ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

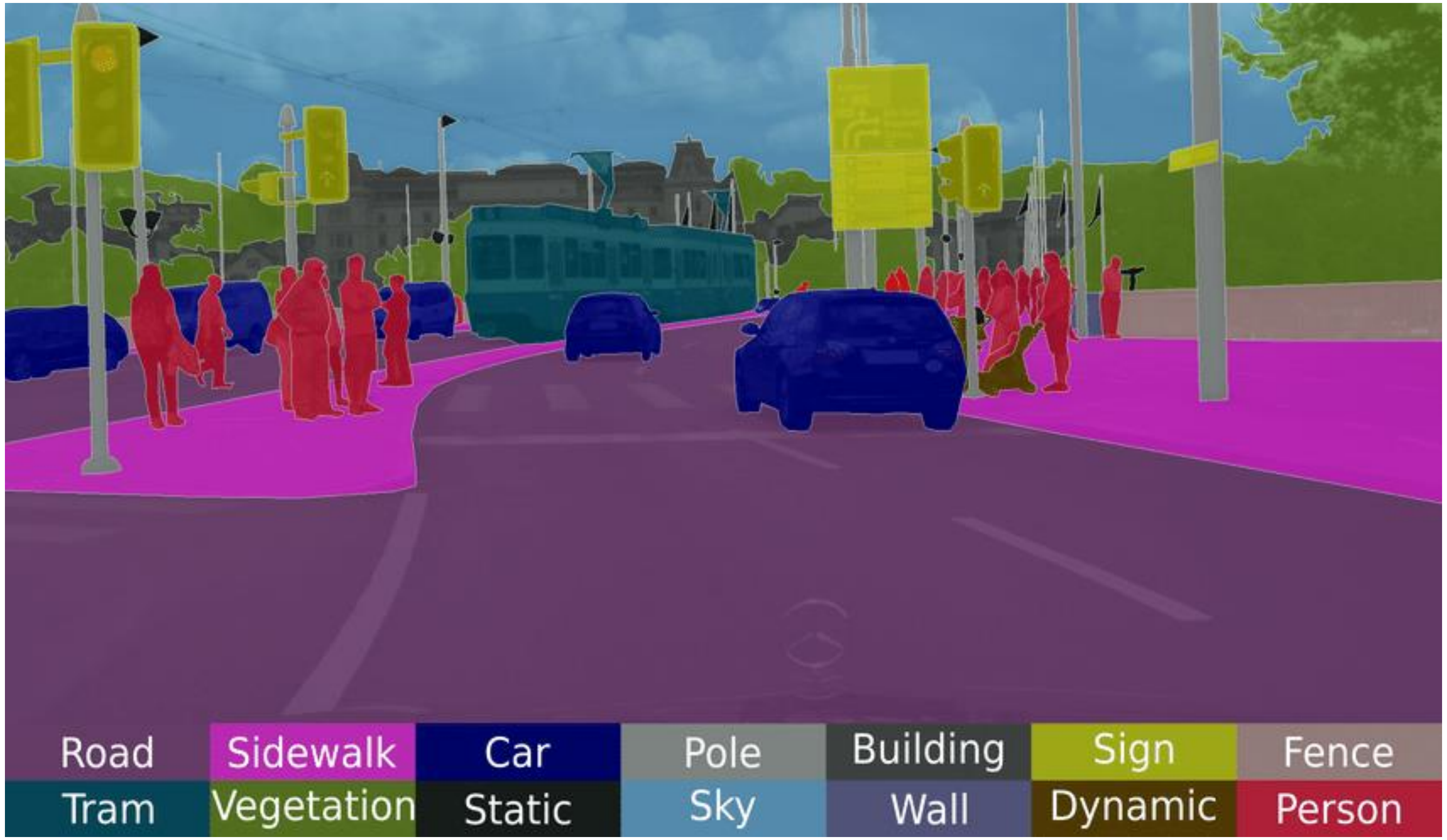
Poster 738

Fast and Accurate Real-Time Semantic Segmentation with Dilated Asymmetric Convolutions

Leonel Rosas-Arias[1], ***Gibran Benitez-Garcia[2]***, Jose Portillo-Portillo[1],
Gabriel Sanchez-Perez[1] and Keiji Yanai[2]

[1]Instituto Politecnico Nacional, Mexico City, Mexico
[2]The University of Electro-Communications, Tokyo, Japan

Real-time semantic segmentation



Find and classify pixels belonging to each objects in the scene.

