Cars Can’t Fly up in the Sky: Improving Urban-Scene Segmentation via Height-driven Attention Networks

Sungha Choi, Joanne T. Kim, Jaegul Choo

LG Electronics, Korea University, KAIST

CVPR 2020

Presenter: Minho Park

Motivation

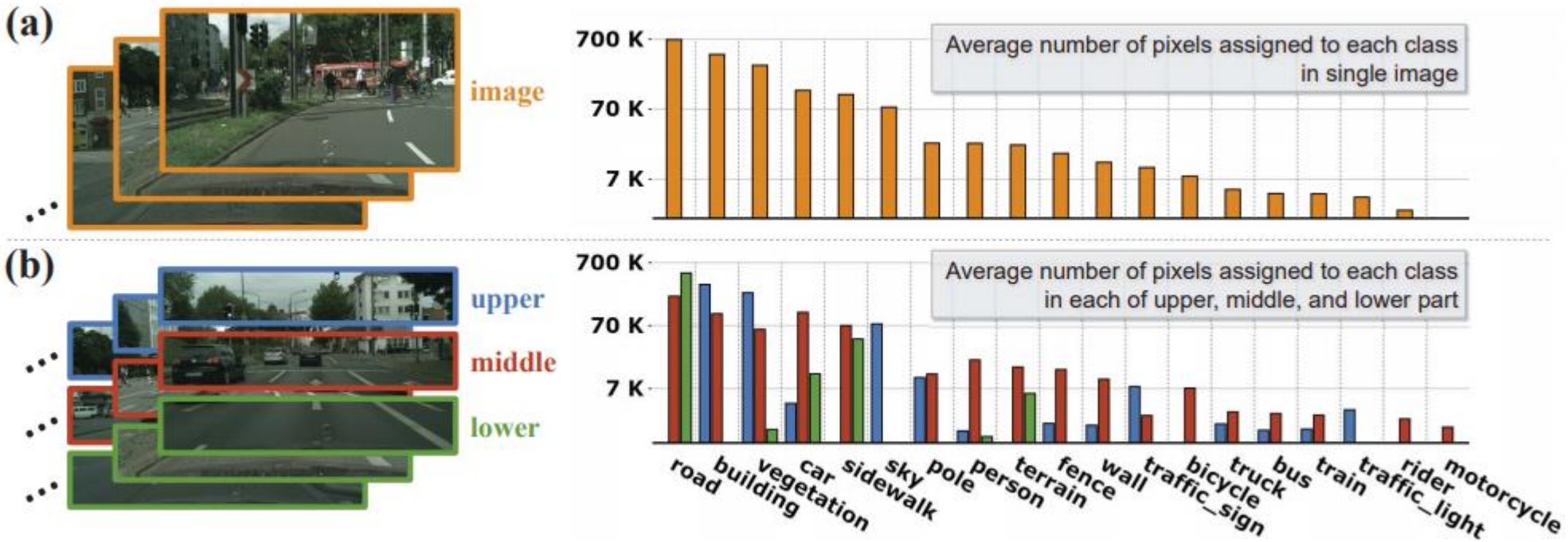

Figure 1: Motivation of our approach, the pixel-wise class distributions.

Motivation

Uncertainty is reduced

Given	Probabilities of top-5 classes 88%					Entire class entropy		
	p_{road}	$p_{\text{build.}}$	p_{veg}	p_{car}	p_{swalk}			
Image	36.9	22.8	15.9	7.0	6.1	1.84		$H(X)$
Upper	0.006	47.8	35.4	0.6	0.009	1.24	1.26 (avg)	$H(X\|\text{upper})$
Middle	31.4	16.6	9.4	17.4	10.7	2.04		$H(X\|\text{middle})$
Lower	87.9	0.1	0.3	2.2	7.9	0.51		$H(X\|\text{lower})$

Table 1: Comparison of the probability distributions (%) of pixels being assigned to each class when an entire image is separated on upper, middle, and lower regions of the Cityscapes dataset.

SENet

- Explicitly modelling the interdependencies between the channels of its convolutional features

