Improving Deep Network Performance via Model Compression

Yuenan Hou
Multimedia Lab, The Chinese University of Hong Kong

About me

Yuenan HOU:

- A 4th year Ph.D. candidate in MMLab, CUHK (graduate in July, 2021)
- Supervised by Prof. Chen Change Loy (NTU associate professor) and Prof. Xiaoou Tang (CUHK professor)
- Google Scholar citation is 175 and h-index is 5
- Has 6 first-author papers published in / submitted to top conferences, e.g., CVPR, ICCV, and 1 second-author paper submitted to MICCAI
- Reviewers for several top conferences and journals, e.g., CVPR, AAAI, TIP
- Internship in Sensetime Research and has 5 patents in total
- Received national scholarship (2014), first prize in MCM (2016), special award for undergraduate thesis (2017), etc

Publications

1. Network Pruning via Resource Reallocation
submitted to International Conference on Computer Vision (ICCV), 2021
Yuenan Hou, Zheng Ma, Chunxiao Liu, Zhe Wang, Chen Change Loy

2. Patchwise Contrastive Distillation for Generative Adversarial Networks
submitted to International Conference on Computer Vision (ICCV), 2021
Yuenan Hou, Xinge Zhu, Chen Change Loy

3. Inter-Region Affinity Distillation for Road Marking Segmentation
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, acceptance rate: 22.1% (1470/6656)
Yuenan Hou, Zheng Ma, Chunxiao Liu, Tak-Wai Hui, Chen Change Loy

4. Learning Lightweight Lane Detection CNNs by Self Attention Distillation
International Conference on Computer Vision (ICCV), 2019, acceptance rate: 25.0% (1077/4304)
Yuenan Hou, Zheng Ma, Chunxiao Liu, Chen Change Loy

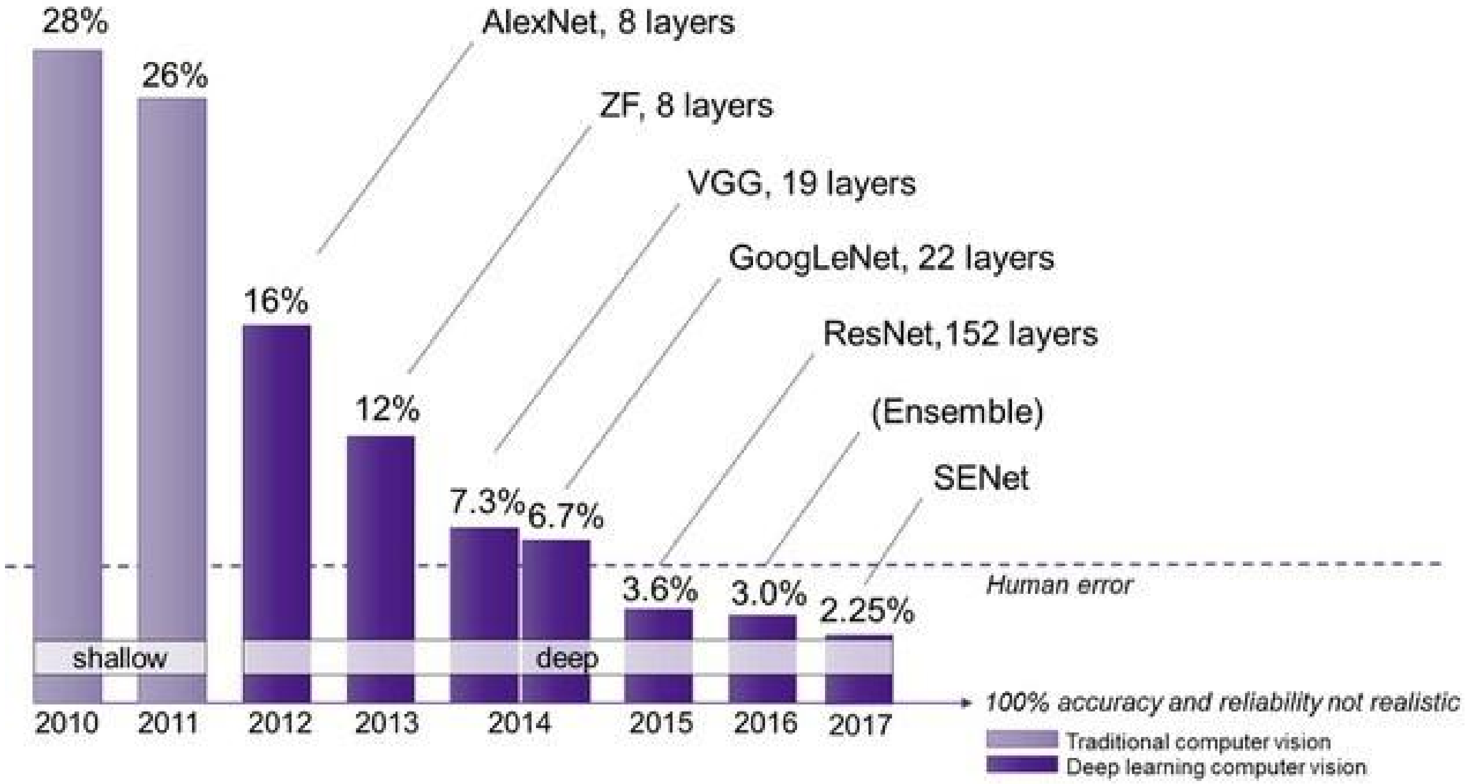
5. Learning to Steer by Mimicking Features from Heterogeneous Auxiliary Networks
AAAI Conference on Artificial Intelligence (AAAI, Oral), 2019, acceptance rate: 16.2% (1150/7095)
Yuenan Hou, Zheng Ma, Chunxiao Liu, Chen Change Loy

6. A Novel DDPG Method with Prioritized Experience Replay
IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2017
Yuenan Hou, Lifeng Liu, Qing Wei, Xudong Xu, Chunlin Chen

7. Categorical Relation-Preserving Contrastive Knowledge Distillation for Skin Lesion Classification
submitted to International Conference on Medical Image Computing and Computer Assisted Intervension (MICCAI), 2021
Xiaohan Xing, **Yuenan Hou**, Hang Li, Yixuan Yuan, Hongsheng Li, Max Q.-H. Meng

Motivation

Convolutional neural networks (CNNs) have achieved remarkable success in computer vision fields with an drastic increase in network complexity.



Motivation

The huge memory footprint and lengthy inference time have made these cumbersome networks prohibitive from deployment in many resource-limited mobile systems and edge devices.

Model	Parameter
LeNet-5	1 M
AlexNet	240 M
VGG-16	552 M

Motivation

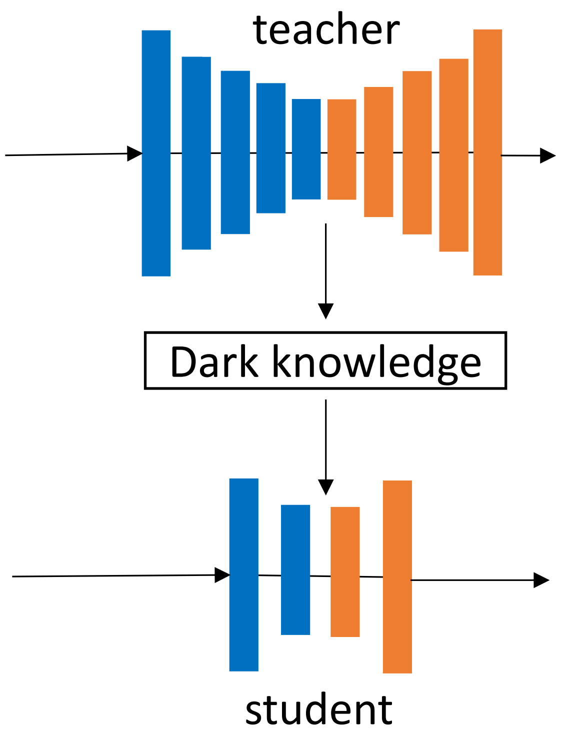
Model compression [Buciluǎ et al. 2006] is widely adopted to alleviate the demand of deep models on memory storage, and speed up the model inference without incurring severe performance degradation.

Main categories of model compression:

- Knowledge distillation [Hinton et al. 2015]
- Network pruning [Hassibi et al. 1992]
- Quantization [Zhou et al. 2017], low rank factorization [Sainath et al. 2013], …

Objective of Knowledge Distillation

Transfer the dark knowledge from the large cumbersome network (teacher) to the small compact model (student) so as to enhance the representation learning of the small model [Hinton et al. 2015]



Categories

1) Knowledge forms: class probability vectors [Hinton et al. 2015], feature maps [Romero et al. 2015], attention maps [Zagoruyko et al. 2017], inter-layer similarity maps [Yim et al. 2017], etc

2) Architectures: teacher and student can have similar/different architectures [Tian et al. 2020]

3) Distillation strategy: vanilla distillation [Hinton et al. 2015], selective / top-k distillation [Ge et al. 2019]

Challenges

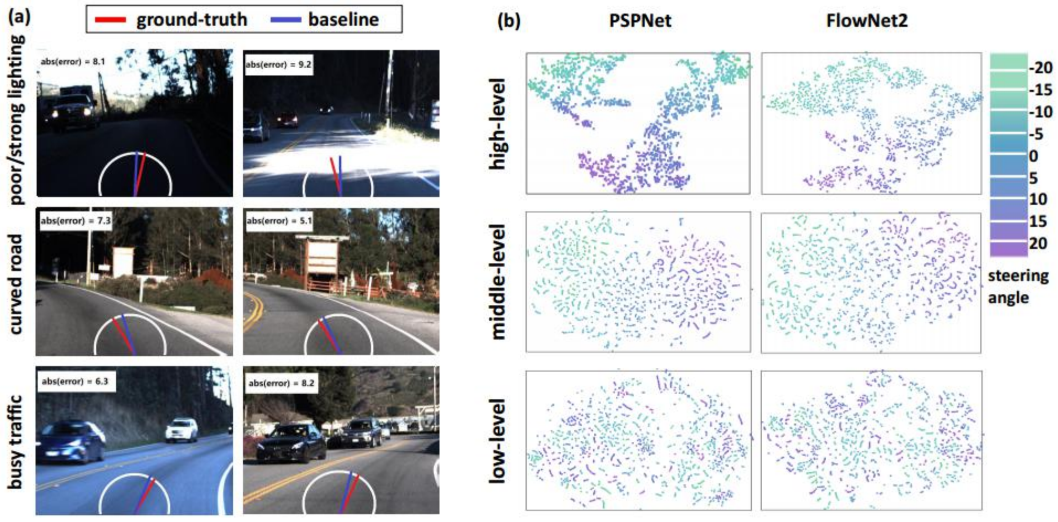
1) How to leverage rich contextual information when training signals are sparse? [Hou et al. 2019]