Problem:

- **High-accuracy** semantic segmentation is extremely **expensive to compute**.
- Networks for **real-time semantic segmentation** sacrifice a lot of accuracy.

Objective:

- Reduce the accuracy gap between *real-time* and *non-real-time for* semantic segmentation.

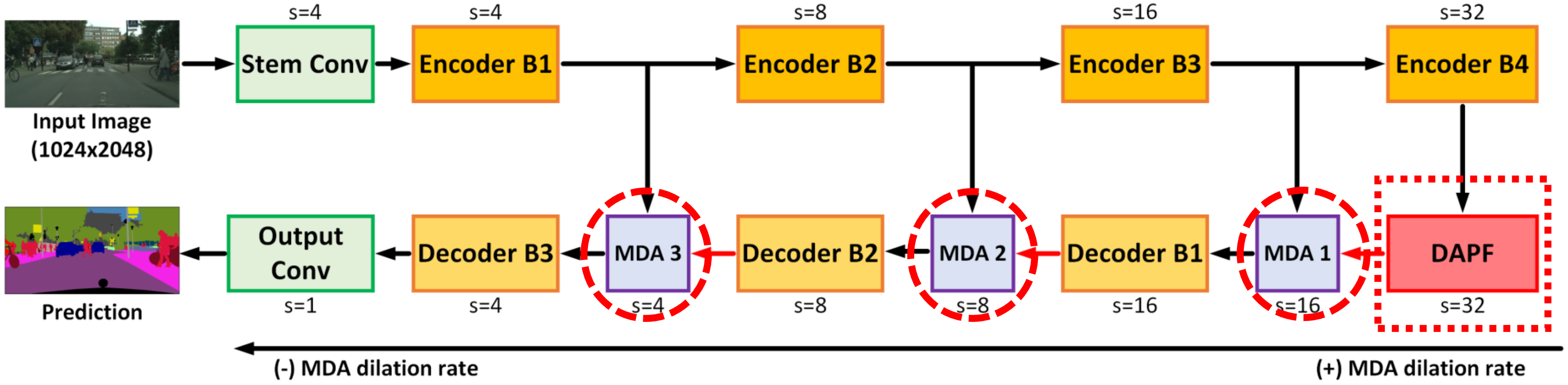
Image from: Janai, Joel & Guney, Fatma & Behl, Aseem & Geiger, Andreas. Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art, 2017.

Contributions

- ***FASSD-Net:***
  - SOTA performance of ***speed and accuracy*** with high resolution images (1024x2048).
  - ***Two additional variations*** to balance the speed and accuracy trade-offs.
- ***Dilated Asymmetric Pyramidal Fusion module (DAPF):***
  - Obtains feature maps rich in ***contextual information***.
  - Requires considerably ***fewer floating-point operations*** compared with similar method, such as ASPP [1].
- ***Multi-resolution Dilated Asymmetric module (MDA):***
  - Offers an improved way to ***fuse two set of feature maps*** of different resolution.
  - Simultaneously ***refines spatial and contextual information*** from input feature maps.
  - Can be used in **all decoder stages**.

