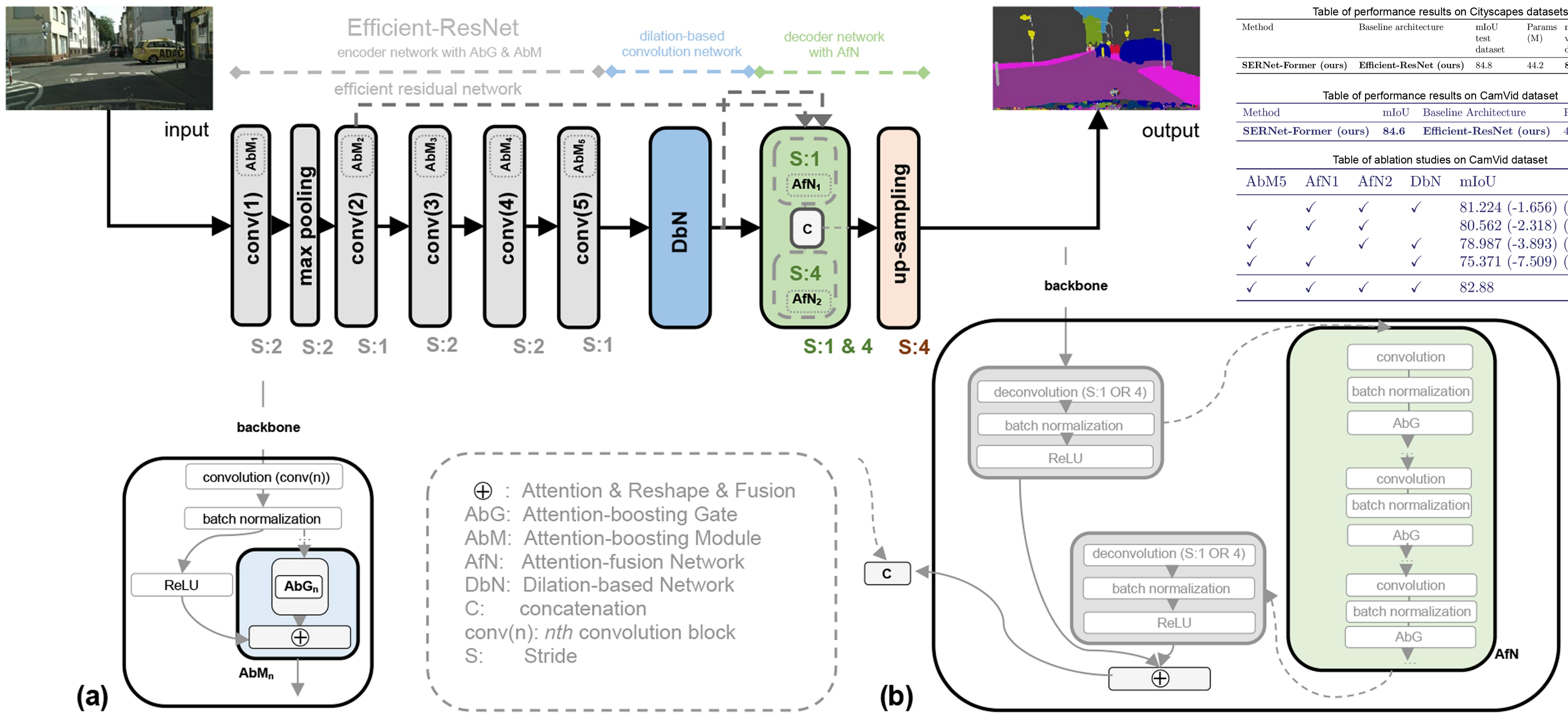
CVPR 2024 Workshops. Equivariant Vision: From Theory to Practice 18th June 2024

Author: Serdar Erişen
Hacettepe University
Ankara, Turkey
serdarerisen@ieee.org; serdarch@gmail.com; serdarerisen@hacettepe.edu.tr