2) How to avoid the expensive training of the cumbersome teacher model? [Hou et al. 2019]

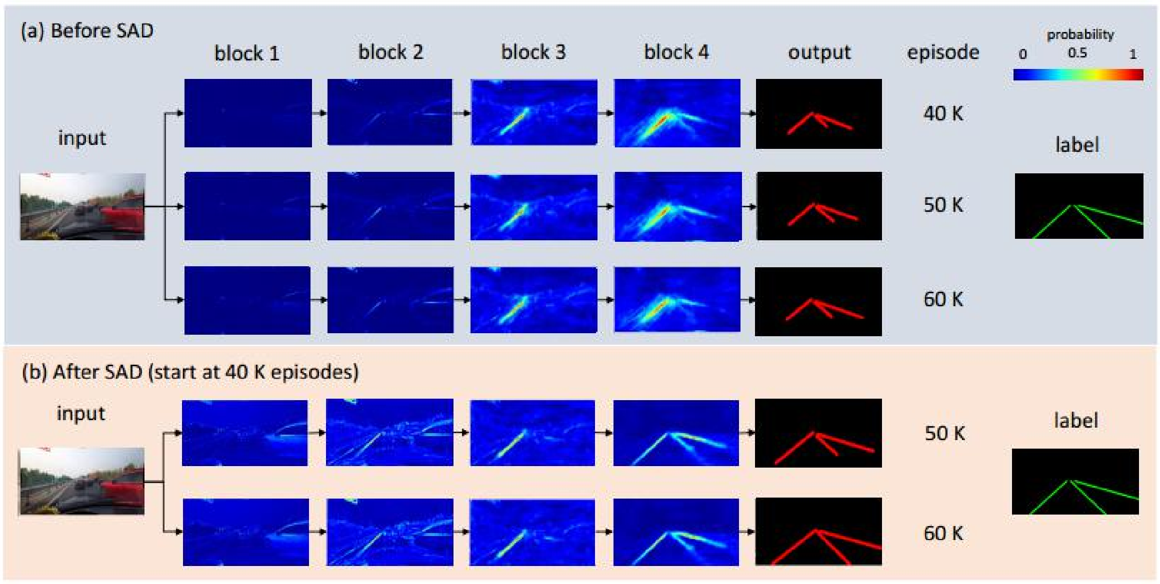
3) How to transfer the structural knowledge effectively in segmentation tasks? [Hou et al. 2020]

4) How to apply knowledge distillation to real-world tasks? [Hou et al. 2021]

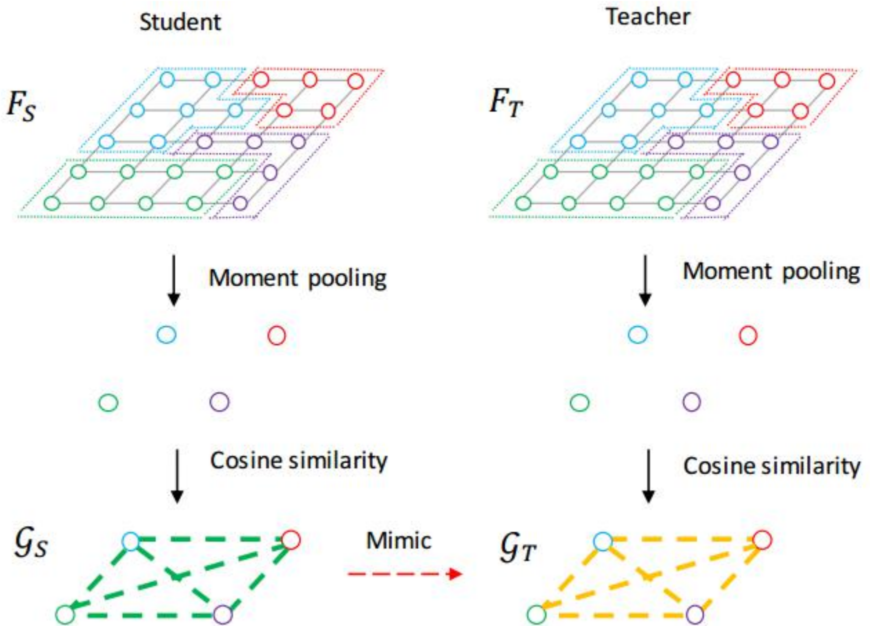
1) Heterogeneous auxiliary network feature mimicking AAAI 2019 Oral



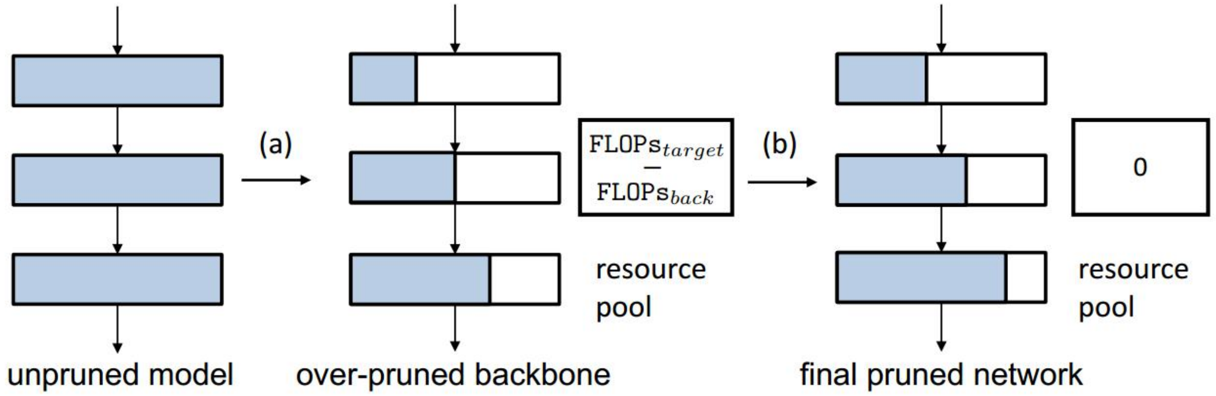
2) Self attention distillation, ICCV 2019



3) Inter-region affinity distillation, CVPR 2020



4) Network pruning via resource reallocation submitted to ICCV 2021

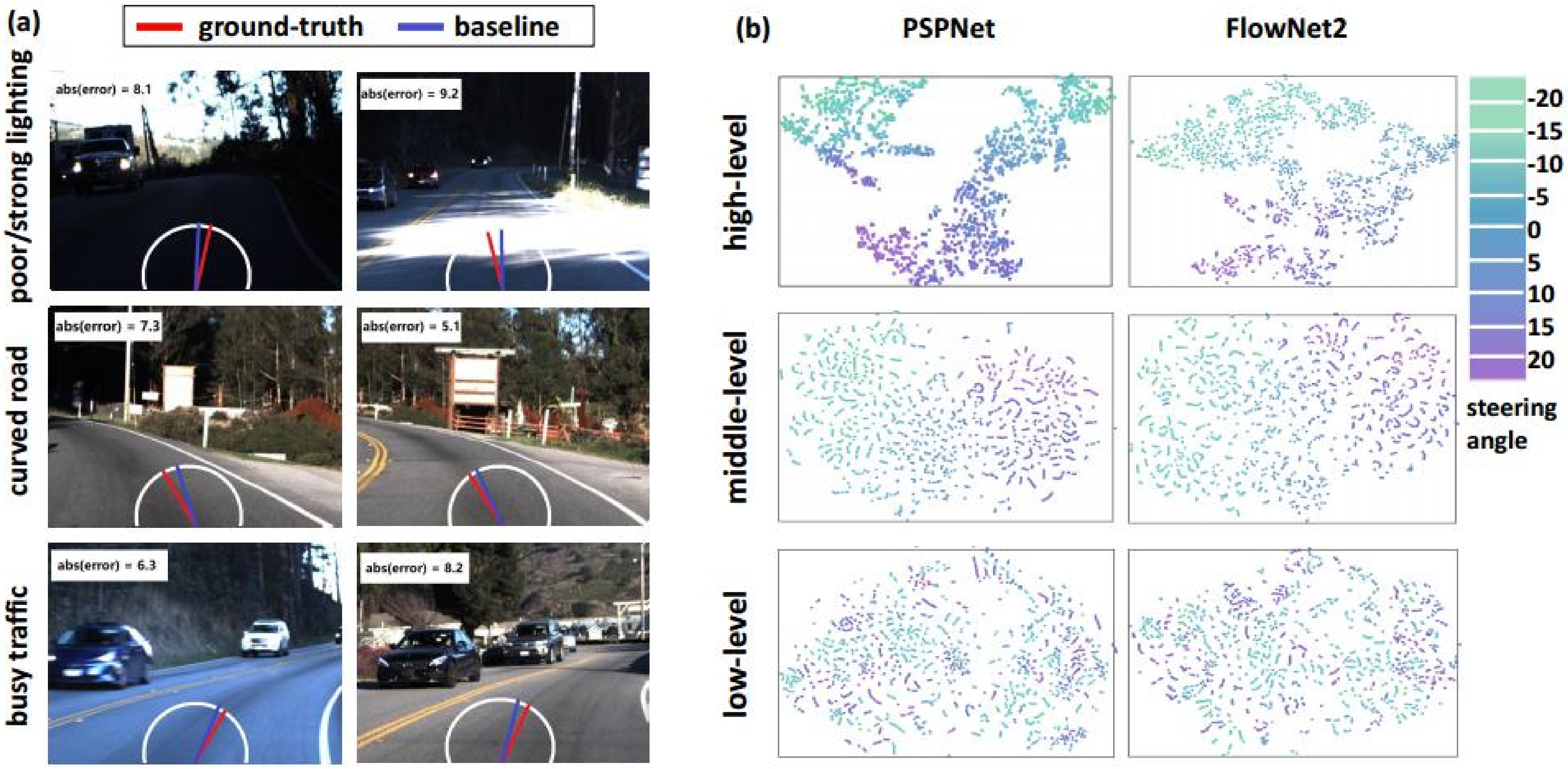


Heterogeneous Auxiliary Network Feature Mimicking

Motivation:

