Dual Super-Resolution Learning for Semantic Segmentation

Li Wang[*,1], Dong Li[1], Yousong Zhu[2], Lu Tian[1], Yi Shan[1]

Emails: {liwa, dongl, lutian, yishan}@xilinx.com, yousong.zhu@nlpr.ia.ac.cn

[1] Xilinx Inc., Beijing, China.

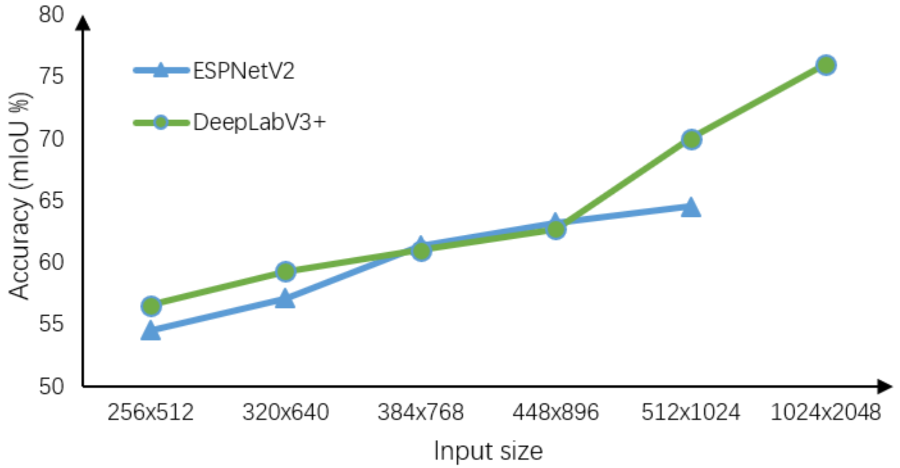
[2] Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Motivation

- Large computation budgets for SOTA methods
- Performance degradation for light-weight methods
- High-resolution input (*e.g.*, 1024x2048)

Method	GFLOPs	Cityscapes mIoU (val)
PSPNet (ResNet101) [1]	1149.92	79.70%
DeepLabv3+ (Xception-65) [2]	837.28	78.79%
DeepLabv3+ (MoileNetv2) [2]	42.54	70.71%
ESPNetv2 [3]	5.4	64.50%

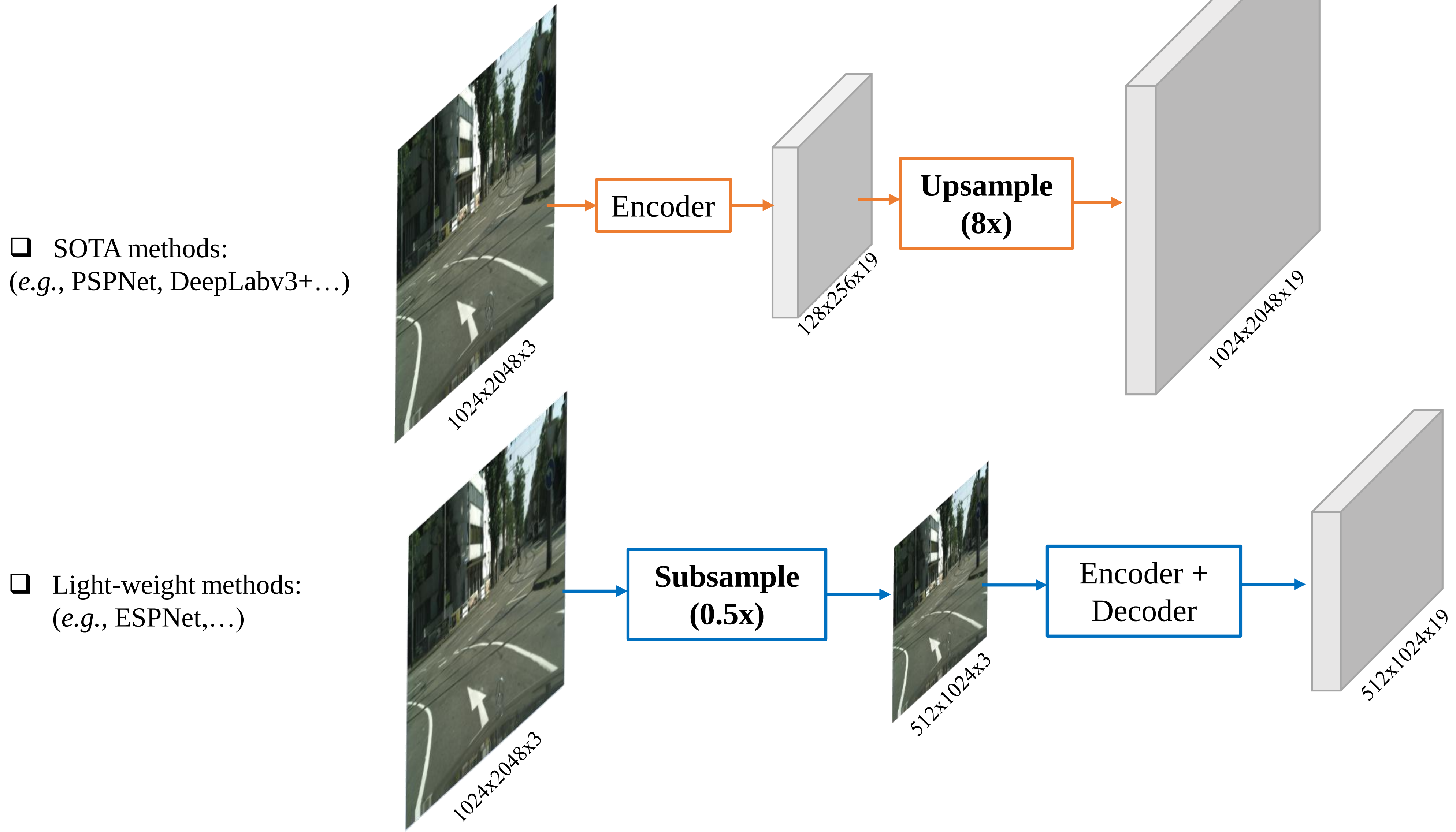
Accuracy vs. GFLOPs of current state-of-the-art methods on Cityscapes