SERNet-Former: Semantic Segmentation by Efficient Residual Network with Attention-Boosting Gates and Attention-Fusion Networks

Improving the efficiency of state-of-the-art methods in semantic segmentation requires overcoming the increasing computational cost as well as issues such as fusing semantic information from global and local contexts. Based on the recent success and problems that convolutional neural networks (CNNs) encounter in semantic segmentation, this research proposes an encoder-decoder architecture with a unique efficient residual network, Efficient-ResNet. Attention-boosting gates (AbGs) and attention-boosting modules (AbMs) are deployed by aiming to fuse the equivariant and feature-based semantic information with the equivalent sizes of the output of global context of the efficient residual network in the encoder. Respectively, the decoder network is developed with the additional attention-fusion networks (AfNs) inspired by AbM. AfNs are designed to improve the efficiency in the one-to-one conversion of the semantic information by deploying additional convolution layers in the decoder part. Our network is tested on the challenging CamVid and Cityscapes datasets, and the proposed methods reveal significant improvements on the residual networks. To the best of our knowledge, the developed network, SERNet-Former, achieves state-of-the-art results (84.62 % mean IoU) on CamVid dataset and challenging results (87.35 % mean IoU) on Cityscapes validation dataset.



Schematic illustration of SERNet-Former. (a) Attention-boosting Gate (AbG) and Attention-boosting Module (AbM) are fused into the encoder part. (b) Attention-fusion Network (AfN), introduced into the decoder.

Table of performance results on Cityscapes datasets

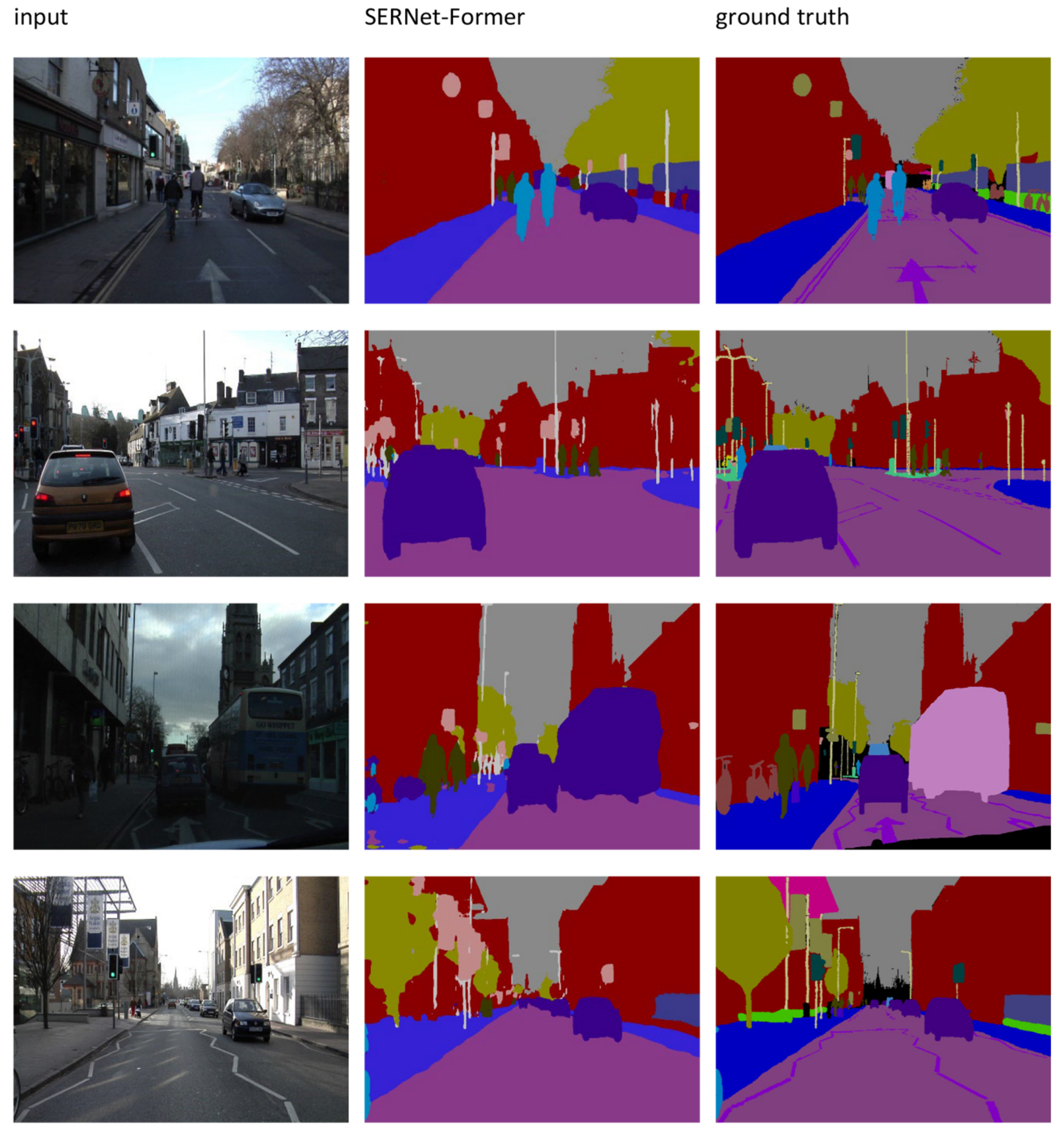
Method	Baseline architecture	mIoU test dataset	Params (M)	mIoU validation dataset
SERNet-Former (ours)	**Efficient-ResNet (ours)**	84.8	44.2	**87.35**