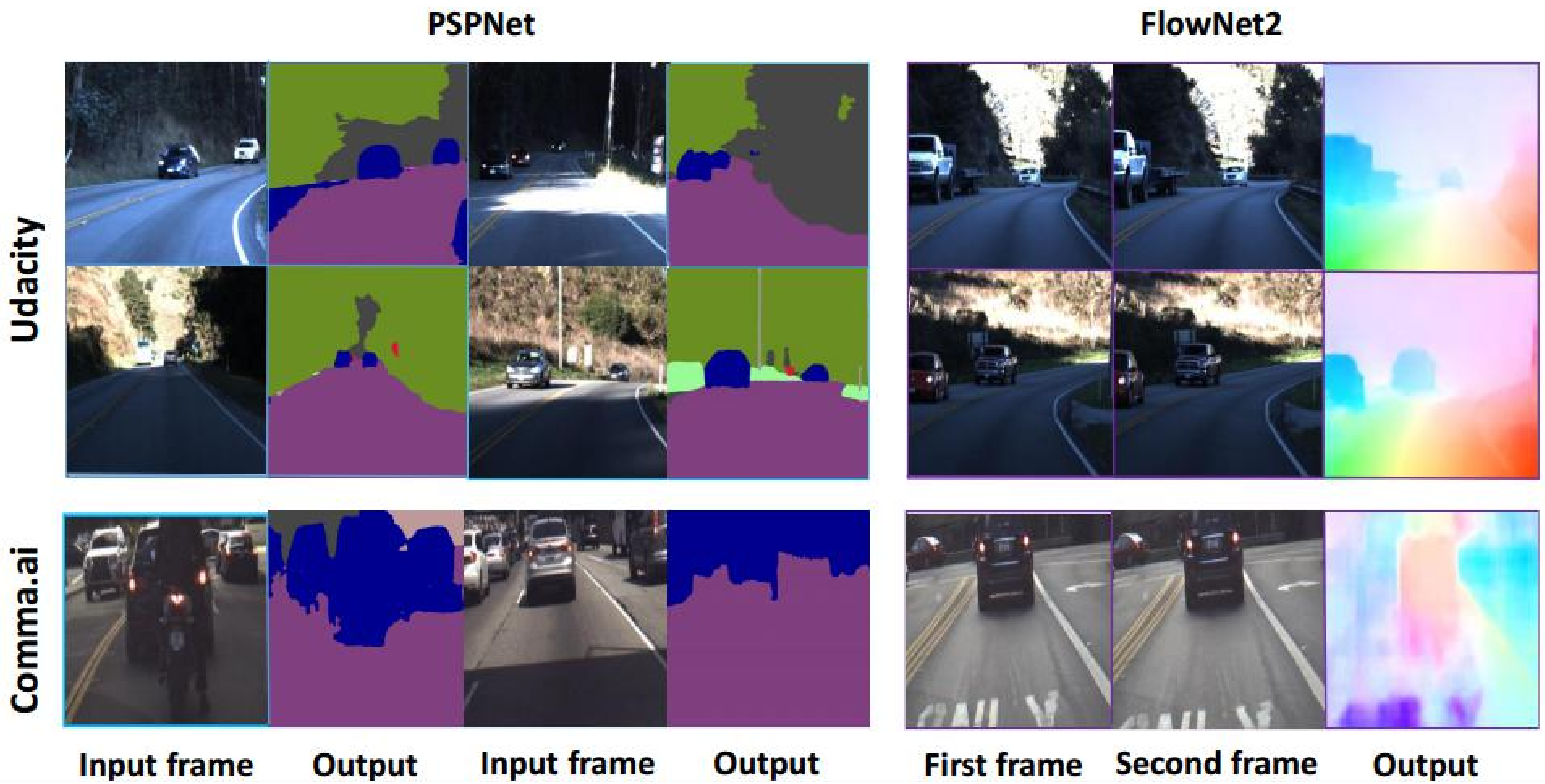
- Learning rich environmental contexts, e.g., physical scene constraints or coexistence of scene objects
- No requirement of additional and expensive annotations
- Not affecting the running efficiency

Challenges and Observations



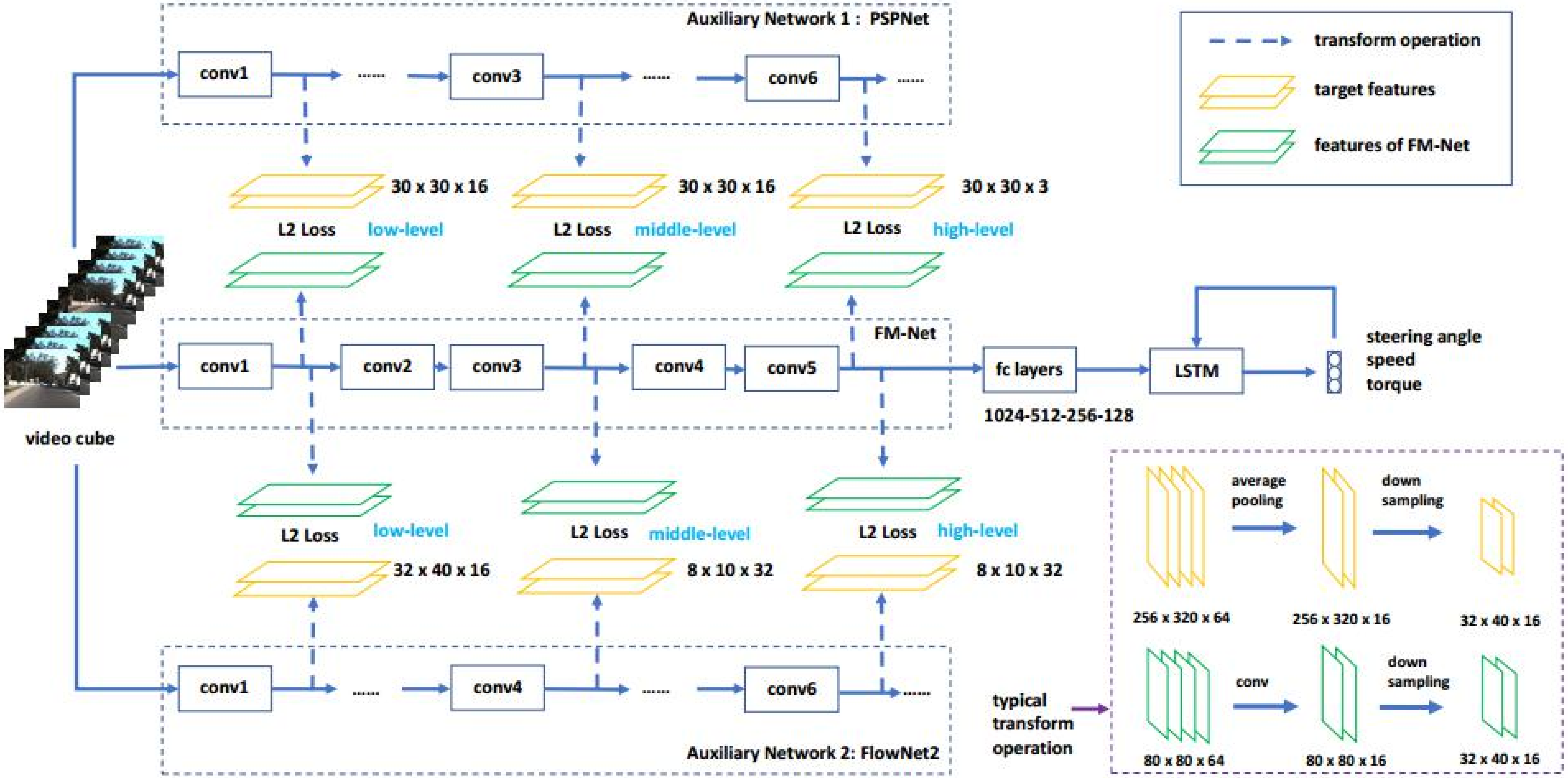
Features of auxiliary networks are highly correlated with steering angles