[1] Chen, L.C, et.al. *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation*. ECCV'18

FASSD-Net architecture

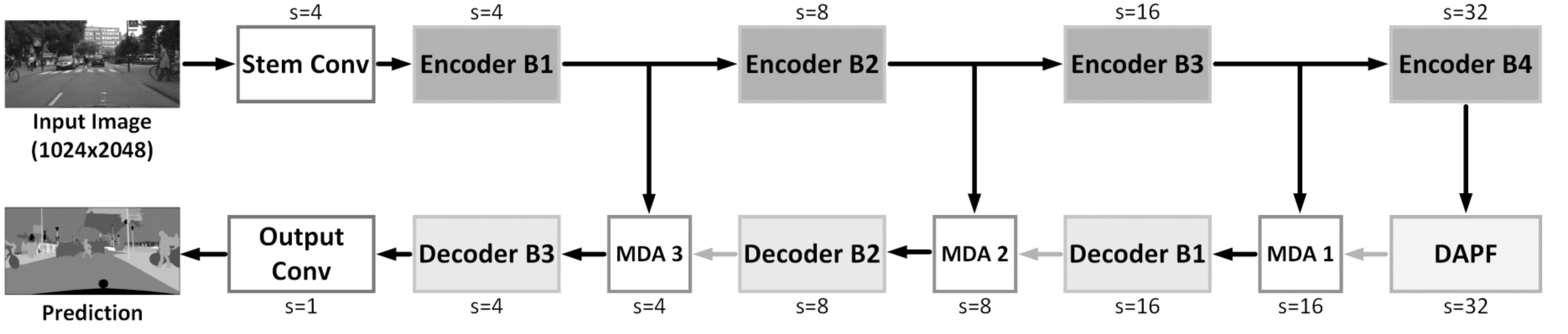


FASSD-Net:

- **Encoder.-** HarDNet [2] (custom version)
- **Decoder.-** **DAPF + MDA**

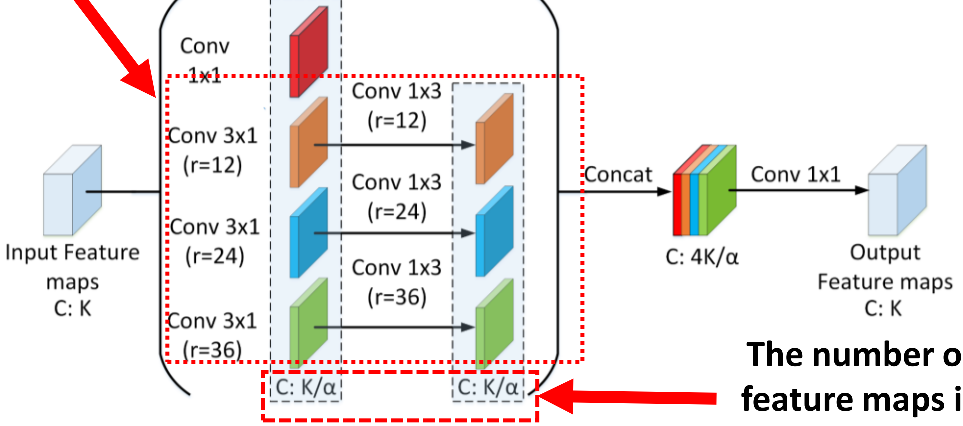
[2] P. Chao, et. al. *HarDNet: A Low Memory Traffic Network*, ICCV 2019.

FASSD-Net architecture



With α=2, our proposed module requires 50% fewer floating-point operations compared to ASPP.

