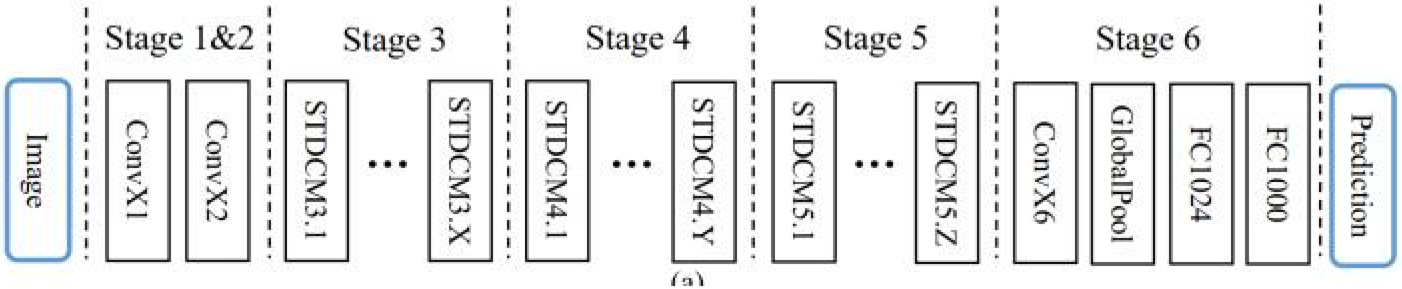
- Motivation
  - Adding an extra path to encode spatial information is time-consuming
  - Backbones from pretrained tasks could be insufficient for seg. task
  - Argue: auxiliary path lacks low-level info. guidance
- Contribution
  - Short-Term Dense Concatenate (STDC) module



Overview

- Encoding Stage
  - STDC Module: preserves scalable receptive fields and multi-scal info.
  - STDC Network: Integrate STDC modules into U-Net architecture
- Decoding Stage
  - Use **Detail Aggregation Module** to generate **binary detail GT** from seg. GT
  - Adopt binary CE loss and dice loss jointly to **optimize training process**
  - Fuse low-level info. to obtain results

Encoding Network



- Network Overview
  - Stage1 & 2: Only one conv. block (conv. layer + BN +ReLU)
  - Stage 3-5
    - STDCMk.1: Stride=2 (size*0.5)
    - Following: Stride=1 (resolution unchanged)
  - Output Channels: 1024

Stages	Output size	KSize	S	STDC1		STDC2	
				R	C	R	C
Image	224×224				3		3
ConvX1	112×112	3×3	2	1	32	1	32
ConvX2	56×56	3×3	2	1	64	1	64
Stage3	28×28		2	1	256	1	256
	28×28		1	1		3	
Stage4	14×14		2	1	512	1	512
	14×14		1	1		4	
Stage5	7×7		2	1	1024	1	1024
	7×7		1	1		2	
ConvX6	7×7	1×1	1	1	1024	1	1024
GlobalPool	1×1	7×7					
FC1					1024		1024
FC2					1000		1000
FLOPs				813M		1446M	
Params				8.44M		12.47M	

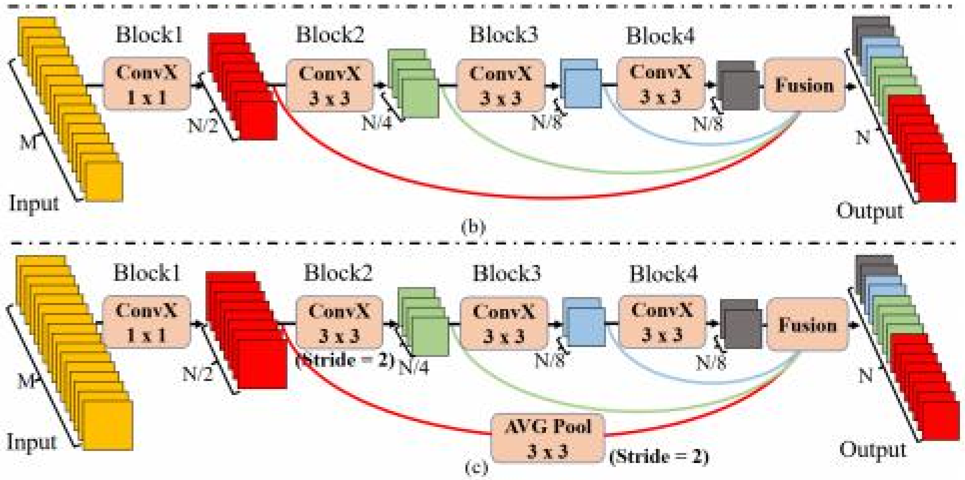
Table 2. Detailed architecture of STDC networks. Note that *ConvX* shown in the table refers to the Conv-BN-ReLU. The basic module of Stage 3, 4 and 5 is STDC module. KSize mean kernel size. S, R, C denote stride, repeat times and output channels respectively.