Table of performance results on CamVid dataset

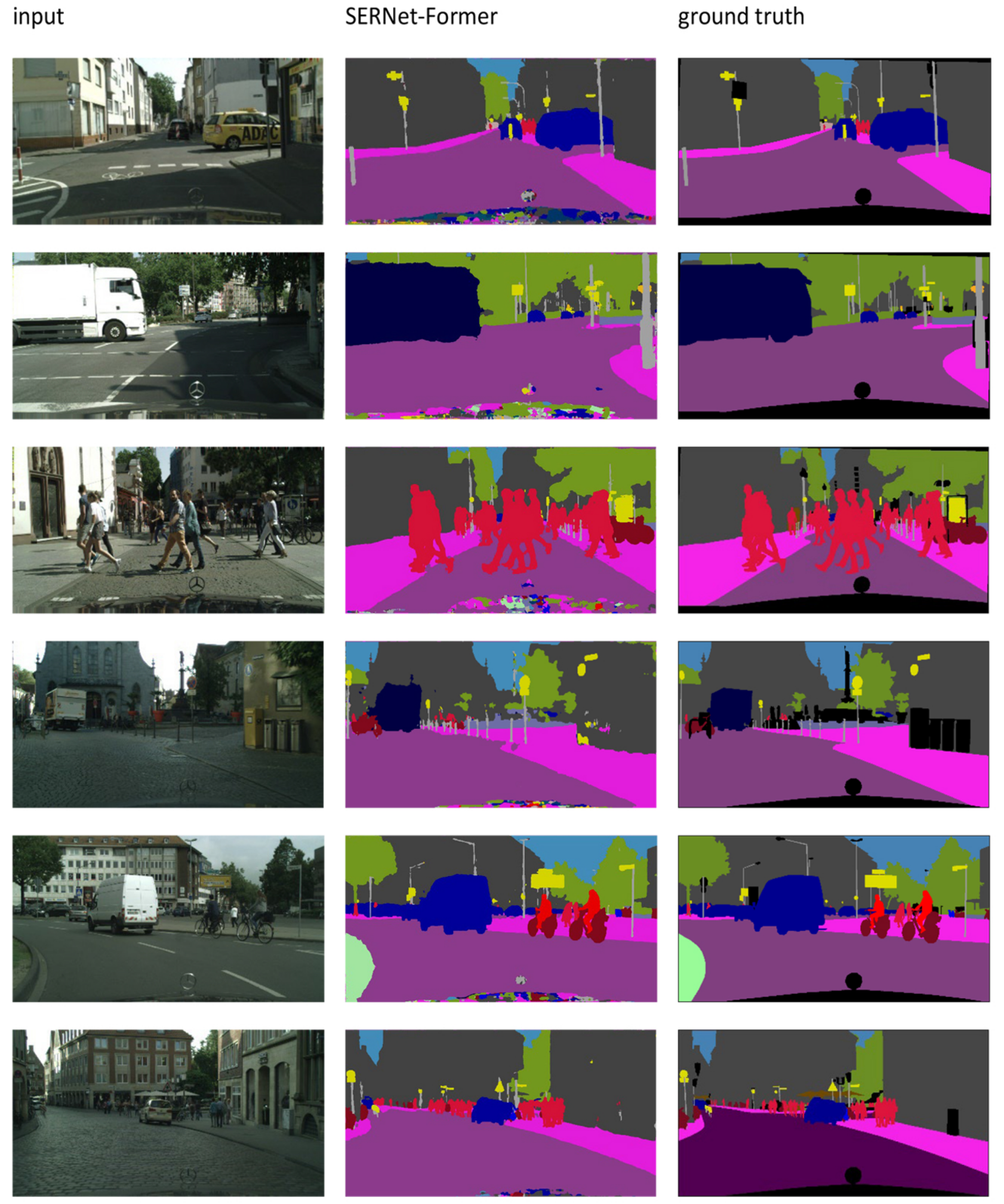
Method	mIoU	Baseline Architecture	Params (M)
SERNet-Former (ours)	**84.6**	**Efficient-ResNet (ours)**	44.2