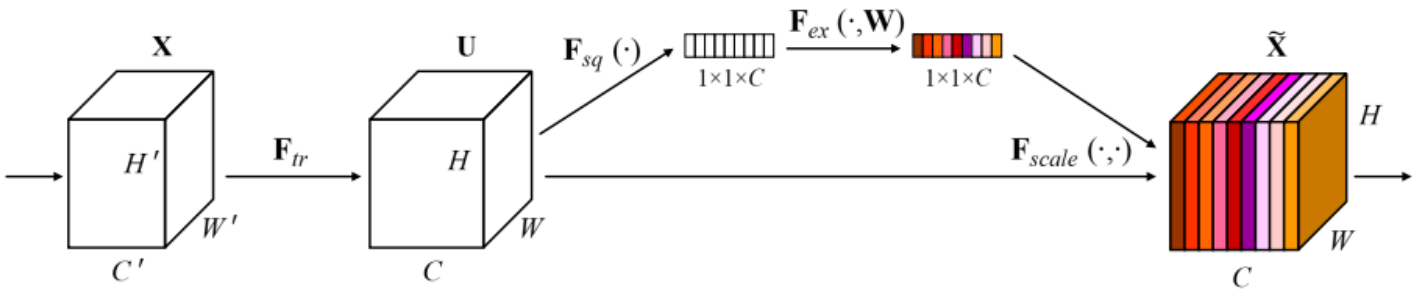

Fig. 1. A Squeeze-and-Excitation block.

Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." CVPR. 2018.

SENet

- Squeeze-and-excitation networks capture the global context of the entire image using global average pooling and predict per-channel scaling factors.

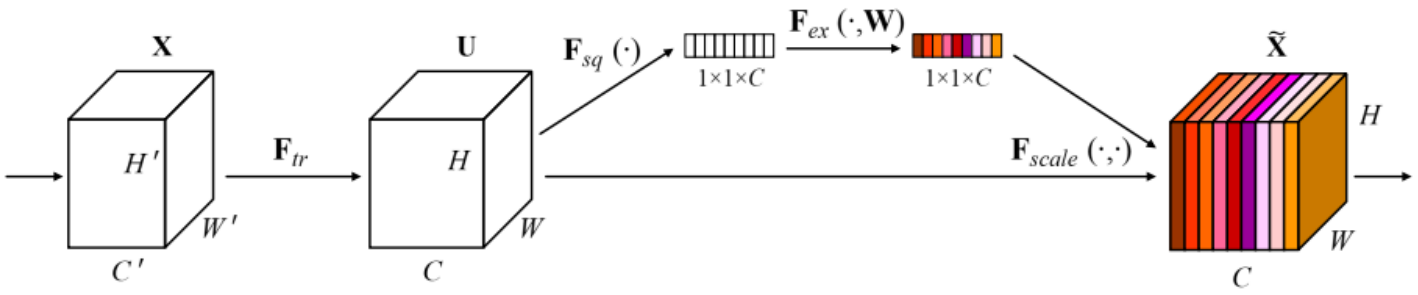

Fig. 1. A Squeeze-and-Excitation block.

Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." CVPR. 2018.

SENet

- However, the urban-scene datasets consist only of road driving pictures, which means that the images share similar class statistics.
- **The global context of the entire image cannot help per-pixel classification in urban-scene images.**

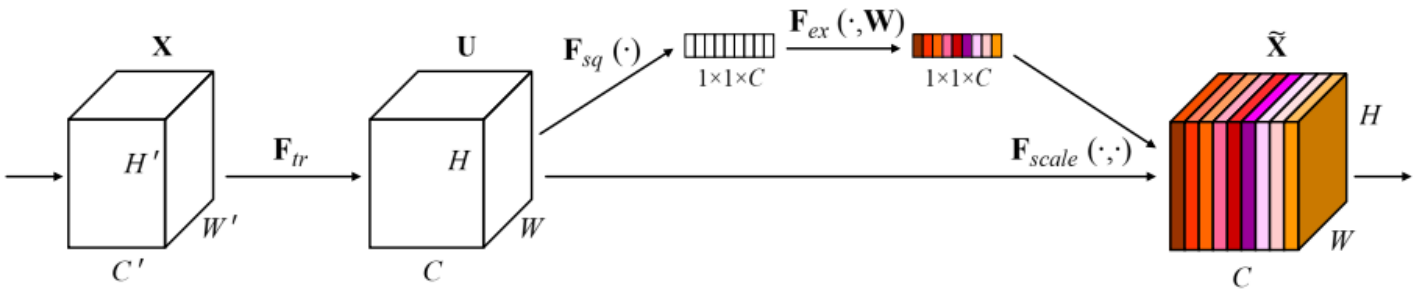

Fig. 1. A Squeeze-and-Excitation block.

Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." CVPR. 2018.