Encoding Network

- STDC Module
  - KS of ConvX1 = 1, following : 3
  - Output channel of each block: N/2^i

  (for the last: same as former's)
  - Computation complexity (rely on M&N)

$$
\begin{aligned}
S_{param} &= M \times 1 \times 1 \times \frac{N}{2^1} + \sum_{i=2}^{n-1} \frac{N}{2^{i-1}} \times 3 \times 3 \times \frac{N}{2^i} + \\
&\frac{N}{2^{n-1}} \times 3 \times 3 \times \frac{N}{2^{n-1}} \\
&= \frac{NM}{2} + \frac{9N^2}{2^3} \times \sum_{i=0}^{n-3} \frac{1}{2^{2i}} + \frac{9N^2}{2^{2n-2}} \\
&= \frac{NM}{2} + \frac{3N^2}{2} \times \left(1 + \frac{1}{2^{2n-3}}\right) \qquad (3)
\end{aligned}
$$

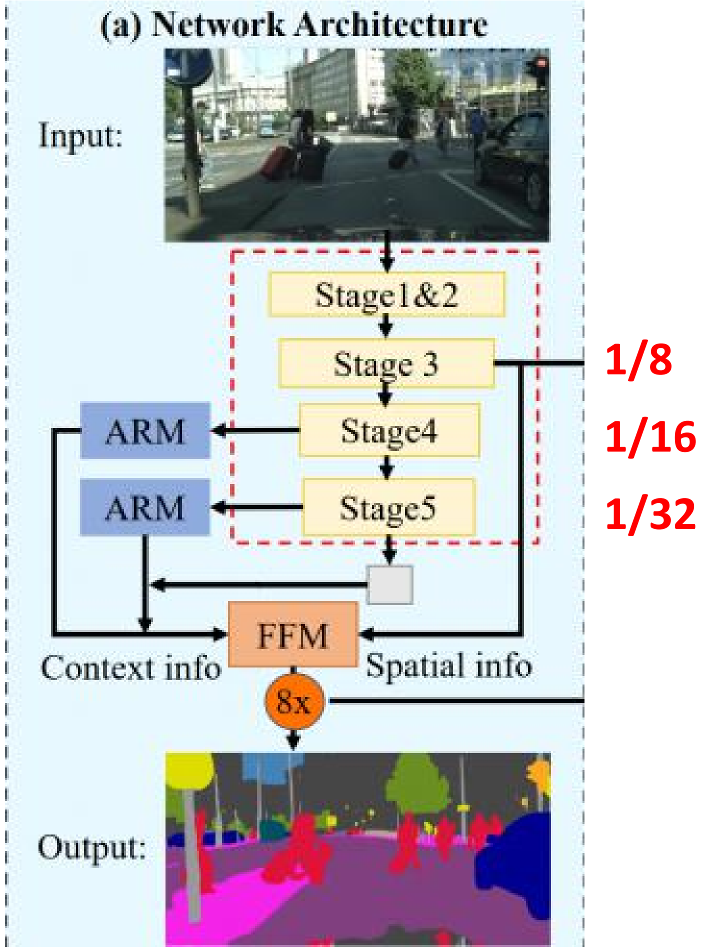


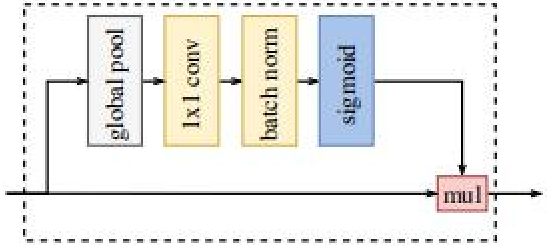
STDC module	Block1	Block2	Block3	Block4	Fusion
RF(S = 1)	1×1	3×3	5×5	7×7	$1 \times 1, 3 \times 3$ $5 \times 5, 7 \times 7$
RF(S = 2)	1×1	3×3	7×7	11×11	3×3 $7 \times 7, 11 \times 11$

Table 1. Receptive Field of blocks in our STDC module. *RF* denotes Receptive Field, *S* means stride, Note that if stride=2, the 1×1 *RF* of Block1 is turned into 3×3 *RF* by Average Pool operation.