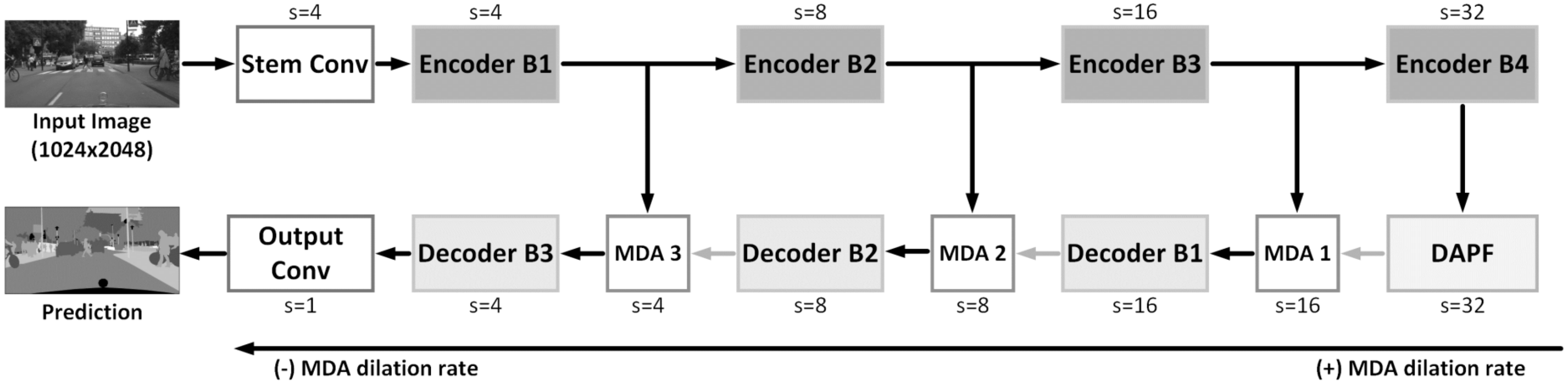
Factorized convolutions

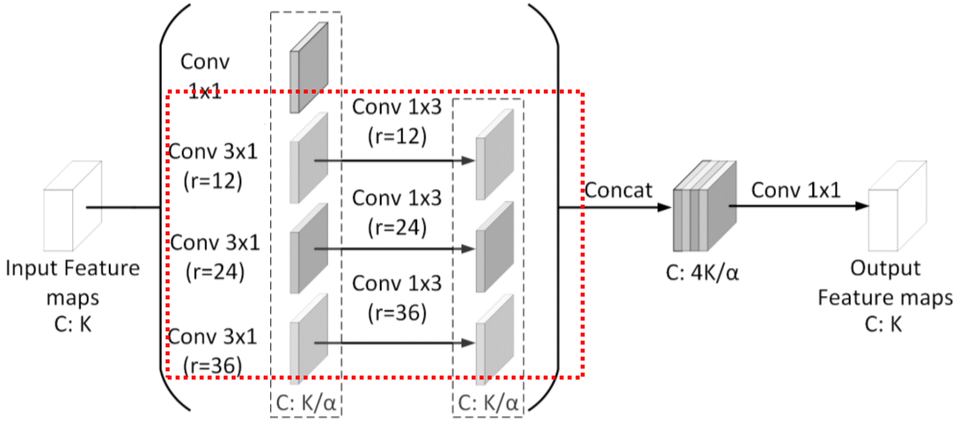


DAPF Module

The number of intermediate feature maps is controlled by the α parameter.