Accuracy vs. Input size for different networks on Cityscapes

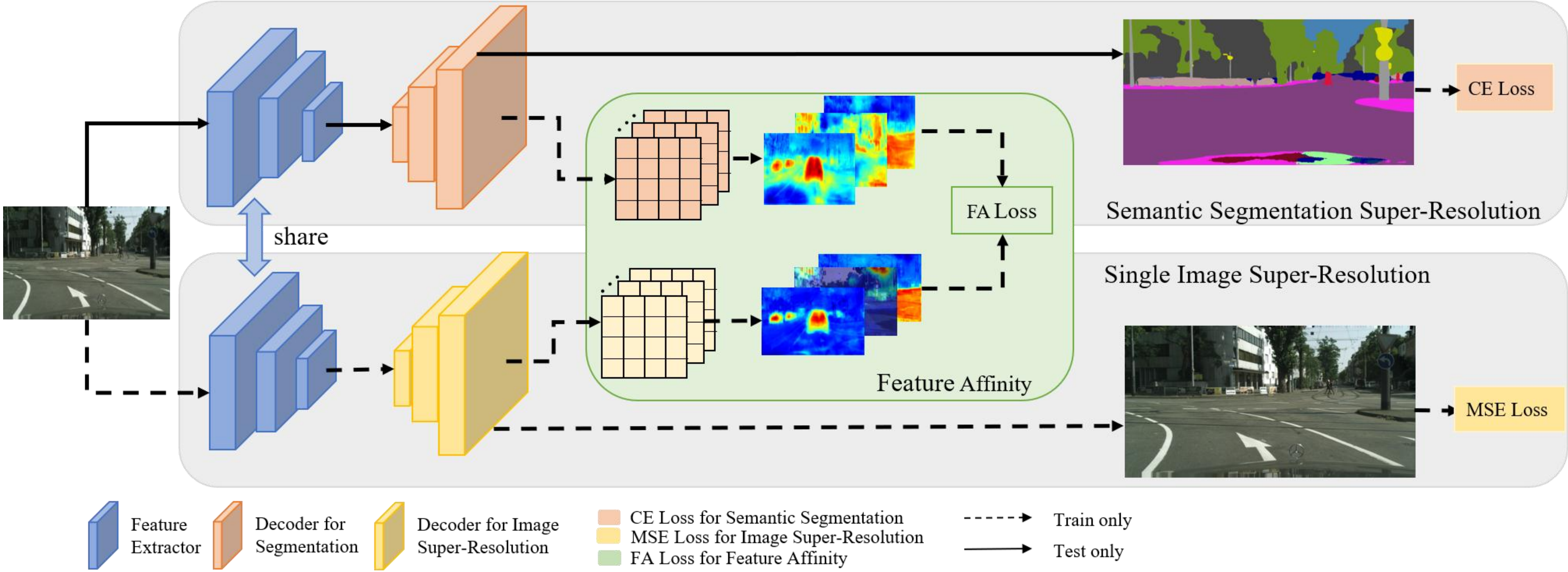
[1] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In CVPR, 2017
[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018
[3] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In CVPR, 2019

Review of Existing Methods



[1] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In CVPR, 2017
[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018
[3] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In CVPR, 2019

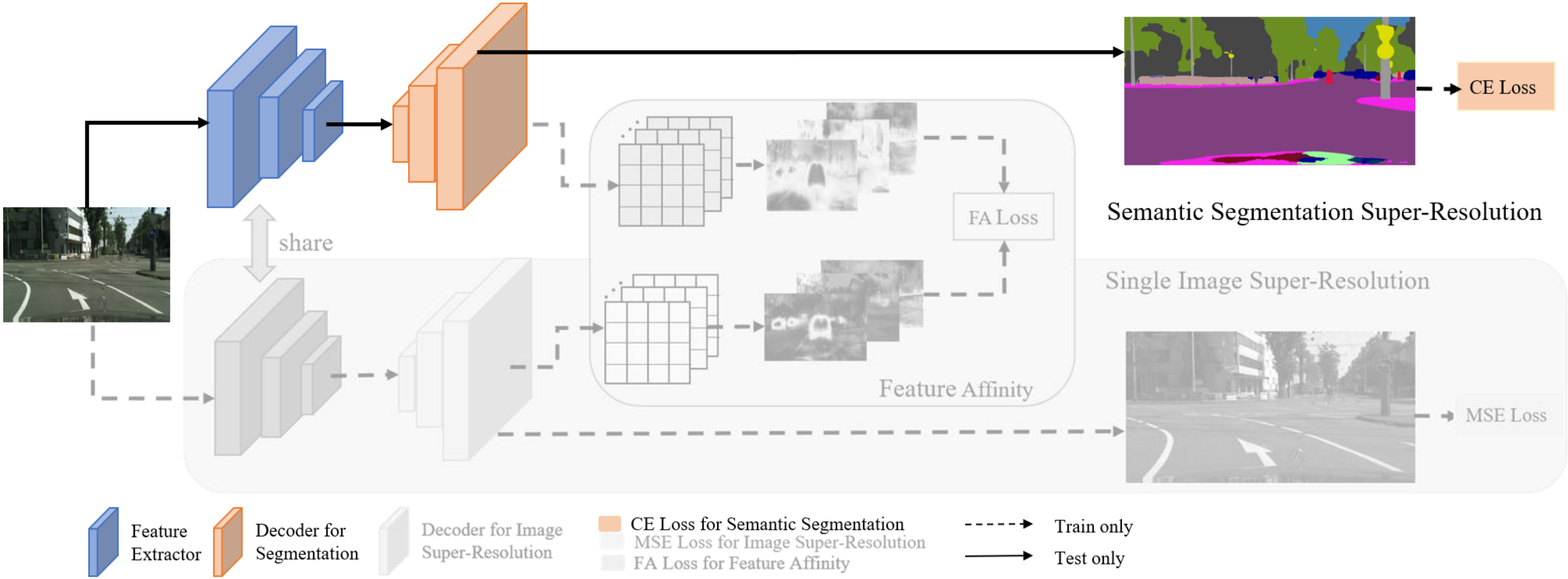
Our Method



share
CE Loss
FA Loss
Semantic Segmentation Super-Resolution
Single Image Super-Resolution
Feature Affinity
MSE Loss
Feature Extractor
Decoder for Segmentation
Decoder for Image Super-Resolution
CE Loss for Semantic Segmentation
MSE Loss for Image Super-Resolution
FA Loss for Feature Affinity
Train only
Test only