Table of ablation studies on CamVid dataset

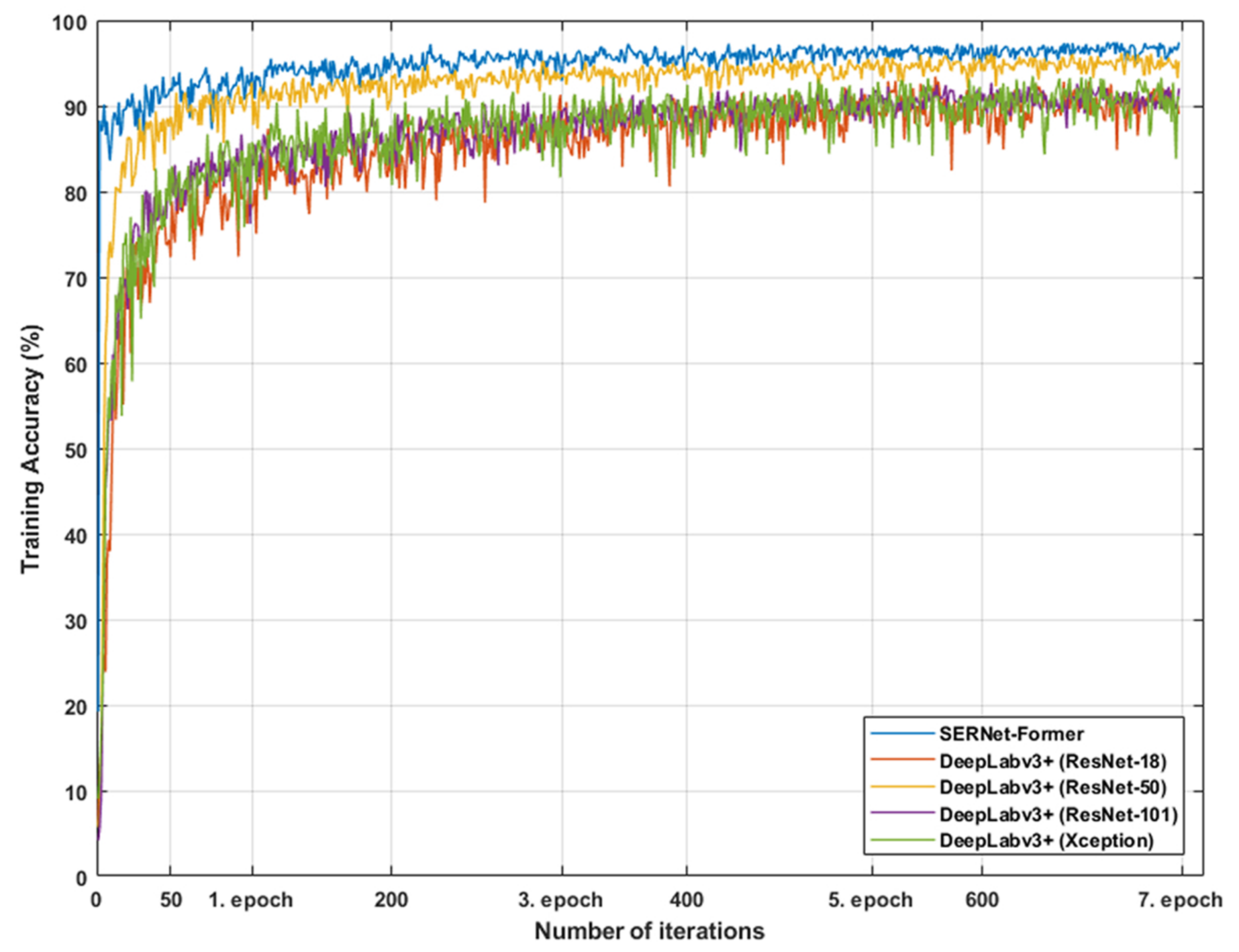
AbM5	AfN1	AfN2	DbN	mIoU
	✓	✓	✓	81.224 (-1.656) (1.99 %)
✓	✓	✓		80.562 (-2.318) (2.79 %)
✓		✓	✓	78.987 (-3.893) (4.69 %)
✓	✓		✓	75.371 (-7.509) (9 %)
✓	✓	✓	✓	82.88