Qualitative results of auxiliary networks



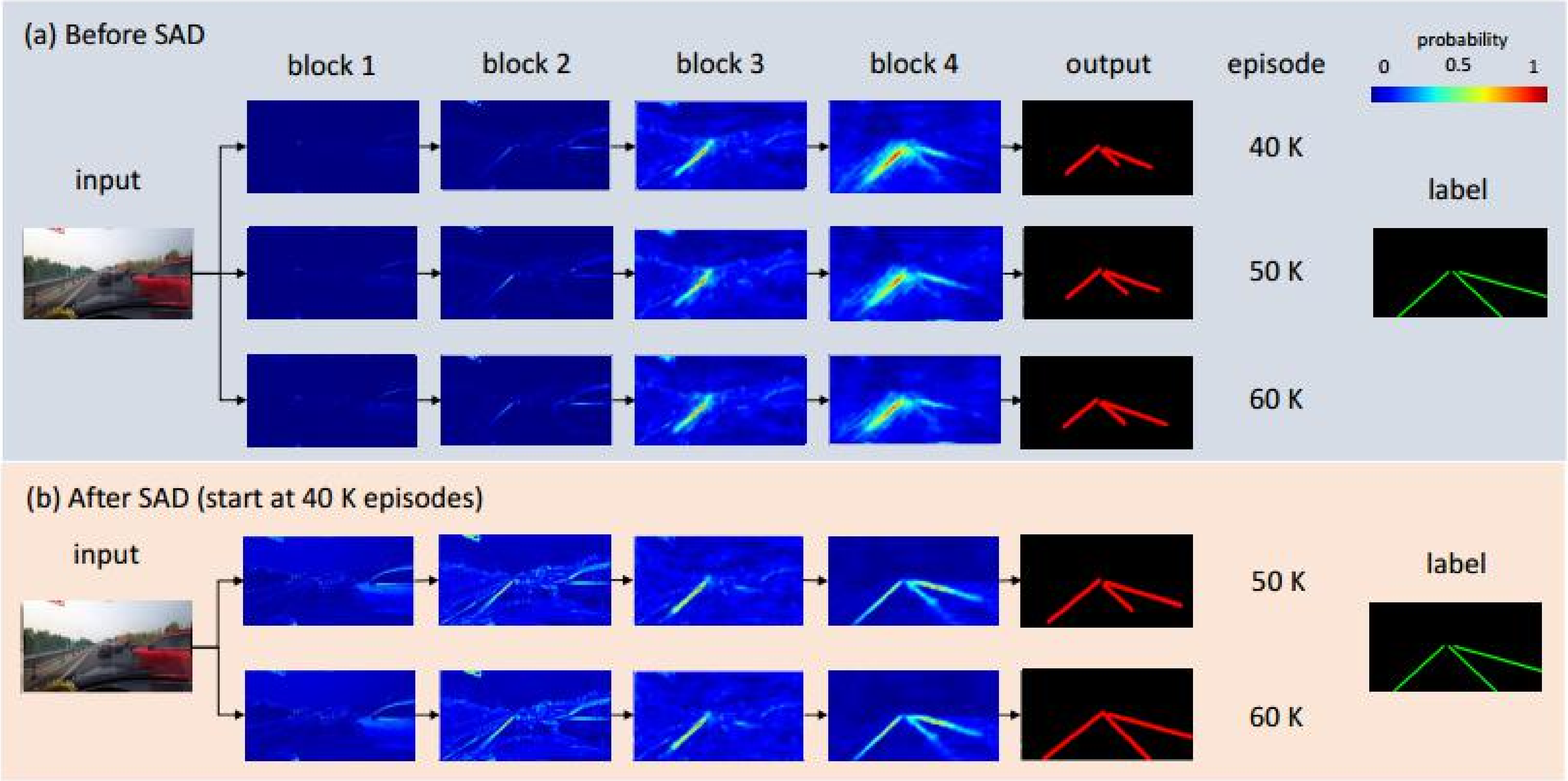
Auxiliary networks show good generalization on unseen target data

Framework overview



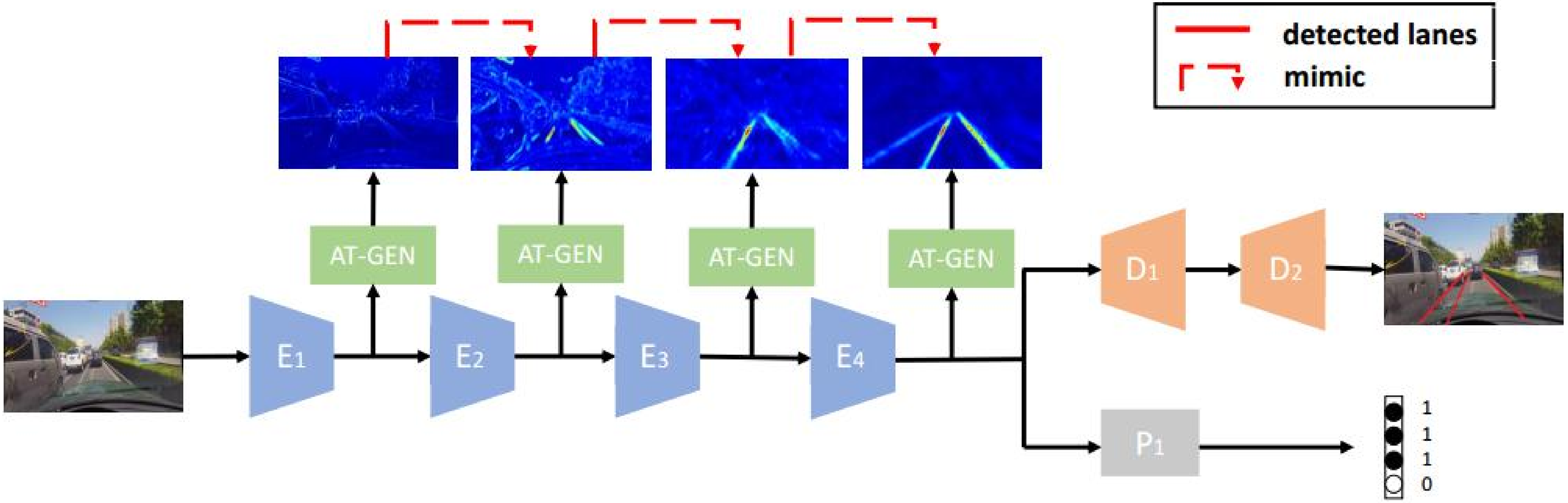
Overview of our main framework

Self Attention Distillation



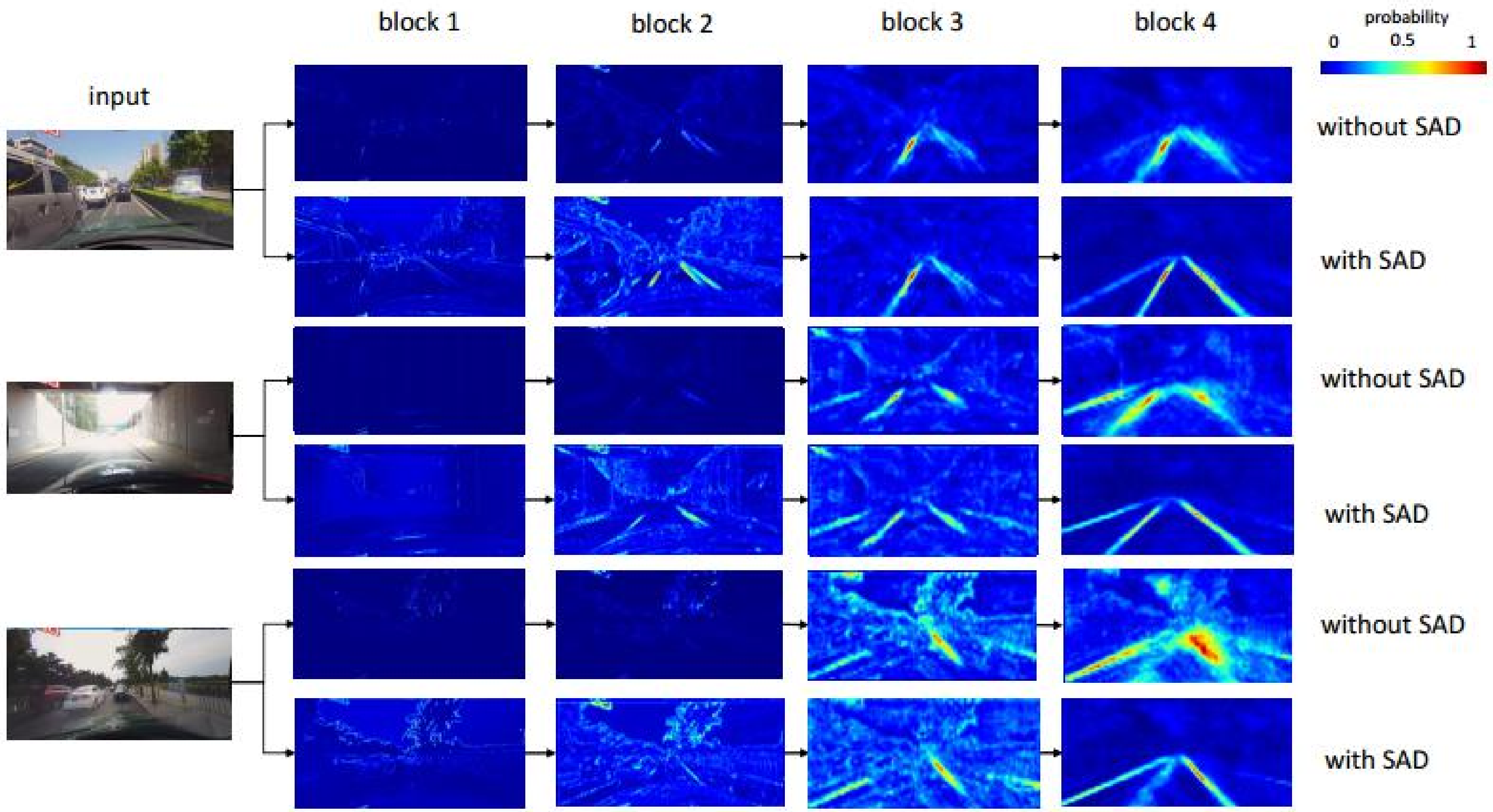
Attention maps derived from different layers of a well-trained network capture diverse and rich contextual information that hints the lane locations and a rough outline of the scene

Framework overview



Overview of our main framework

Visualization of attention maps

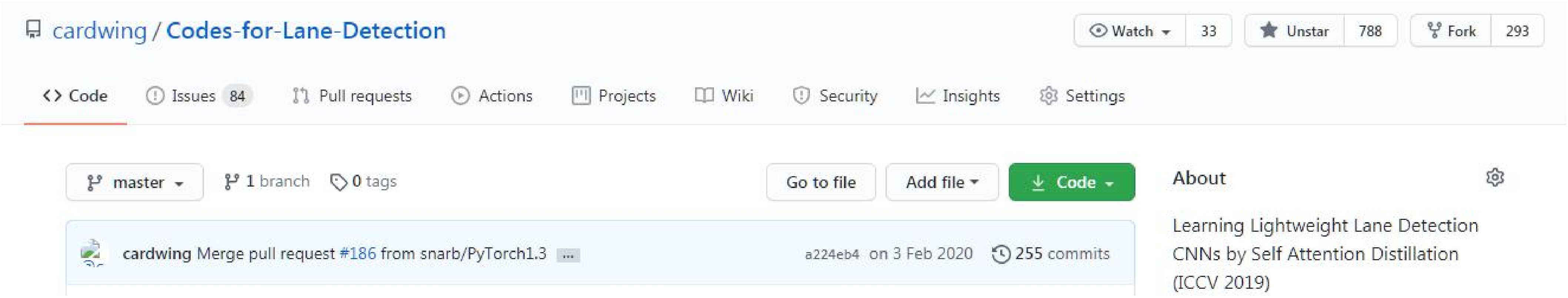


After adding SAD, attention maps of the trained network become more concentrated on task-relevant objects, e.g., lanes, vehicles and road curbs.