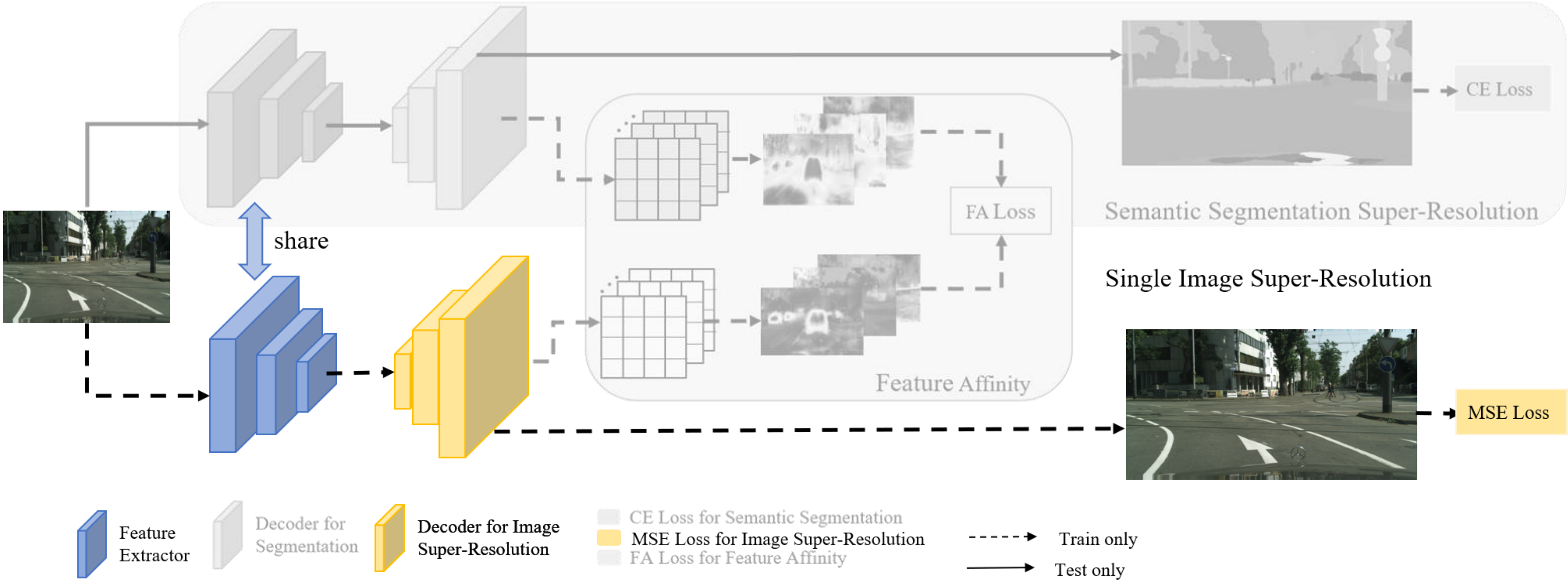
Our Method

- **Semantic Segmentation Super-Resolution (SSSR)**



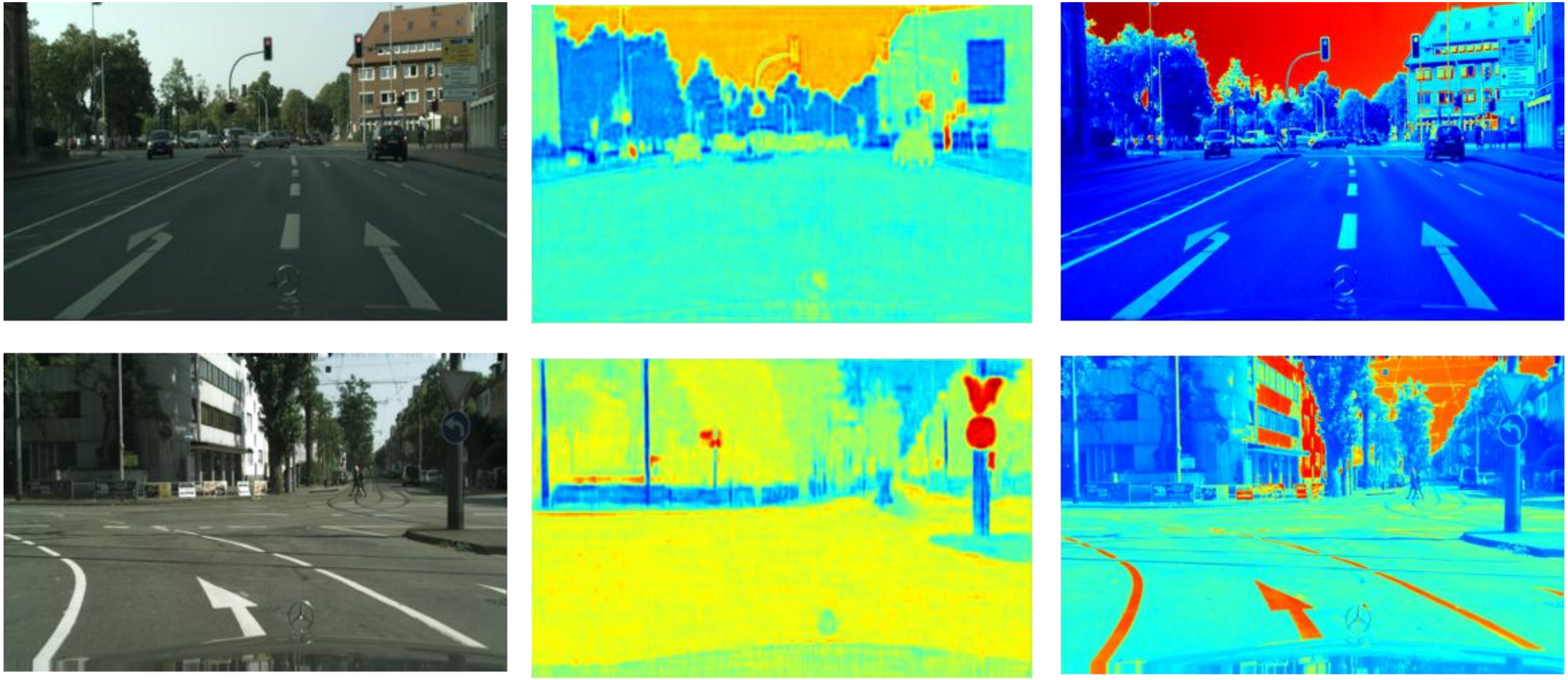
Our Method

➢ **Single Image Super-Resolution (SISR)**



Our Method

➢ Single Image Super-Resolution (SISR)