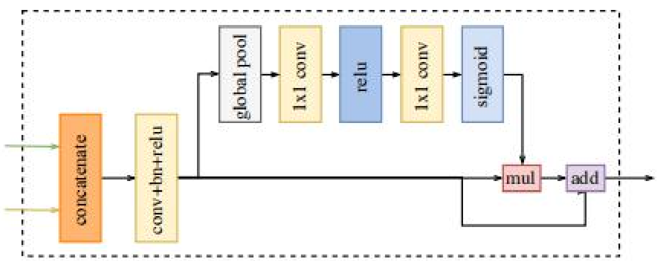
Decoder

- Segmentation Architecture
  - Backbone: pretrained STDC networks
  - Advantage:
    - Feature from backbone preserves rich detail information
    - Feature from decoder contains context information (due to input from global pooling layer)





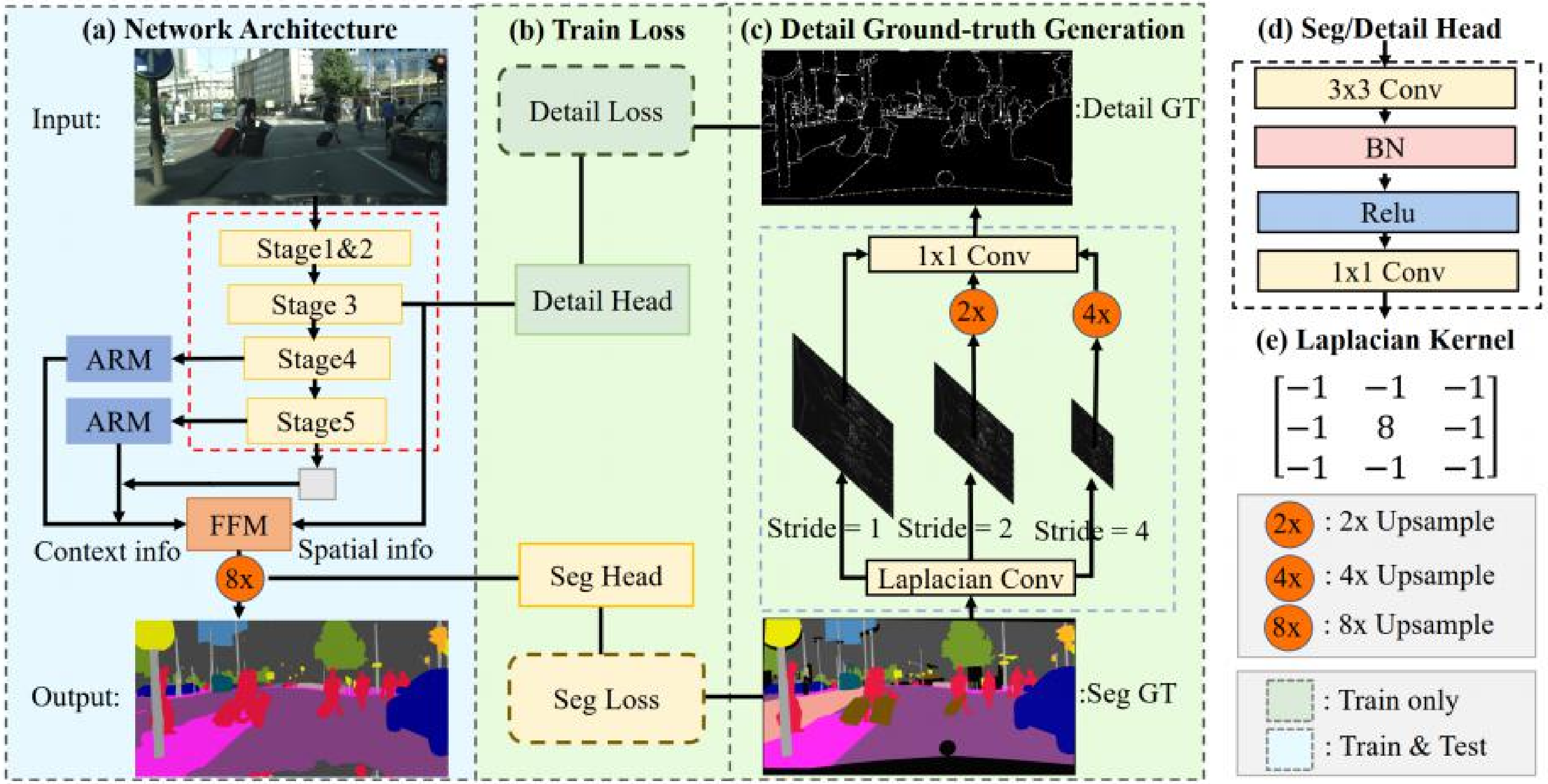
(b) Attention Refinment Module



(c) Feature Fusion Module

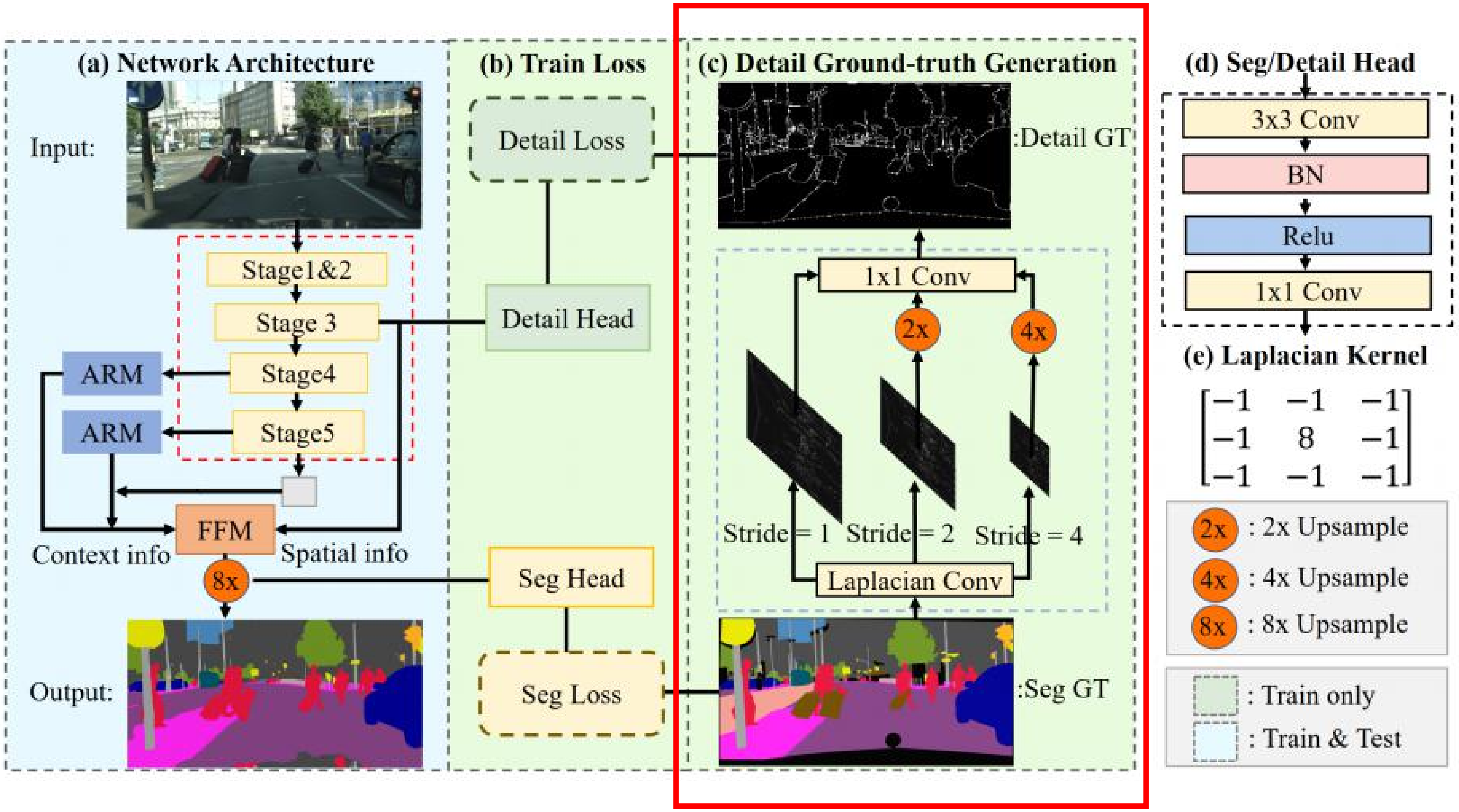
Decoder

- Detail Guidance
  - Treat detail prediction as a binary seg. task
  - Introduce detail loss to optimize training process