Lane detection codebase

Release a lightweight and high-performance codebase:

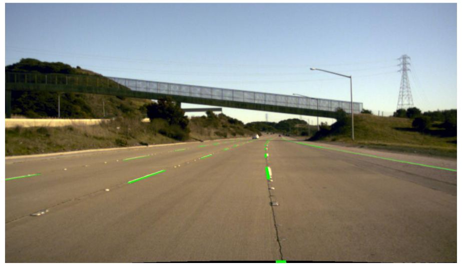
- Has near 800 stars and 300 forks
- Widely used by prestigious academic and industrial community, e.g., MIT, CMU and Huawei
- Rank 2nd in the most popular lane detection code in paperswithcode



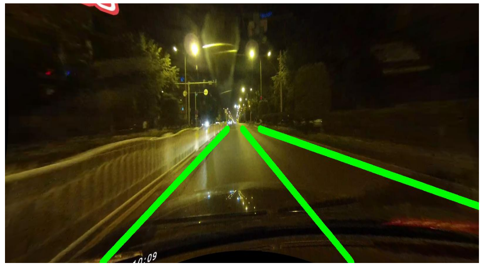
Inter-Region Affinity Distillation

- Challenges from road marking segmentation task:

ground-truth



tiny road elements



poor lighting

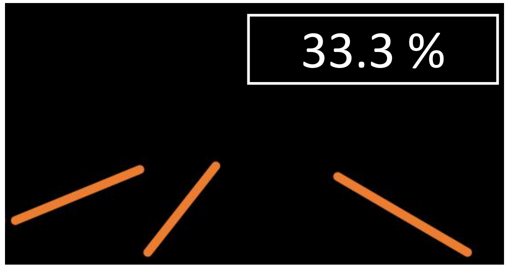
occlusions by vehicles

sparsity

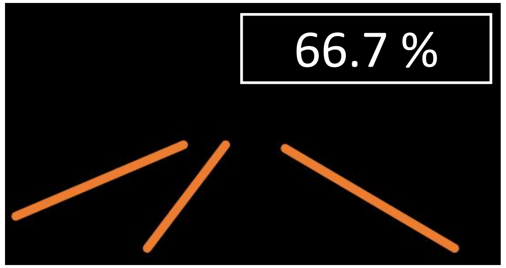
- Challenge becomes crippling when we need a small model for autonomous driving
  - Existing KD methods are effective in many classification tasks
  - Fall short in road marking segmentation

input

ERFNet [Romera et al. 2017]



ERFNet-BiFPN [Zhu et al. 2018]



→ F1

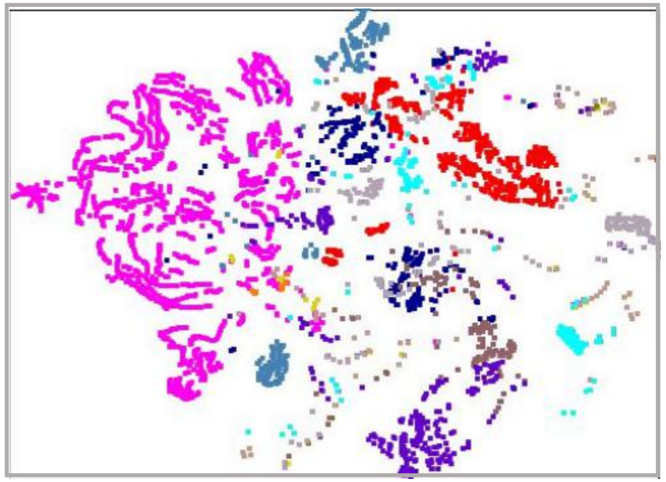
Motivation

- A road scene typically exhibits consistent configuration, i.e. road elements are orderly distributed in a scene. The structural relationship is crucial to provide the necessary constraint or regularization, especially for small networks, to combat against the sparsity of supervision.

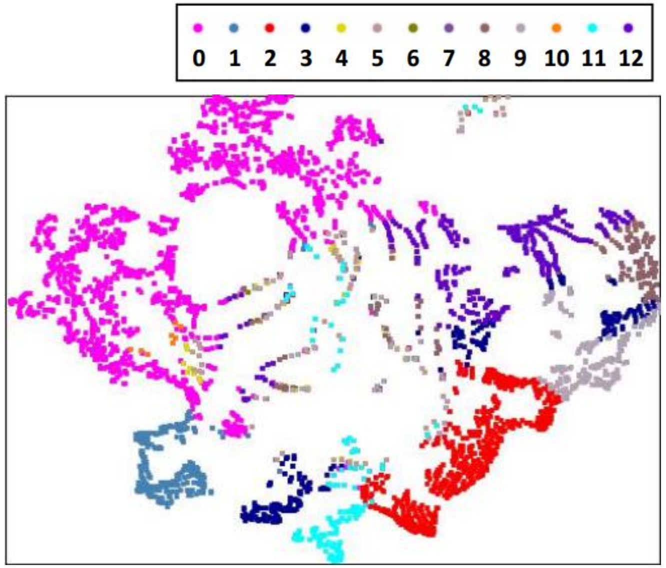
- Feature distribution relationships encoded by teacher on different parts of a deep feature map reveal rich structural connections between different regions.

Visualization of deep feature embeddings

input

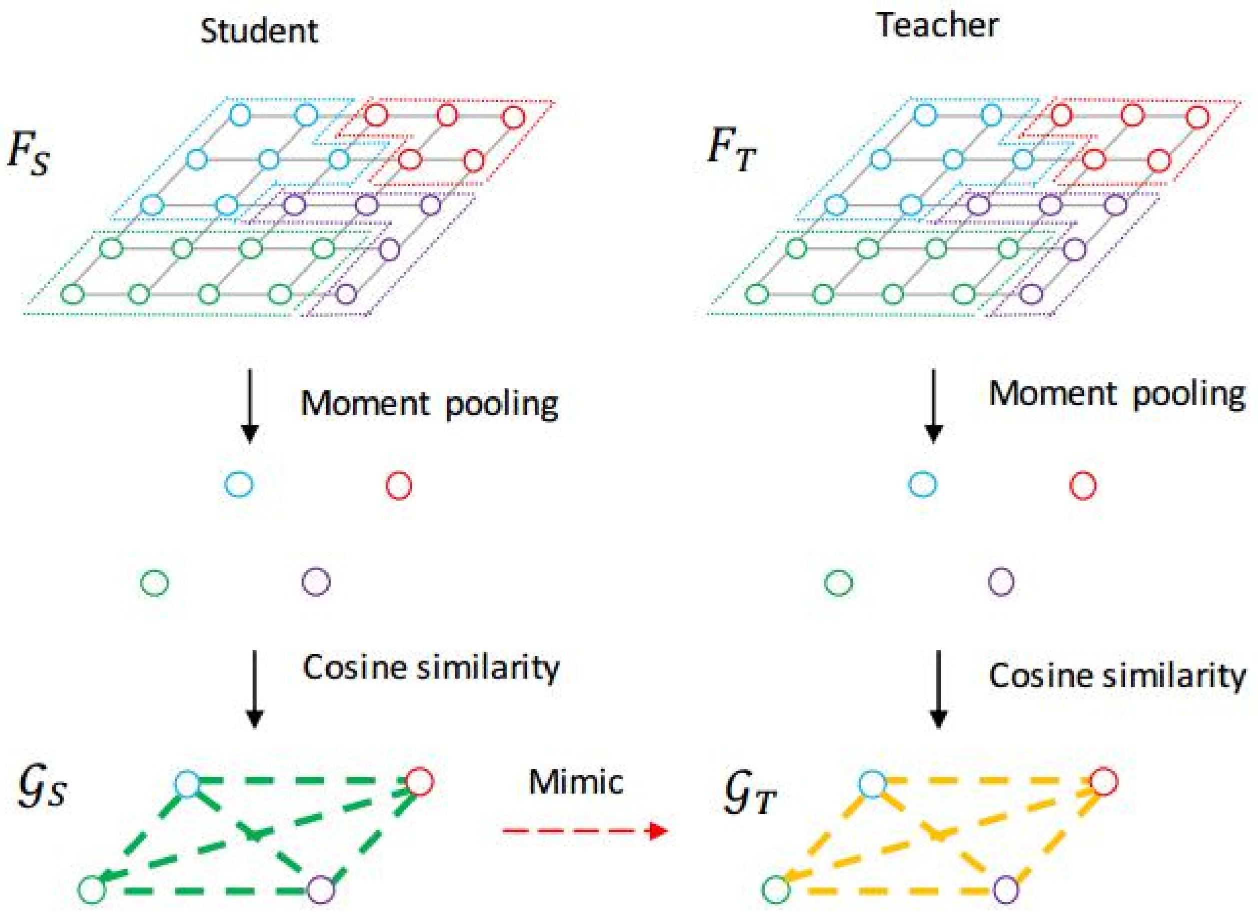


ERFNet (student)