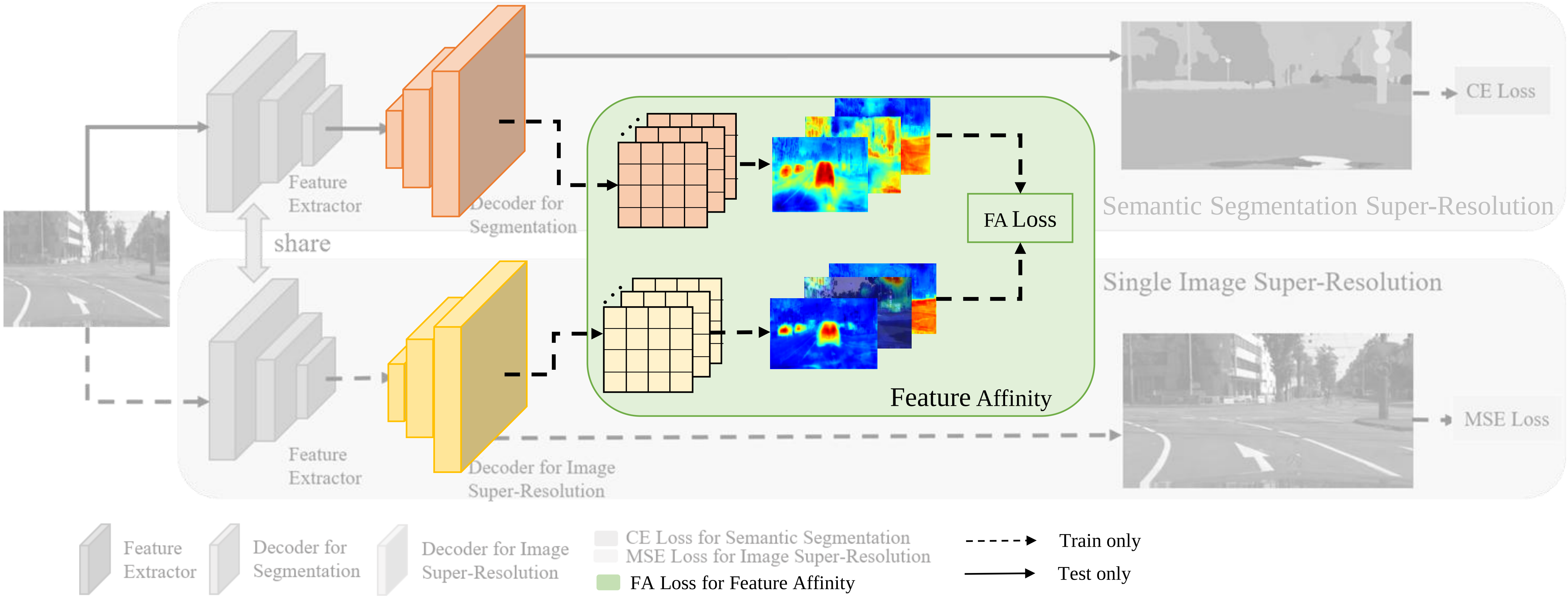


(a) (b) (c)

Feature-level visualization for SSSR and SISR under the same input (0.5x). (a) Input image, (b) SSSR feature visualization, (c) SISR feature visualization.

Our Method

- **Feature Affinity (FA)**



Our Method

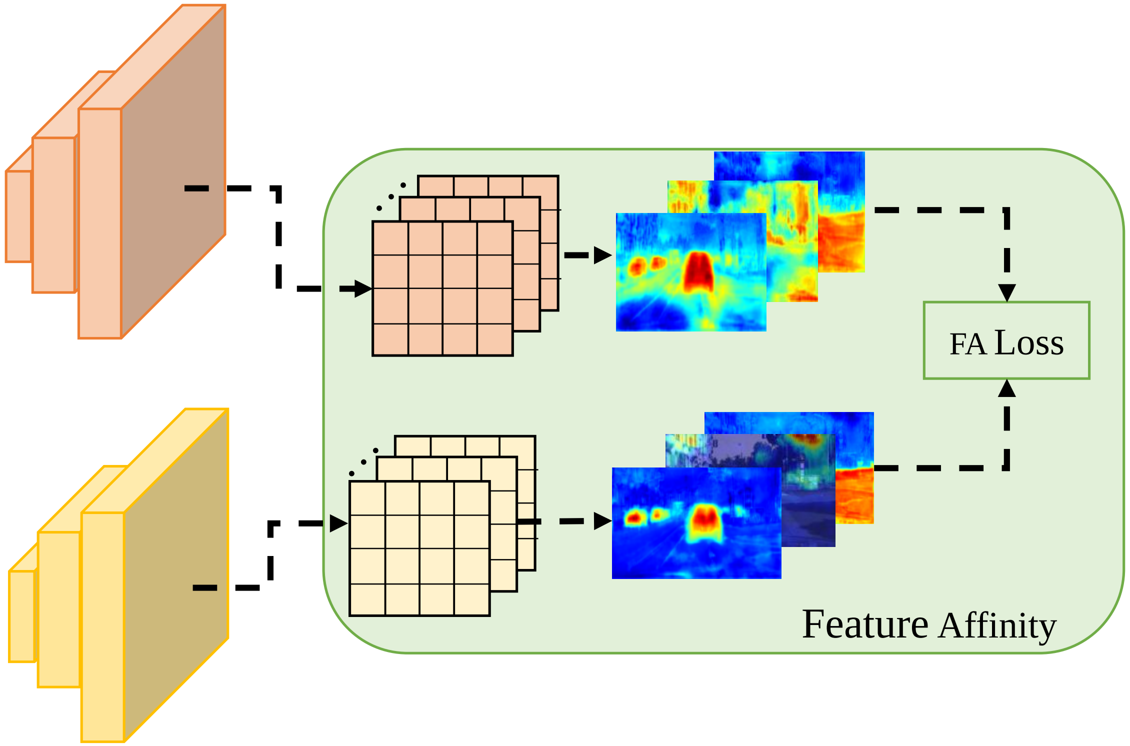
Feature Affinity (FA)

$$L_{fa} = \frac{1}{W'^2 H'^2} \sum_{i=1}^{W'H'} \sum_{j=1}^{W'H'} ||S_{ij}^{seg} - S_{ij}^{sr}||_q$$