FASSD-Net architecture

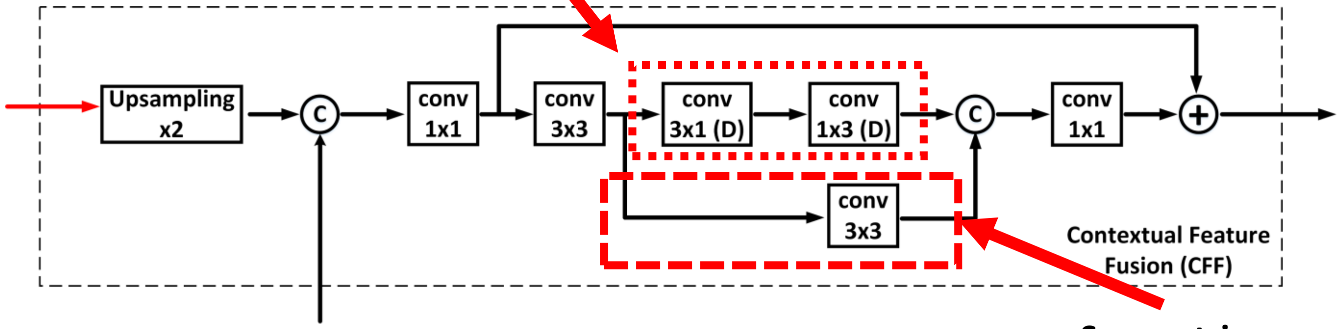




DAPF Module

Asymmetric branch: Context

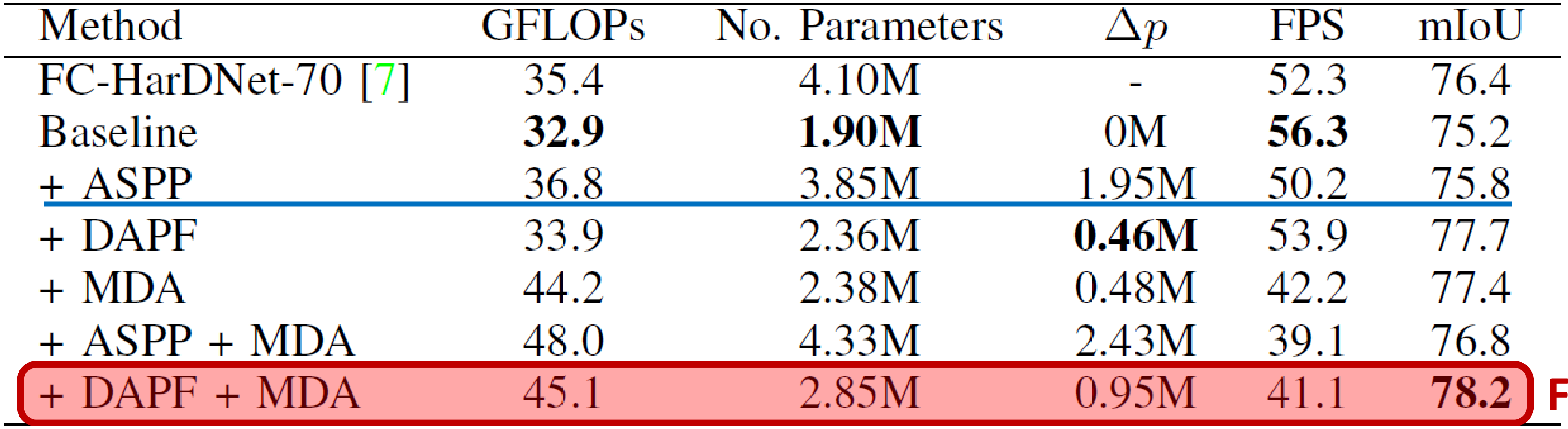
Refines both contextual and detail information



Multi-resolution Dilated Asymmetric module (MDA)