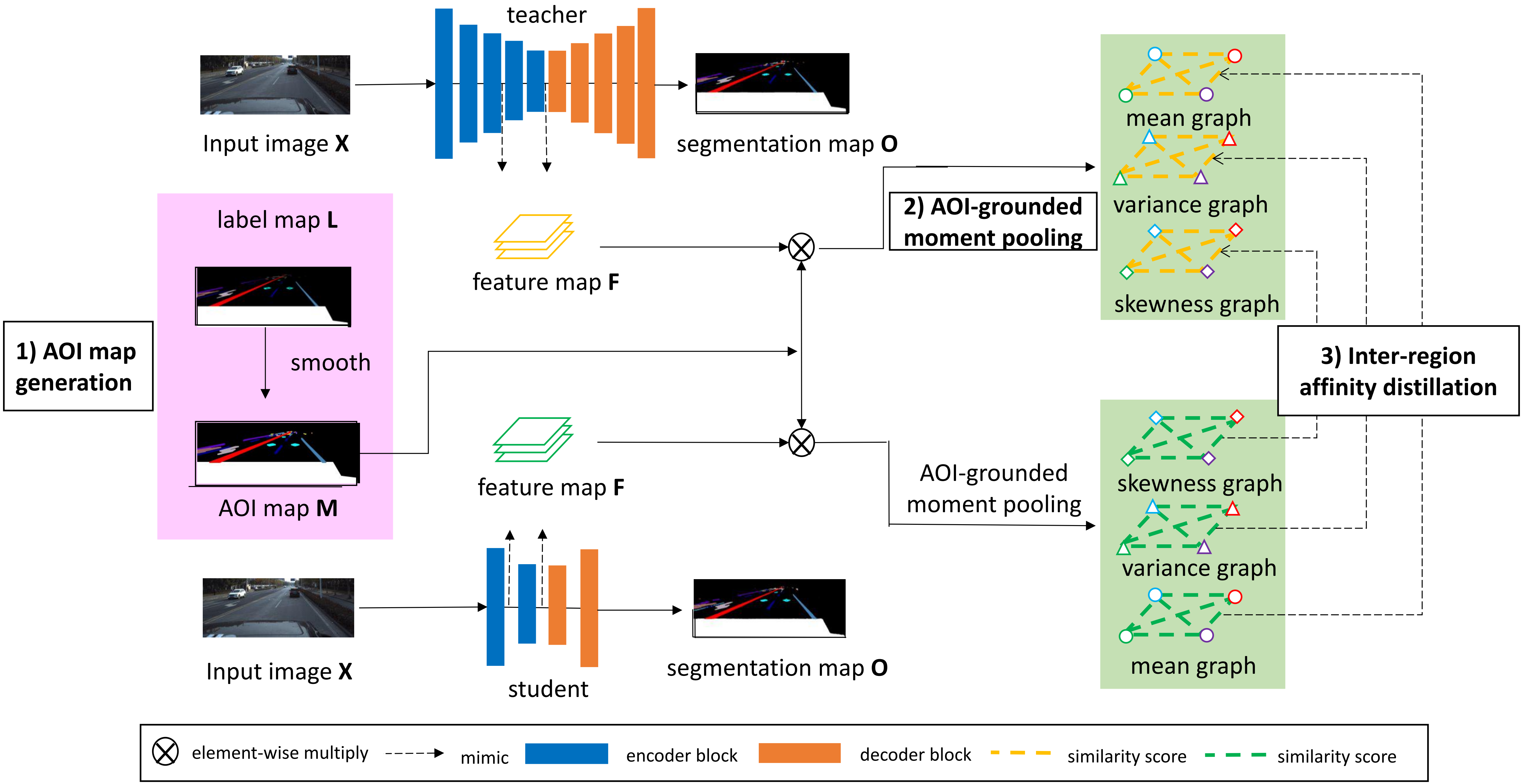
ResNet-101 (teacher)

Method

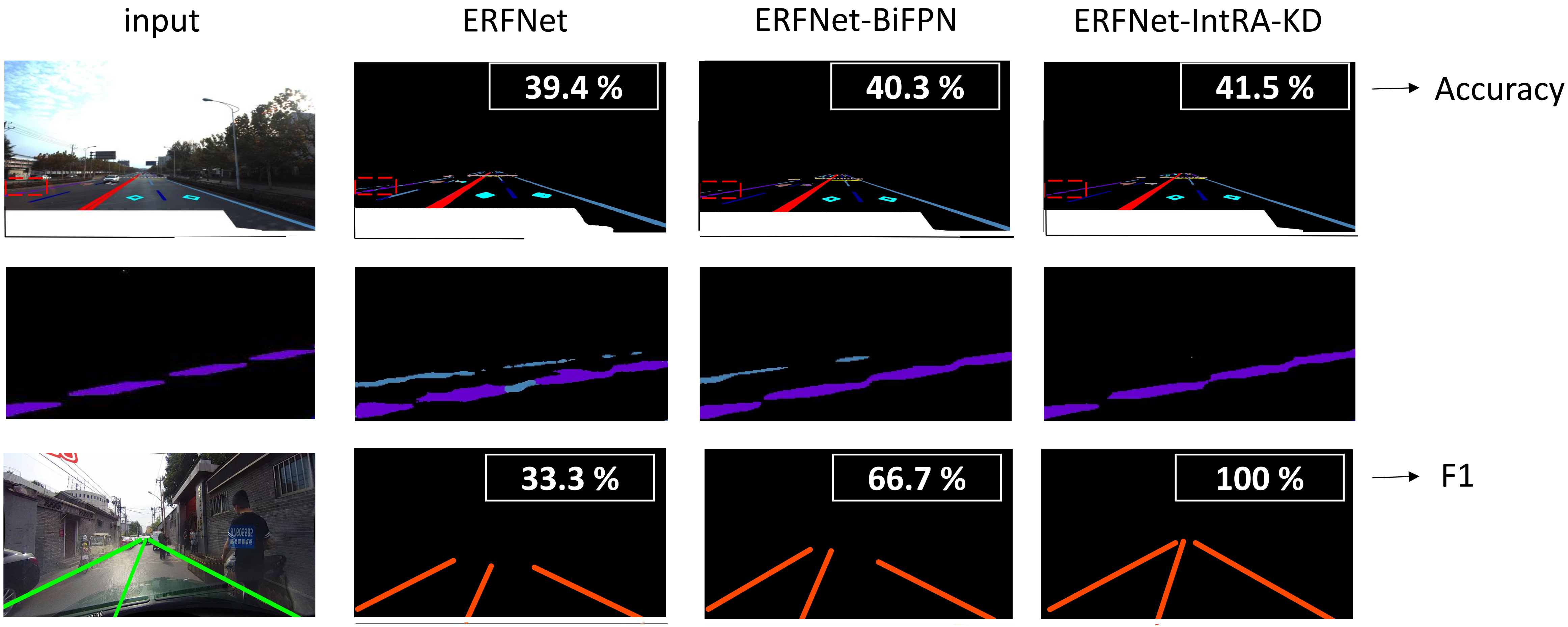


Knowledge on scene structure is represented as inter-region affinity graphs. Through graph matching, a distillation loss on graph consistency is generated to update the student network.