Where:

$$S_{ij} = \left(\frac{F_i}{||F_i||_p}\right)^T \cdot \left(\frac{F_j}{||F_j||_p}\right)$$

$p = 2, \; q = 1$



Loss function

$$L = L_{ce} + w_1 L_{mse} + w_2 L_{fa}$$

Where:

$$L_{ce} = \frac{1}{N} \sum_{i=1}^{N} -y_i log(p_i) \qquad L_{mse} = \frac{1}{N} \sum_{i=1}^{N} ||SISR(X_i) - Y_i||^2$$

$w_1 = 0.1, w_2 = 0.1$

Experiments on Semantic Segmentation

Ablation Study on Cityscapes

- Effect of algorithmic components

Method	Input	Output	Val. mIoU
ESPNetv2 [1]	256x512	256x512	54.5%
+ SSSR	256x512	512x1024	55.7%
+ SSSR + SISR	256x512	512x1024	56.9%
+ SSSR + SISR + FA	256x512	512x1024	**59.5%**

Method	Input	Output	Val. mIoU
DeepLabv3+ [2]	256x512	256x512	56.5%
+ SSSR	256x512	512x1024	57.1%
+ SSSR + SISR	256x512	512x1024	57.4%
+ SSSR + SISR + FA	256x512	512x1024	**59.2%**

- Effect of various input resolutions

Method	256x512	320x640	384x768	448x896	512x1024
ESPNetv2 [1]	54.5%	57.1%	61.4%	63.2%	64.5%
ESPNetv2 (ours)	**59.5%**	**61.9%**	**64.0%**	**65.7%**	**66.9%**
DeepLabv3+ [2]	56.5%	59.3%	62.0%	63.7%	70.0%
DeepLabv3+ (ours)	**59.2%**	**61.7%**	**64.3%**	**65.7%**	**72.0%**

[1] Sachin Mehta, Mohammad Rastegari, Linda Shapiro, and Hannaneh Hajishirzi. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In CVPR, 2019.
[2] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV, 2018

Experiments on Semantic Segmentation

Ablation Study on Cityscapes

Effect of various networks

Method	Val. (%)	Test (%)	GFLOPs
ESPNetv2	64.5	65.1	5.40
ESPNetv2 w/ DSRL	**66.9**	**65.9**	**5.40**
DABNet	62.6	65.0	20.44
DABNet w/ DSRL	**65.4**	**66.2**	**20.44**
BiseNet	62.6	61.8	49.20
BiseNet w/ DSRL	**66.8**	**64.9**	**49.20**
DeepLabv3+	70.0	67.1	974.30
DeepLabv3+ w/ DSRL	**72.0**	**69.3**	**974.30**
PSPNet	71.5	69.1	287.48
PSPNet w/ DSRL	**74.4**	**73.4**	**287.48**

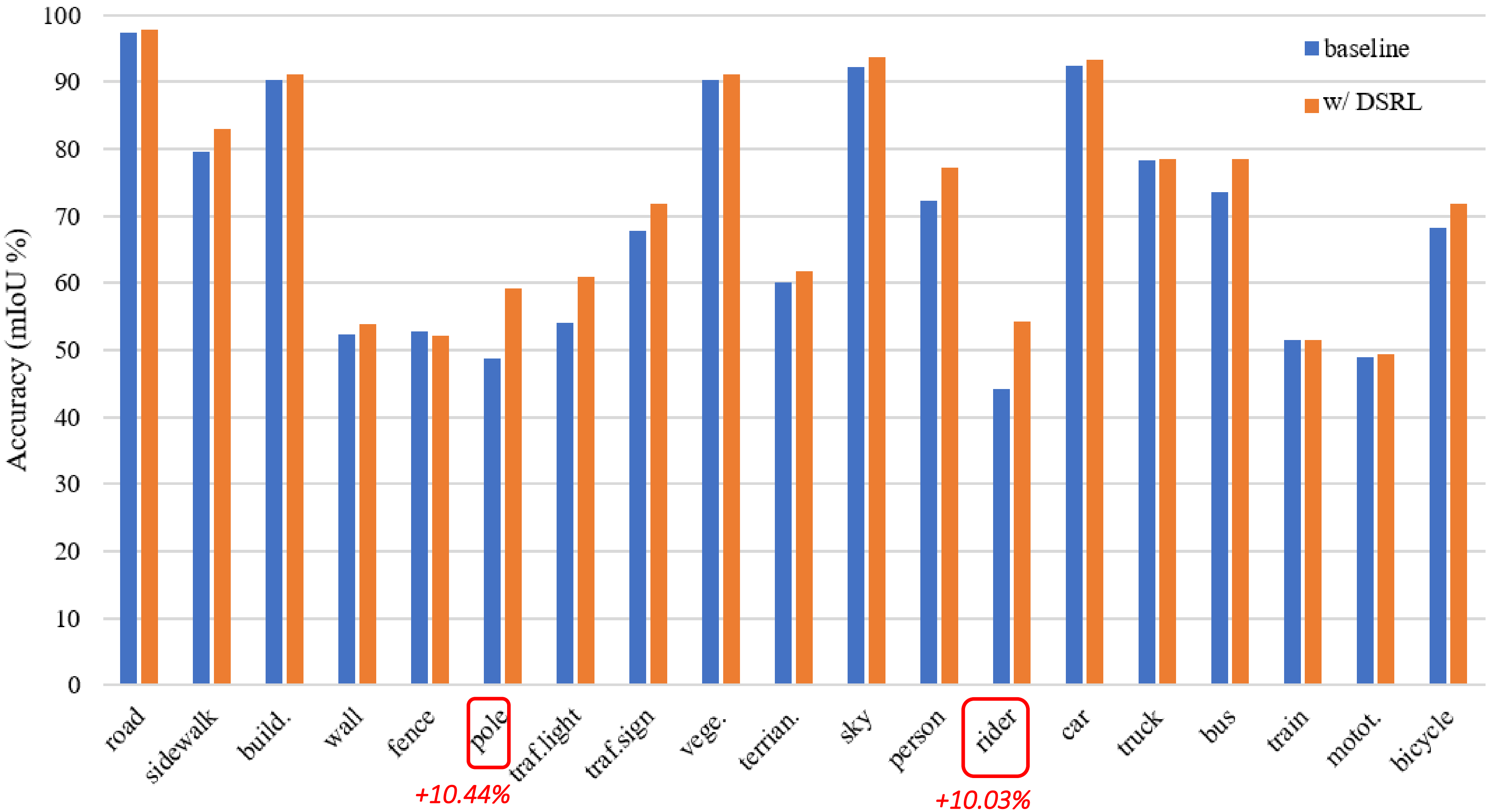
Consistent improvement without extra computation cost