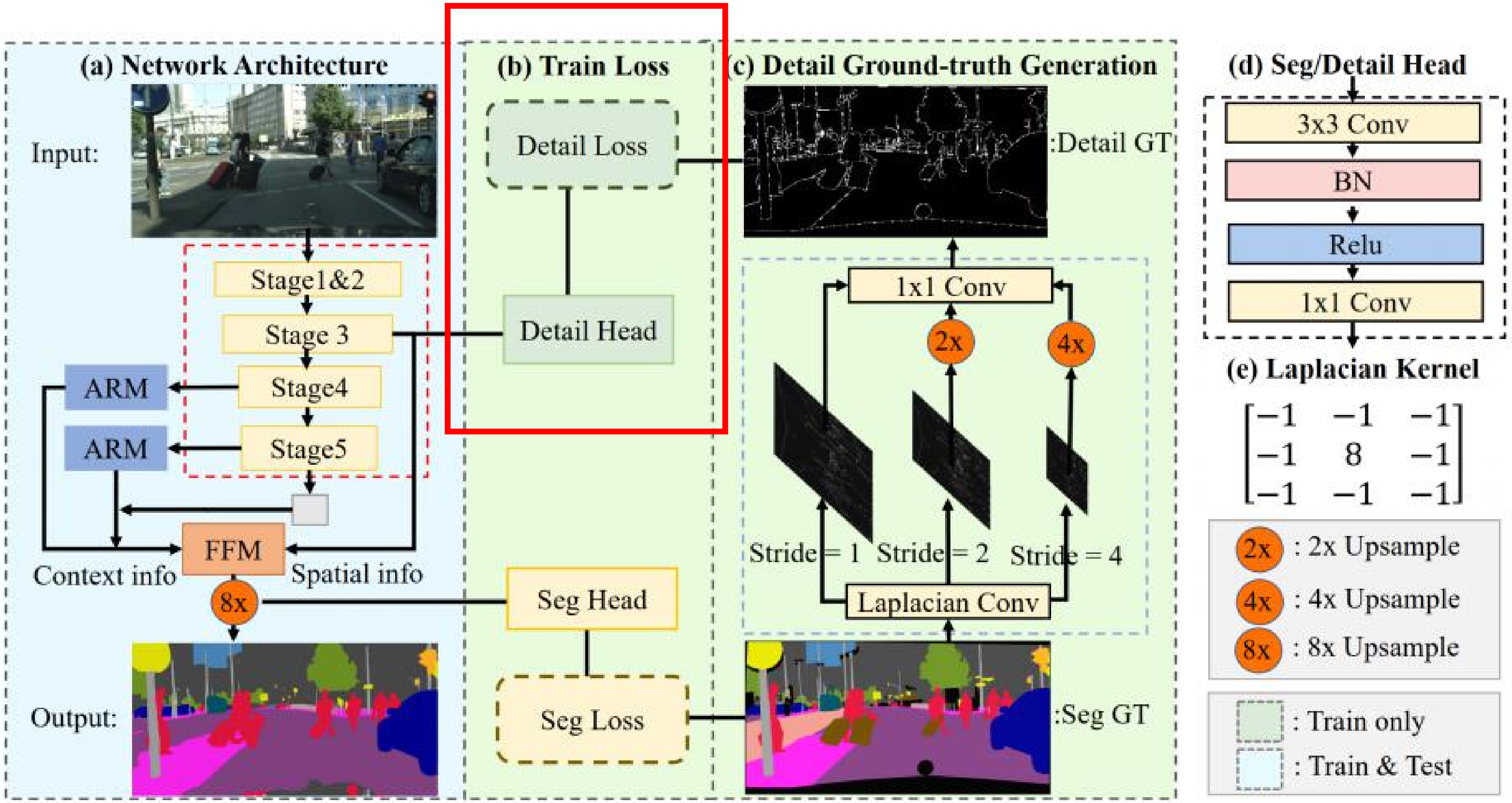
Decoder

- Detail GT Generation
  - Generate binary detail GT from seg. GT by Detail Aggregation Module (**dashed box**)
  - **Thresh: 0.1**. Convert details to binary detail GT