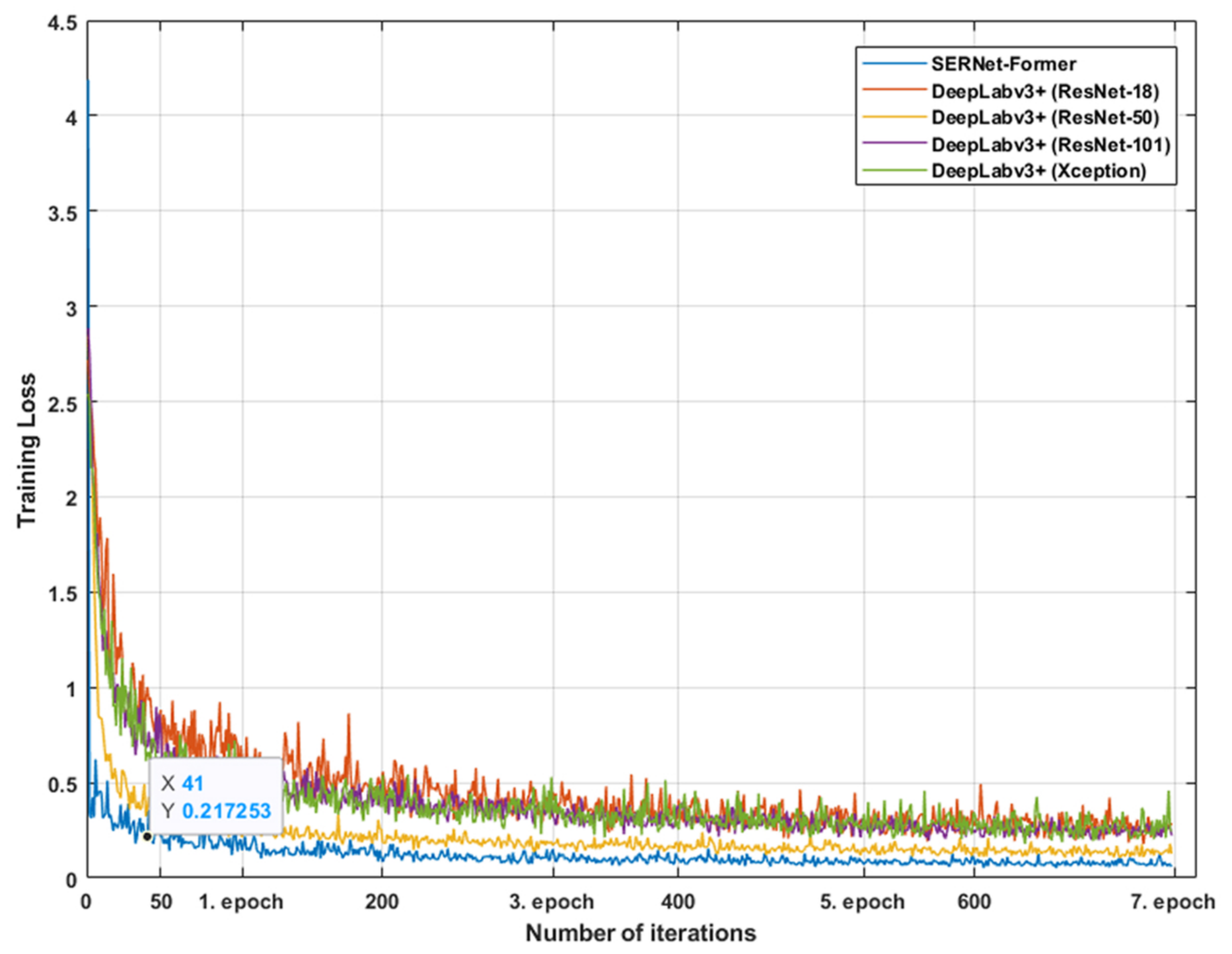
Segmentation results of SERNet-Former on allocated CamVid test dataset. Left column: Image inputs. Middle column: Prediction outputs of SERNet-Former. Right column: Ground truth of annotated labels