Comparisons with state-of-the-art results

Method	**Test (%)**	**GFLOPs**
ENet	58.3	7.24
ESPNet	60.3	8.86
ERFNet	68.0	25.60
PSPNet(ResNet18(0.5))	54.1	133.40
PSPNet(ResNet18(0.5)) w/ Distillation	60.5	133.40
PSPNet(ResNet18(1.0))	67.6	512.80
PSPNet(ResNet18(1.0)) w/ Distillation	71.4	512.80
FCN	65.3	1335.60
RefineNet	73.6	2102.80
ESPNet (ours)	65.1	5.40
DeepLabv3+ (ours)	69.3	974.30
PSPNet (ours)	**73.4**	287.48

Experiments on Semantic Segmentation

➢ Results on Cityscapes

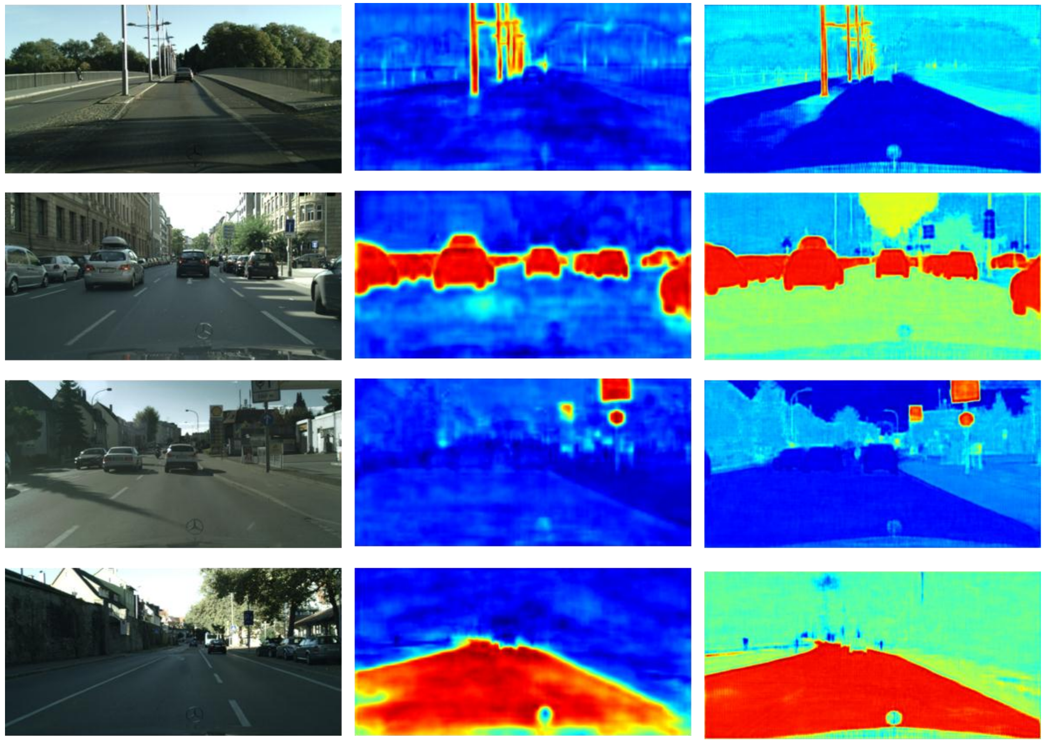


Comparisons of the DeepLabv3+ baseline and our DSRL method in terms of per-class IoU scores on Cityscapes