Decoder

- Detail Loss
  - Binary CE + Dice Loss
  - Dice Loss:
    - Measure overlap between pred. & GT
    - Insensitive to class imbalance



$$L_{dice}(p_d, g_d) = 1 - \frac{2\sum_i^{H\times W} p_d^i g_d^i + \epsilon}{\sum_i^{H\times W}(p_d^i)^2 + \sum_i^{H\times W}(g_d^i)^2 + \epsilon} \quad (5)$$

Laplace smoothing item, 1

Experiments

- Backbones Comparison

Model	Top1 Acc.	Params	FLOPs	FPS
ResNet-18 [9]	69.0%	11.2M	1800M	1058.7
ResNet-50 [9]	75.3%	23.5M	3800M	378.7
DF1 [21]	69.8%	8.0M	746M	1281.3
DF2 [21]	73.9%	17.5M	1770M	713.2
DenseNet121 [15]	75.0%	9.9M	2882M	363.6
DenseNet161 [15]	76.2%	28.6M	7818M	255.0
GhostNet(x1.0) [8]	73.9%	5.2M	141M	699.1
GhostNet(x1.3) [8]	75.7%	7.3M	226M	566.2
MobileNetV2 [25]	72.0%	3.4M	300M	998.8
MobileNetV3 [12]	75.2%	5.4M	219M	661.2
EfficientNet-B0 [26]	76.3%	5.3M	390M	443.0
STDC1	73.9%	8.4M	813M	**1289.0**
STDC2	**76.4%**	12.5M	1446M	813.6

Table 5. Comparisons with other popular networks on ImageNet Classification.

Classification

Backbone	Resolution	mIoU(%)	FPS
GhostNet [8]	512×1024	67.8	135.0
MobileNetV3 [12]	512×1024	70.1	148.3
EfficientNet-B0 [26]	512×1024	72.2	99.9
STDC2	512×1024	**74.2**	**188.6**
GhostNet [8]	768×1536	71.3	60.9
MobileNetV3 [12]	768×1536	73.0	70.4
EfficientNet-B0 [26]	768×1536	73.9	45.9
STDC2	768×1536	**77.0**	**97.0**

Table 3. Lightweight backbone comparison on Cityscapes *val* set. All experients utilize the same decoder and same experiment settings.