Pipeline of Inter-Region Affinity Knowledge Distillation

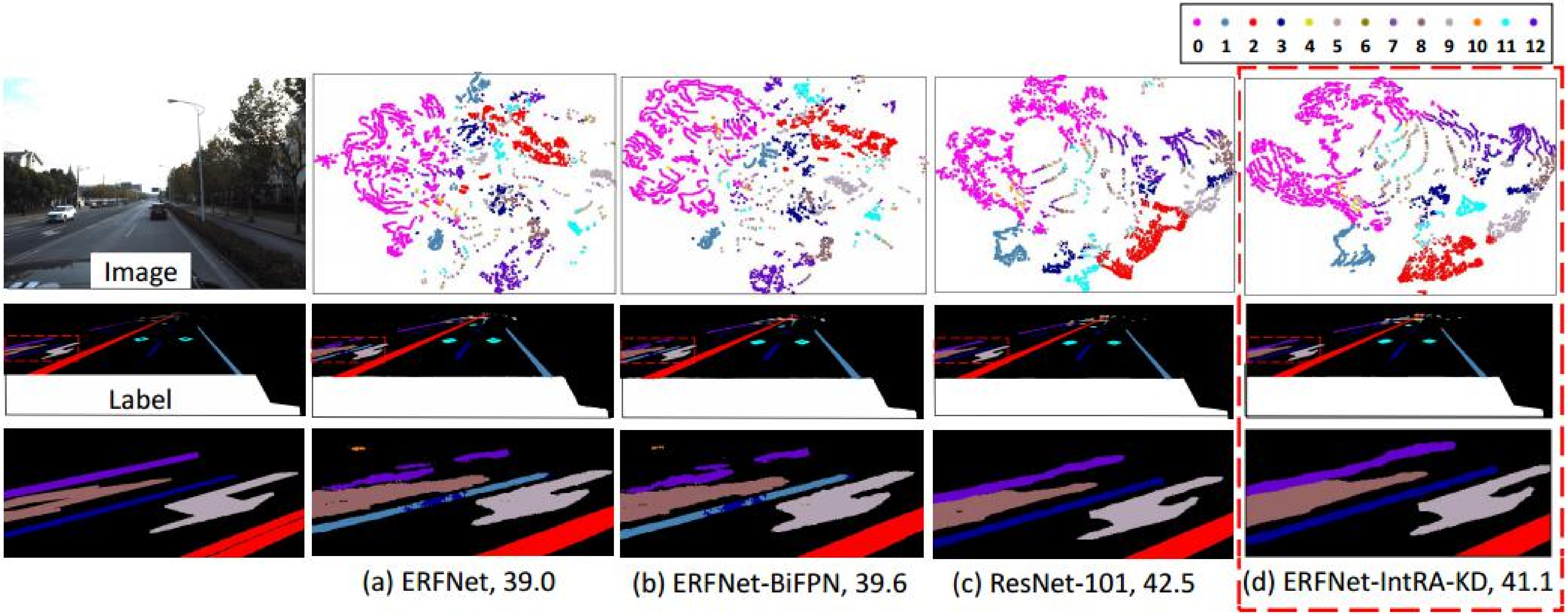


Qualitative Results



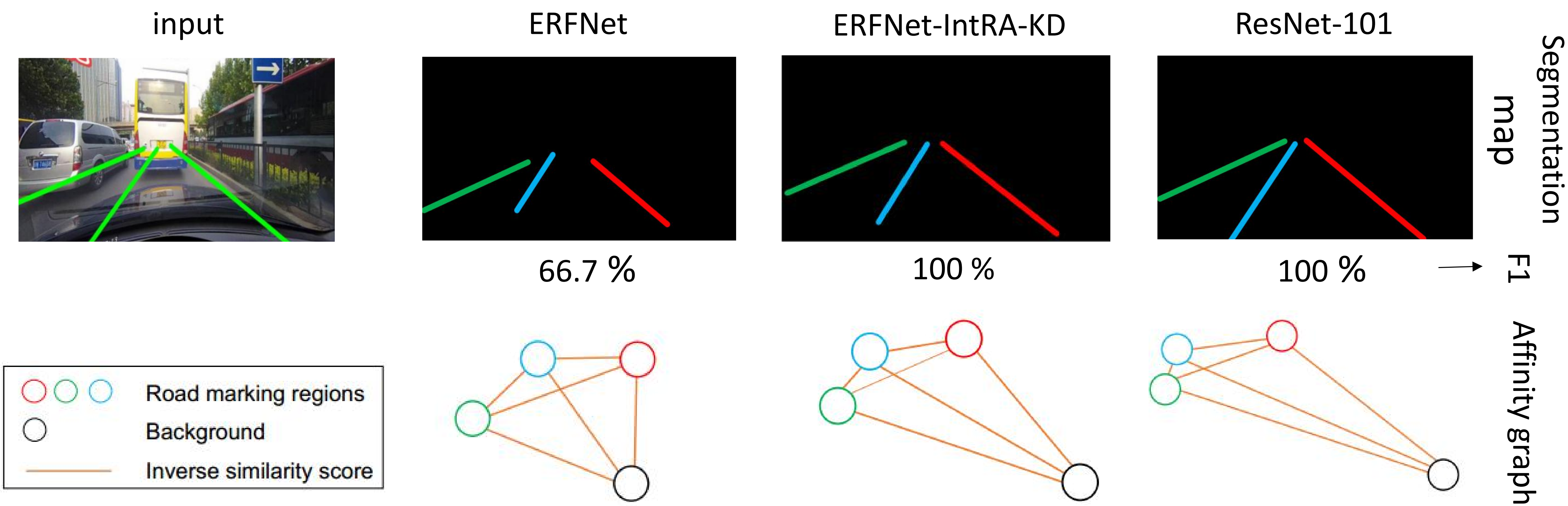
- IntRA-KD makes predictions of long, thin lanes smoother.
- ERFNet-IntRA-KD makes more accurate predictions under severe occlusion.

Qualitative Results



IntRA-KD makes feature embeddings of different classes more distinctly clustered.

Visualization of the Affinity Graph



IntRA-KD not only improves the predictions of ERFNet, but also causes a closer feature structure between the student model and the ResNet-101 teacher model.

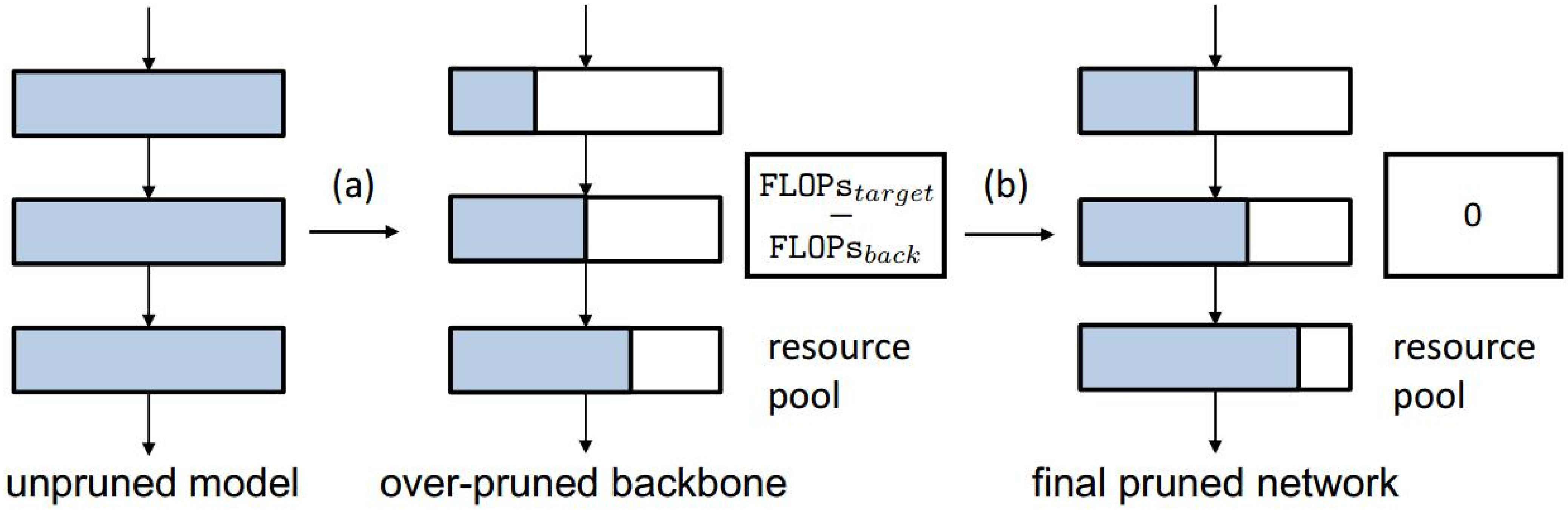
Network Pruning via Resource Reallocation

Motivation:

- Contemporary pruning approaches perform Iterative pruning procedure from the original over-parameterized model, which is both tedious and expensive, especially when the pruning is aggressive

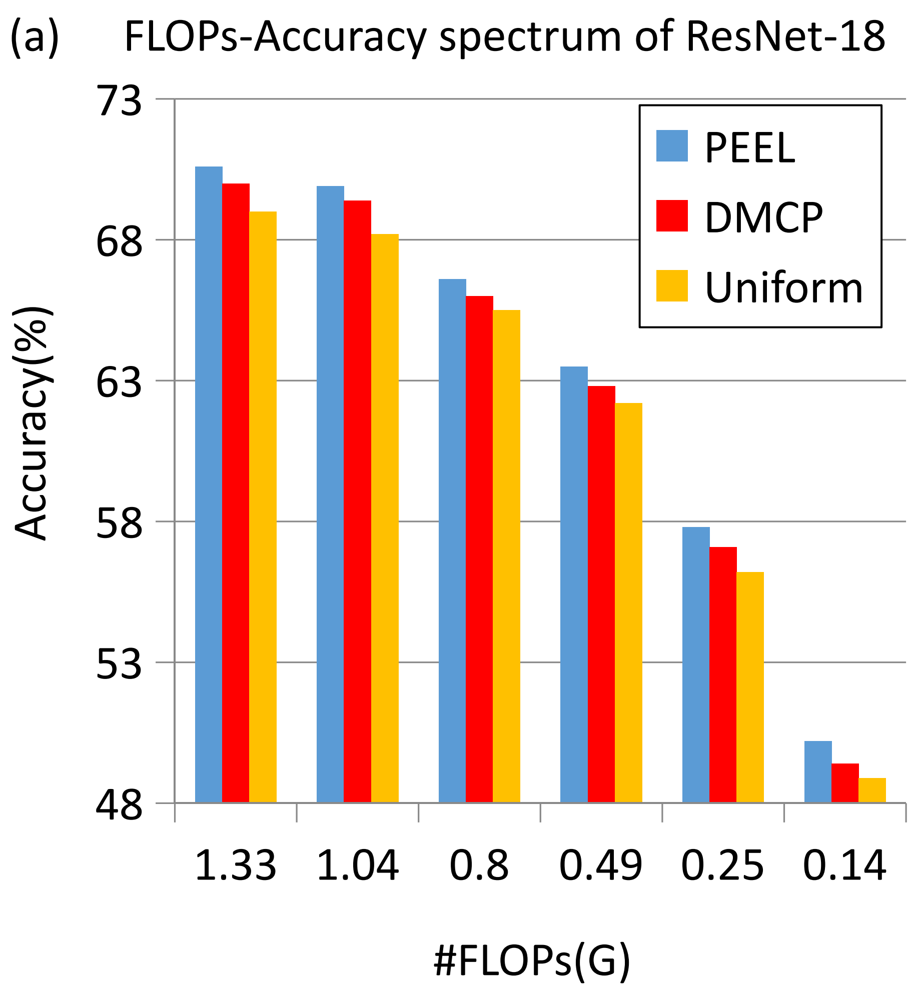
- Previous methods typically ignore the value of the original cumbersome model