Remarkable improvement on small classes

Experiments on Semantic Segmentation

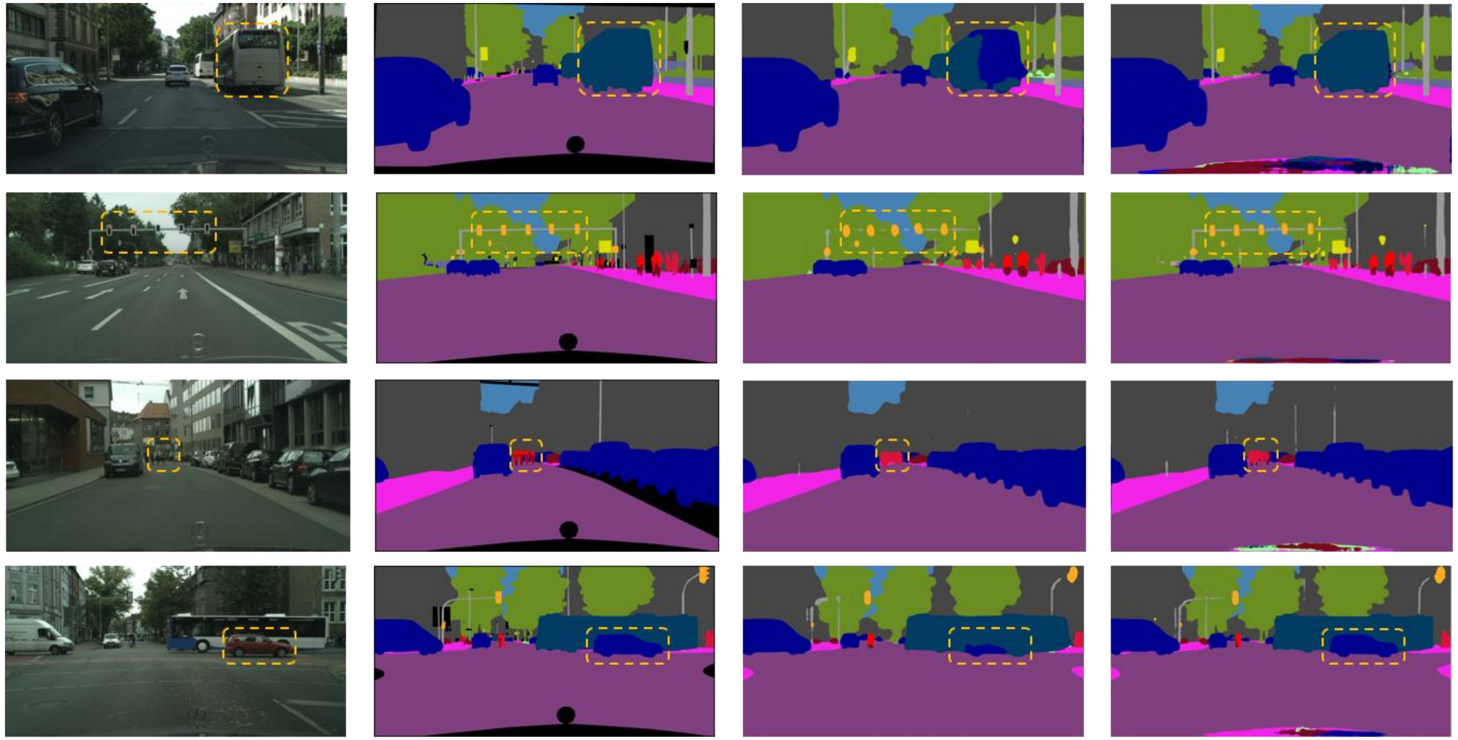
- Results on Cityscapes



(a) (b) (c)

Visualization of segmentation features, (a) input image, (b) the ESPNetv2 baseline method, (c) our DSRL method.

Capture more fine-grained structure information of objects



(a) (b) (c) (d)

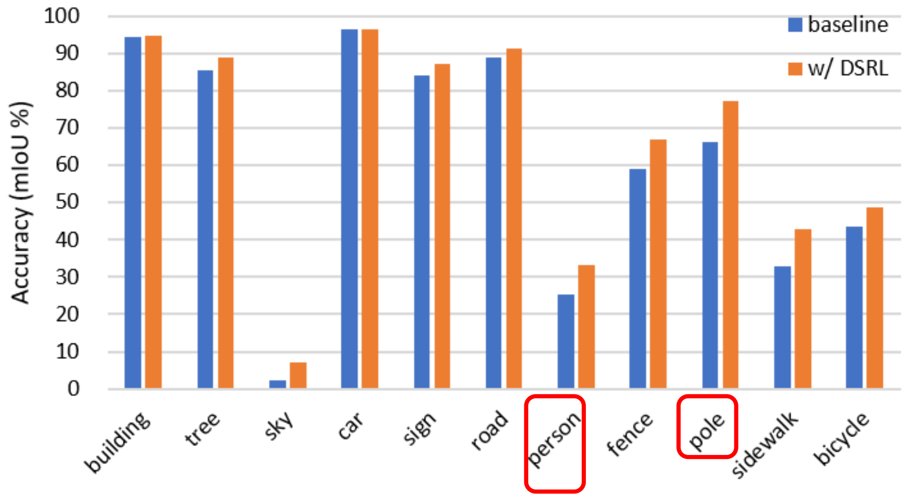
Comparisons of segmentation results. (a) Input image. (b) Ground truth. (c) The DeepLabv3+ baseline method. (d) Our DSRL method.