Symmetric branch: Details

Ablation study on the Cityscapes dataset



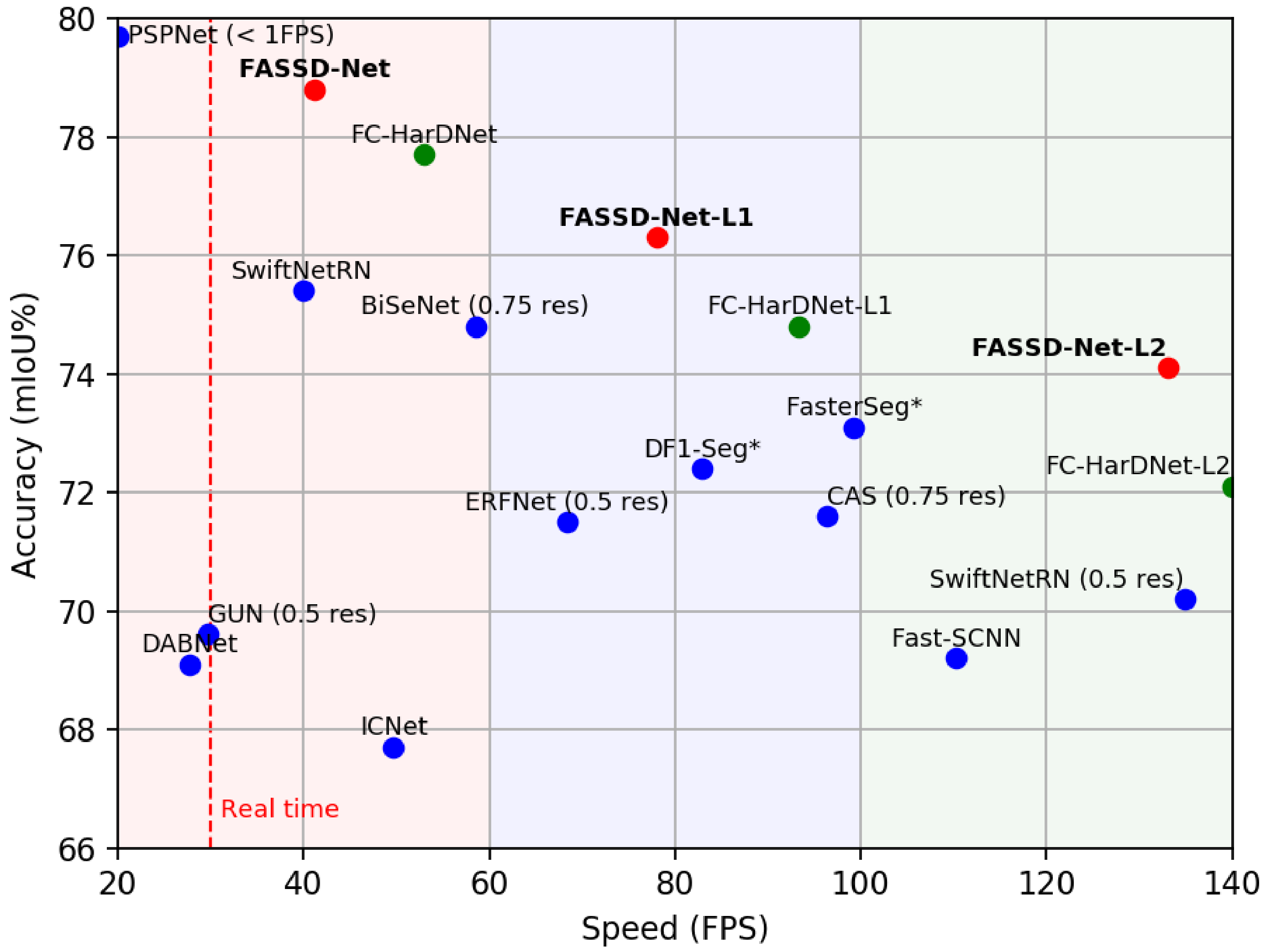
Method	GFLOPs	No. Parameters	Δp	FPS	mIoU
FC-HarDNet-70 [7]	35.4	4.10M	-	52.3	76.4
Baseline	**32.9**	**1.90M**	0M	**56.3**	75.2
+ ASPP	36.8	3.85M	1.95M	50.2	75.8
+ DAPF	33.9	2.36M	**0.46M**	53.9	77.7
+ MDA	44.2	2.38M	0.48M	42.2	77.4
+ ASPP + MDA	48.0	4.33M	2.43M	39.1	76.8
+ DAPF + MDA	45.1	2.85M	0.95M	41.1	**78.2**

FASSD-Net

Experimental setup:

- All networks pretrained with ImageNet.
- Training during 90k iterations.
- Batch size = 16.
- GFLOPs measured for high resolution images at 1024x2048.
- Speed in FPS measured with an *Nvidia GTX 1080Ti* card.