Segmentation results of SERNet-Former on Cityscapes validation dataset. Left column: Image inputs. Middle column: Prediction outputs of SERNet-Former. Right column: Ground truth of annotated labels



Training accuracy of SERNet-Former with Efficient-ResNet baseline on CamVid dataset



Training loss of SERNet-Former with Efficient-ResNet baseline on CamVid dataset

- **Efficient-ResNet** is developed as an encoder with the additional excitation of **attention-boosting gates (AbGs)** and **attention-boosting modules (AbMs)** for the optimal training performance and computational cost of CNNs
- The capacity of the decoder part of our network is improved via **attention-fusion networks (AfNs)**, increasing the efficiency of acquiring and processing the feature-rich semantic information
- **Skip connections**, turning the decoder part into a **superposition network**, are designed to fuse and concatenate multi-scale information from the global and local contexts
- The network achieves **state-of-the-art** performances on **CamVid** and **Cityscapes validation** datasets.

Table of class accuracies of SERNet-Former on CamVid dataset

Method	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Bicycle	mIoU
SERNet-Former (ours)	93.0	88.8	95.1	91.9	73.9	97.7	76.4	83.4	57.3	90.3	83.1	**84.6**

Table of class accuracies of SERNet-Former on Cityscapes dataset

Method	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	mIoU
SERNet-Former* (ours)	96.8	76.3	90.0	57.2	54.6	52.9	60.5	66.0	90.9	64.6	93.9	79.0	61.6	93.5	69.7	85.3	74.7	59.7	65.6	73.3
SERNet-Former† (ours)	98.2	90.2	94.0	67.6	68.2	73.6	78.2	82.1	94.6	75.9	96.9	90.0	77.7	96.9	86.1	93.9	91.7	70.0	82.9	84.8

*: ResNet-50 baseline without AbM, DbN, AfN. †: SERNet-Former with AbM, DbN, AfN