Better segmentation results on various classes

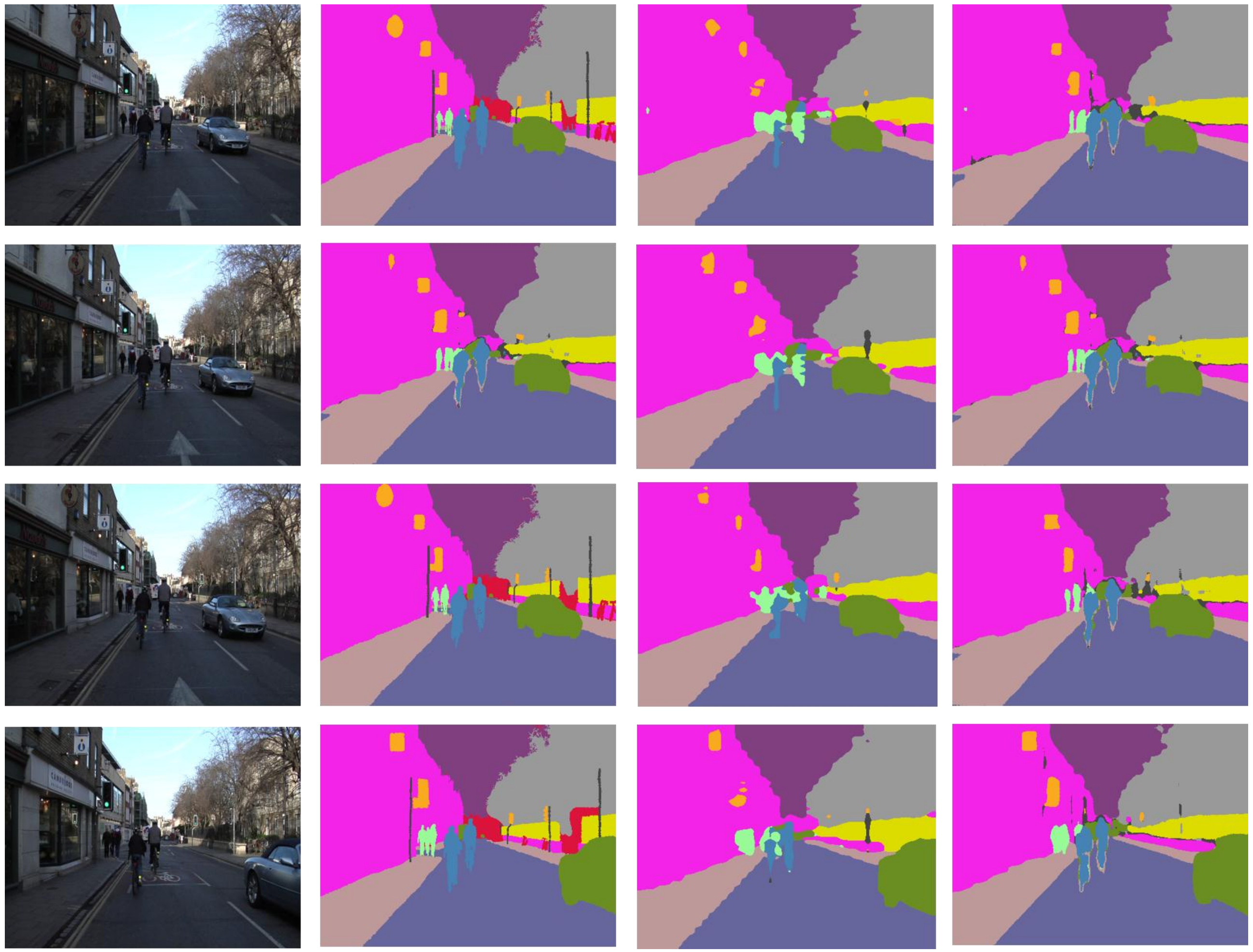
Experiments on Semantic Segmentation

➢ Results on CamVid

Method	Test (%)	GFLOPs
ESPNetv2	50.9	1.82
ESPNetv2 w/ DSRL	**54.4**	1.82
BiSeNet	53.4	4.14
BiSeNet w/ DSRL	**57.0**	4.14
DeepLabv3+	60.4	326.13
DeepLabv3+ w/ DSRL	**63.7**	326.13



Comparisons of the DeepLabv3+ baseline and our DSRL method in terms of per-class IoU scores on CamVid.



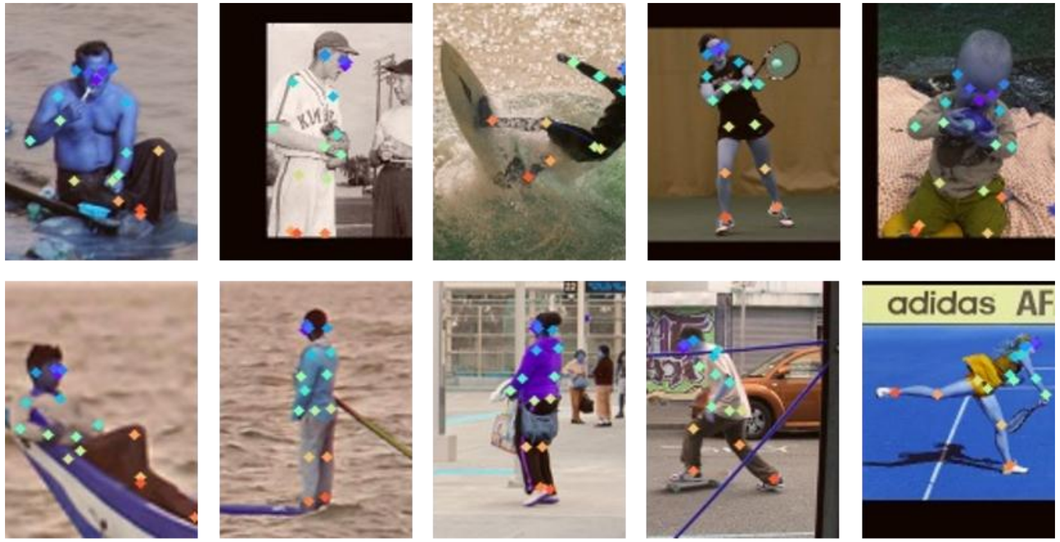
(a) (b) (c) (d)

Examples of segmentation results on CamVid. (a) Input image. (b) Ground Truth. (c) The DeepLabv3+ baseline method. (d) Our DSRL method.

Experiments on Human Pose Estimation

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

- Architecture: HRNet-w32 [1]
- Dataset: MS COCO2017 [2]



Consistent improvement on different resolutions

Method	Input	mAP	AP@0.5	AP@0.75	AR	AR@0.5	AR@0.75	FLOPs
HRNet-w32	256x192	74.4%	90.5%	81.9%	79.8%	94.2%	86.5%	7.12G
HRNet-w32(ours)	256x192	**75.6%**	**92.2%**	**83.0%**	**81.2%**	93.8%	**88.5%**	7.12G
HRNet-w32	160x128	69.2%	89.3%	78.1%	75.7%	93.6%	83.7%	2.97G
HRNet-w32(ours)	160x128	**71.5%**	**89.6%**	**79.4%**	**77.5%**	**93.7%**	**84.5%**	2.97G
HRNet-w32	128x96	64.6%	87.8%	73.9%	71.7%	92.8%	80.2%	1.78G
HRNet-w32(ours)	128x96	**67.9%**	**88.3%**	**76.7%**	**74.5%**	**92.8%**	**82.4%**	1.78G

[1] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. 2019.
[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014.

Conclusion

- Dual Super-Resolution Learning (DSRL)
  - Semantic Segmentation Super-Resolution (SSSR)
    - *Learn high-resolution representations for prediction*
  - Single Image Super-Resolution (SISR)
    - *Capture fine-grained structural representations without extra annotations*
  - Feature Affinity (FA)
    - *Learn similarity between SSSR and SISR features for better knowledge transfer*
- Effectiveness and versatility
  - Improve the performance while keeping the same computation cost
  - Reduce the computation cost while keeping the similar performance
  - Generalized to other dense prediction tasks (*e.g.*, human pose estimation)

Thanks!

Project Page: https://github.com/wanglixilinx/DSRL