Quantitative results

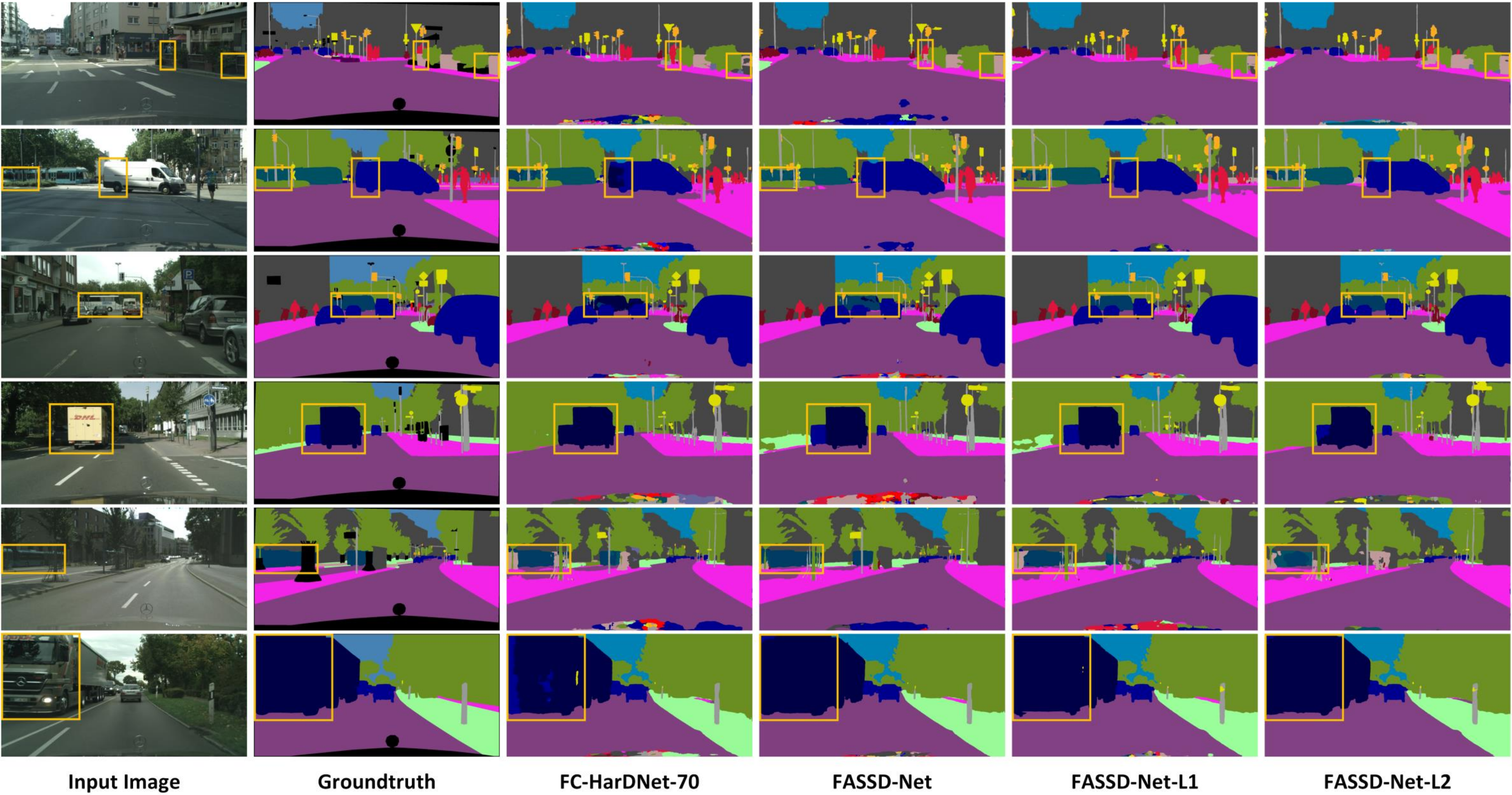


1) For fair comparison, the speed of methods marked by (*) are approximated ***without TensorRT*** acceleration [3].

2) ***PSPNet and FC-HarDNet-L2*** speeds are placed on the x-axis edges for the sake of better visualization.

[3] NVIDIA, “TensorRT,” [Online; accessed December 1, 2020] https://developer.nvidia.com/tensorrt

Qualitative results



Conclusion

- ***Conclusions:***
  - *We proposed two modules (DAPF & MDA) for **reducing the accuracy gap between real-time and non-real-time** semantic segmentation networks.*
  - *With FASSD-Net, we set a new **SOTA accuracy for real-time** semantic segmentation on the Cityscapes validation set.*

- ***Future work:***
  - *Include more **backbone networks** and different datasets for evaluation.*
  - *Evaluate on different scenarios, such as **indoor parsing and medical images**.*

Thank You

Code & Models