Segmentataion

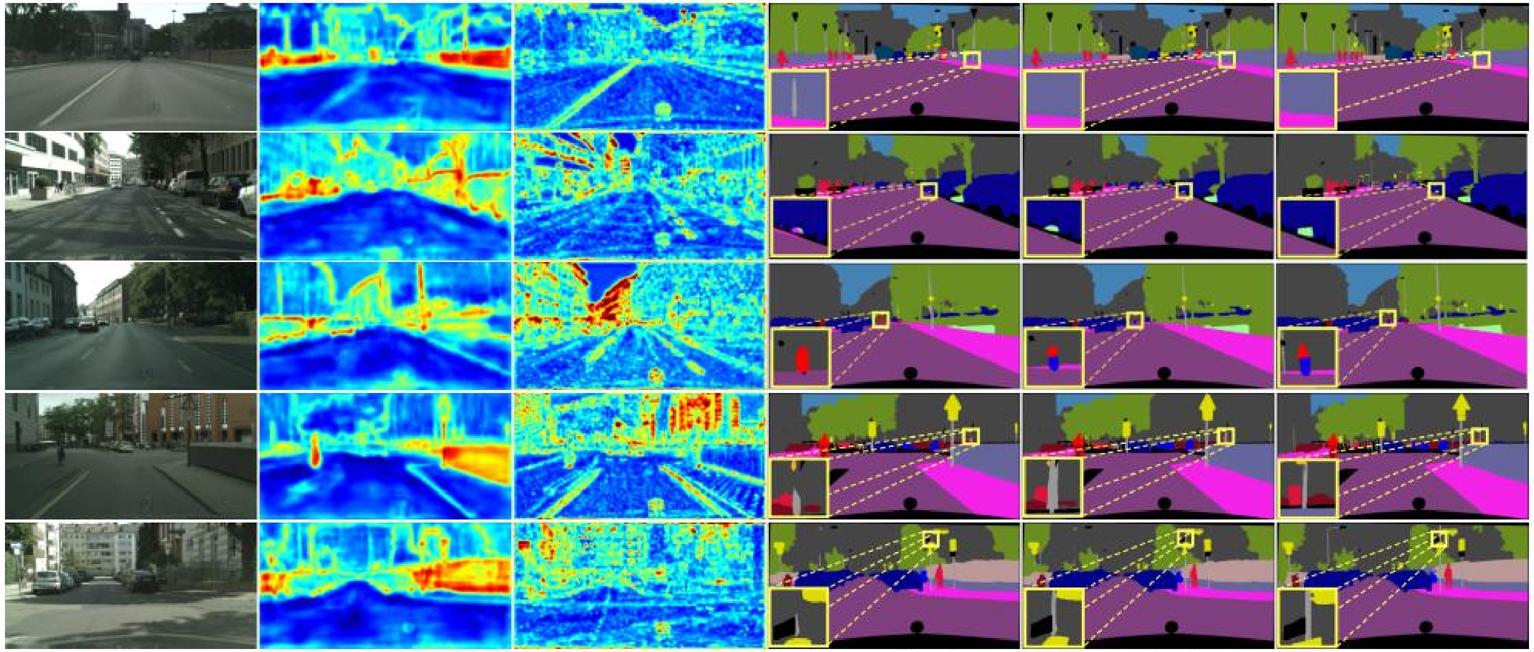
Experiments

Segmentation Results

Model	Resolution	Backbone	mIoU(%) val	mIoU(%) test	FPS
ENet [24]	512 × 1024	no	-	58.3	76.9
ICNet [31]	1024 × 2048	PSPNet50	-	69.5	30.3
DABNet [17]	1024 × 2048	no	-	70.1	27.7
DFANet B [18]	1024 × 1024	Xception B	-	67.1	120
DFANet A’ [18]	512 × 1024	Xception A	-	70.3	160
DFANet A [18]	1024 × 1024	Xception A	-	71.3	100
BiSeNetV1 [28]	768 × 1536	Xception39	69.0	68.4	105.8
BiSeNetV1 [28]	768 × 1536	ResNet18	74.8	74.7	65.5
CAS [30]	768 × 1536	no	-	70.5	108.0
GAS [22]	769 × 1537	no	-	71.8	108.4
DF1-Seg-d8 [21]	1024 × 2048	DF1	72.4	71.4	136.9
DF1-Seg[21]	1024 × 2048	DF1	74.1	73.0	106.4
DF2-Seg1[21]	1024 × 2048	DF2	75.9	74.8	67.2
DF2-Seg2[21]	1024 × 2048	DF2	76.9	75.3	56.3
SFNet [20]	1024 × 2048	DF1	-	74.5	121
HMSeg [19]	768 × 1536	no	-	74.3	83.2
TinyHMSeg [19]	768 × 1536	no	-	71.4	172.4
BiSeNetV2 [27]	512 × 1024	no	73.4	72.6	156
BiSeNetV2-L [27]	512 × 1024	no	75.8	75.3	47.3
FasterSeg [4]	1024 × 2048	no	73.1	71.5	163.9
STDC1-Seg50	512 × 1024	STDC1	72.2	71.9	**250.4**
STDC2-Seg50	512 × 1024	STDC2	74.2	73.4	188.6
STDC1-Seg75	768 × 1536	STDC1	74.5	75.3	126.7
STDC2-Seg75	768 × 1536	STDC2	**77.0**	**76.8**	97.0

Table 6. Comparisons with other state-of-the-art methods on Cityscapes. *no* indicates the method do not have a backbone.

Visualization



(a) Input (b) Stage3 (c) $Stage3_D$ (d) Prediction (e) $Prediction_D$ (f) Groundtruth

Figure 6. Visual comparison of our *Detail Guidance* on Cityscapes *val* set. The column with subscript **D** denotes results with *Detail Guidance*. The first row (a) shows the input images. (b) and (c) illustrate the heatmap of Stage 3 without and with *Detail Guidance*. (d) and (e) demonstrate the predictions without and with *Detail Guidance*. (f) is the ground-truth of input images.