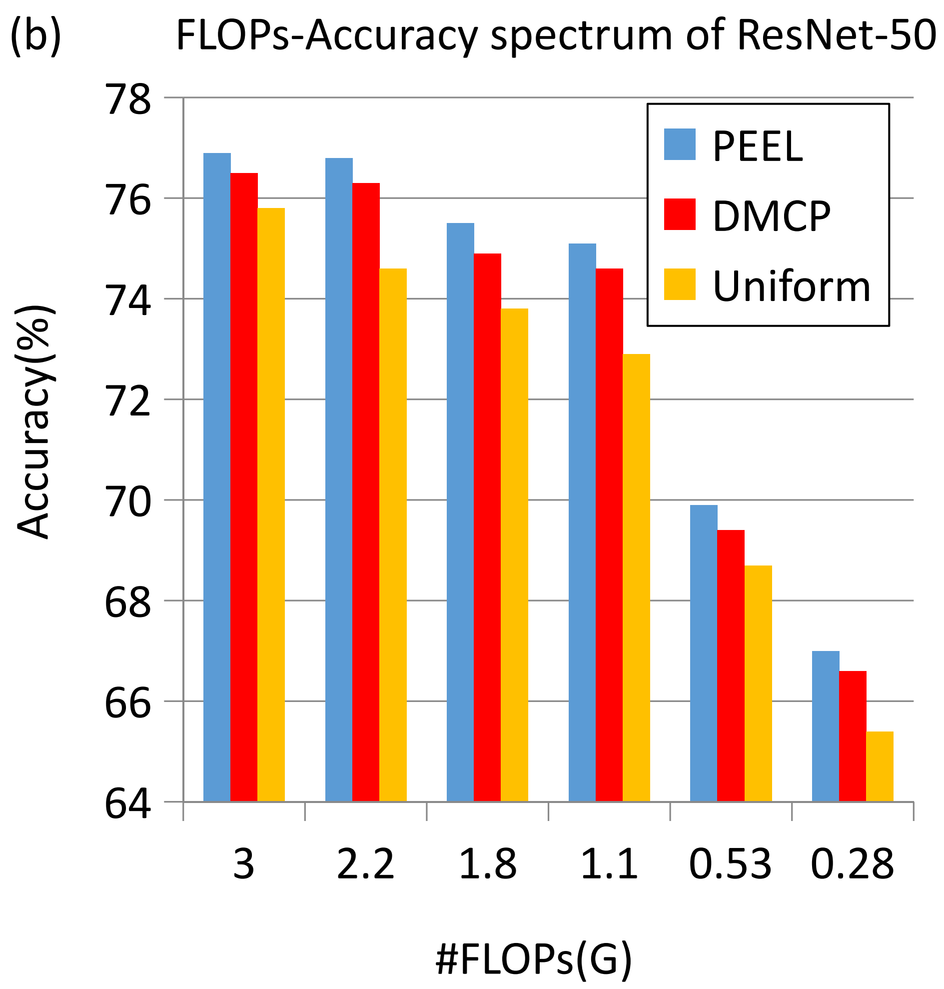
Network Pruning via Resource Reallocation



Our method is comprised of three components, i.e., constructing an over-pruned backbone model, estimating layer importance and reallocating resources, and performing knowledge distillation.

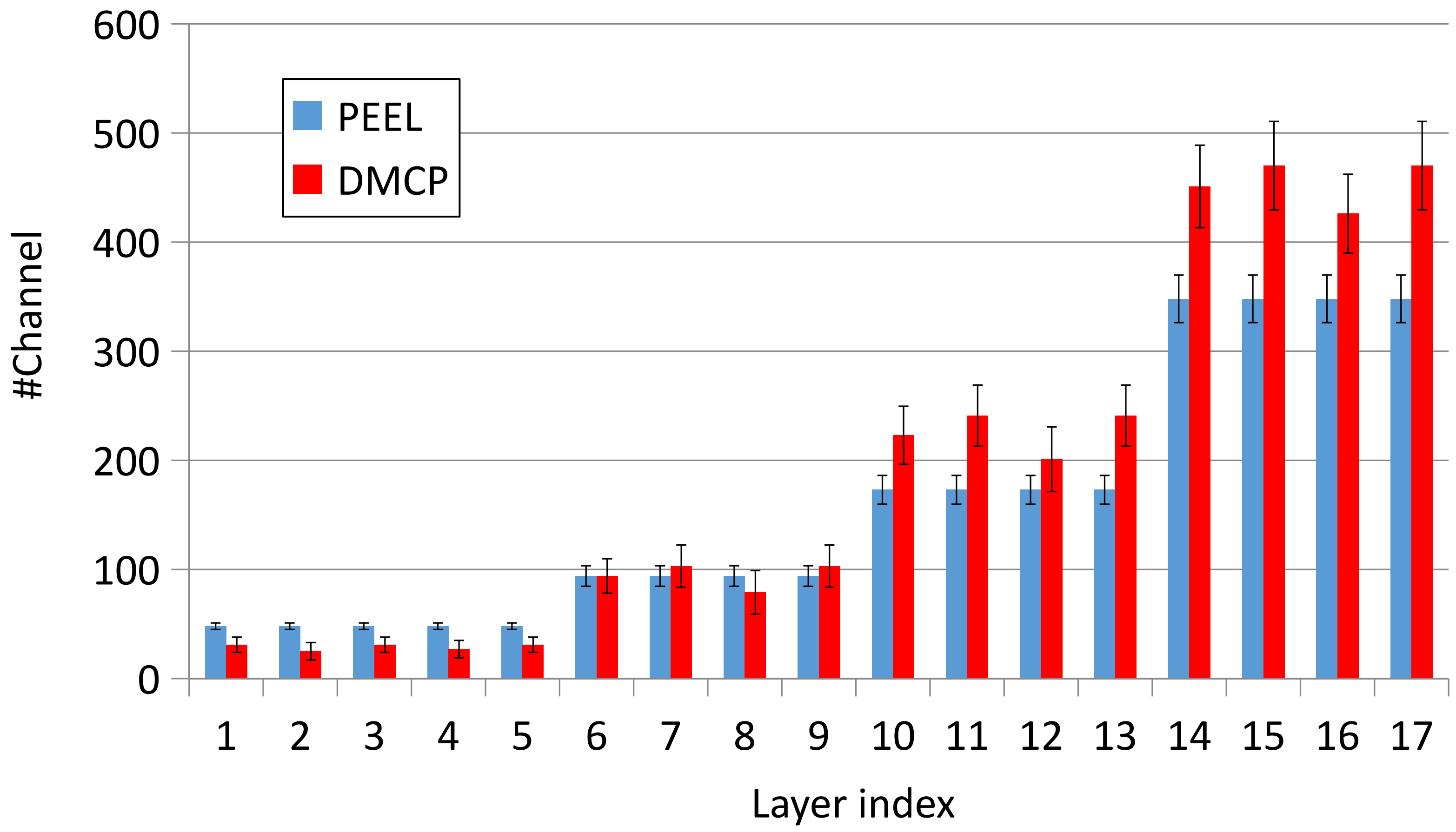
Qualitative results





Our method, i.e., PEEL, consistently finds better architectures whose performance outperforms those searched by DMCP [Guo et al. 2020] and uniform pruning [Liu et al. 2019].

Searching stability



The variance of the number of channels in each layer of PEEL is much smaller than that of DMCP.

Future work

Task:

- Compression of GANs [Hou et al, 2021], 3D detection/segmentation networks, etc

Efficiency:

- Combining model compression with active learning, few shot learning, self supervised learning, etc

Theory:

- Theoretical understanding about why KD works, e.g., regularization or providing inter-class similarity information

References

[1] Hinton, G., Vinyals, O. and Dean, J.. Distilling the knowledge in a neural network. In NIPS workshop, 2015.

[2] Buciluǎ, C., Caruana, R. and Niculescu-Mizil, A.. Model compression. In KDD, 2006.

[3] Hassibi, B. and Stork, D.. Second order derivatives for network pruning: Optimal brain surgeon. In NIPS, 1992.

[4] Zhou, A., Yao, A., Guo, Y., Xu, L. and Chen, Y.. Incremental network quantization: Towards lossless cnns with low-precision weights. In ICLR, 2017.

[5] Sainath, T.N., Kingsbury, B., Sindhwani, V., Arisoy, E. and Ramabhadran, B.. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In ICASSP, 2013.

[6] Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C. and Bengio, Y.. Fitnets: Hints for thin deep nets. In ICLR, 2015.

[7] Zagoruyko, S. and Komodakis, N.. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In ICLR, 2017.

[8] Yim, J., Joo, D., Bae, J. and Kim, J.. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In CVPR, 2017.

[9] Tian, Y., Krishnan, D. and Isola, P.. Contrastive representation distillation. In ICLR, 2020.

[10] Ge, S., Zhao, S., Li, C. and Li, J.. Low-resolution face recognition in the wild via selective knowledge distillation. In TIP, 2019.

[11] Romera, E., Alvarez, J.M., Bergasa, L.M. and Arroyo, R.. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. In T-ITS, 2017.

[12] Zhu, L., Deng, Z., Hu, X., Fu, C.W., Xu, X., Qin, J. and Heng, P.A.. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In ECCV, 2018.

[13] Guo, S., Wang, Y., Li, Q. and Yan, J.. DMCP: Differentiable Markov Channel Pruning for Neural Networks. In CVPR, 2020.

[14] Liu, Z., Sun, M., Zhou, T., Huang, G. and Darrell, T.. Rethinking the value of network pruning. In ICLR, 2019.

[15] Hou, Y., Ma, Z., Liu, C. and Loy, C.C.. Learning to steer by mimicking features from heterogeneous auxiliary networks. In AAAI, 2019.

[16] Hou, Y., Ma, Z., Liu, C. and Loy, C.C.. Learning lightweight lane detection cnns by self attention distillation. In ICCV, 2019.

[17] Hou, Y., Ma, Z., Liu, C., Hui, T.W. and Loy, C.C.. Inter-Region Affinity Distillation for Road Marking Segmentation. In CVPR, 2020.

[18] Hou, Y., Zhu, X. and Loy, C.C., Patchwise Contrastive Distillation for Generative Adversarial Networks, submitted to ICCV, 2021.

[19] Hou, Y., Ma, Z., Liu, C., Wang, Z. and Loy, C.C., Network Pruning via Resource Reallocation, submitted to ICCV, 2021.