DANet

- Dual Attention Network is also similar to this work, but
- expensive cost
- Depend on inpu

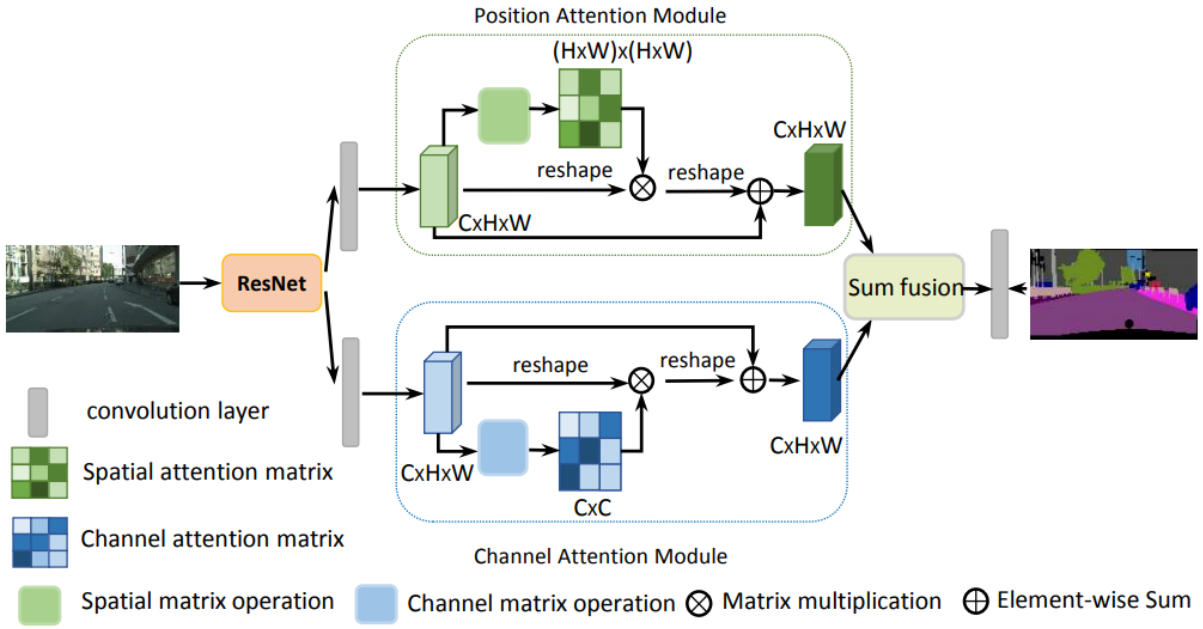

Figure 2: An overview of the Dual Attention Network. (Best viewed in color)

Fu, Jun, et al. "Dual attention network for scene segmentation." CVPR. 2019.

Method (HANet)

- Overview

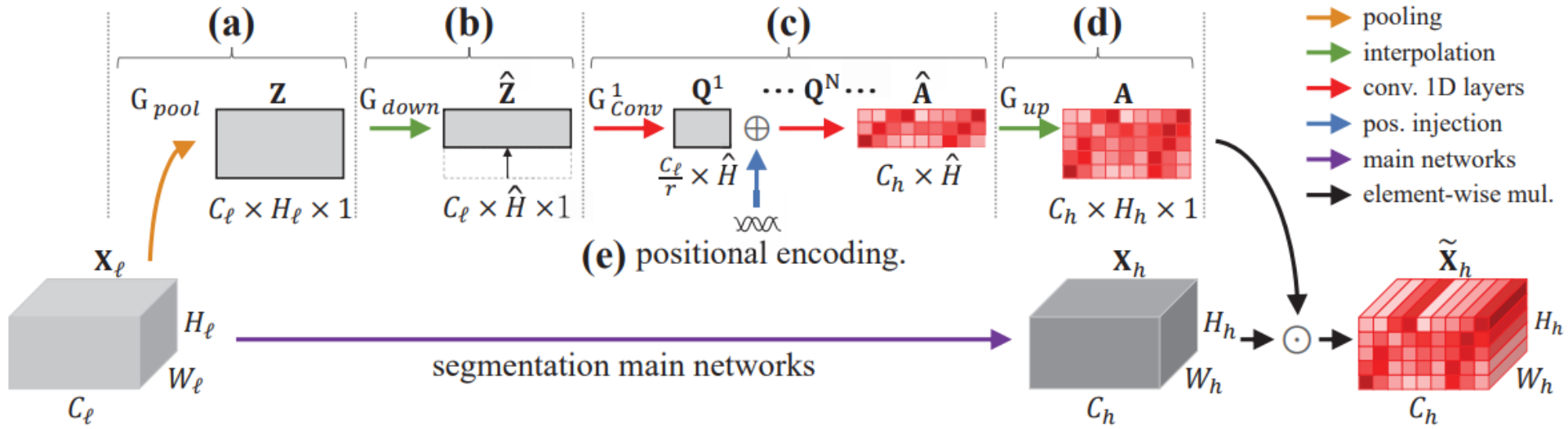

Figure 2: Architecture of our proposed HANet.

Method (HANet)

- Width-wise pooling (Fig. 2(a))

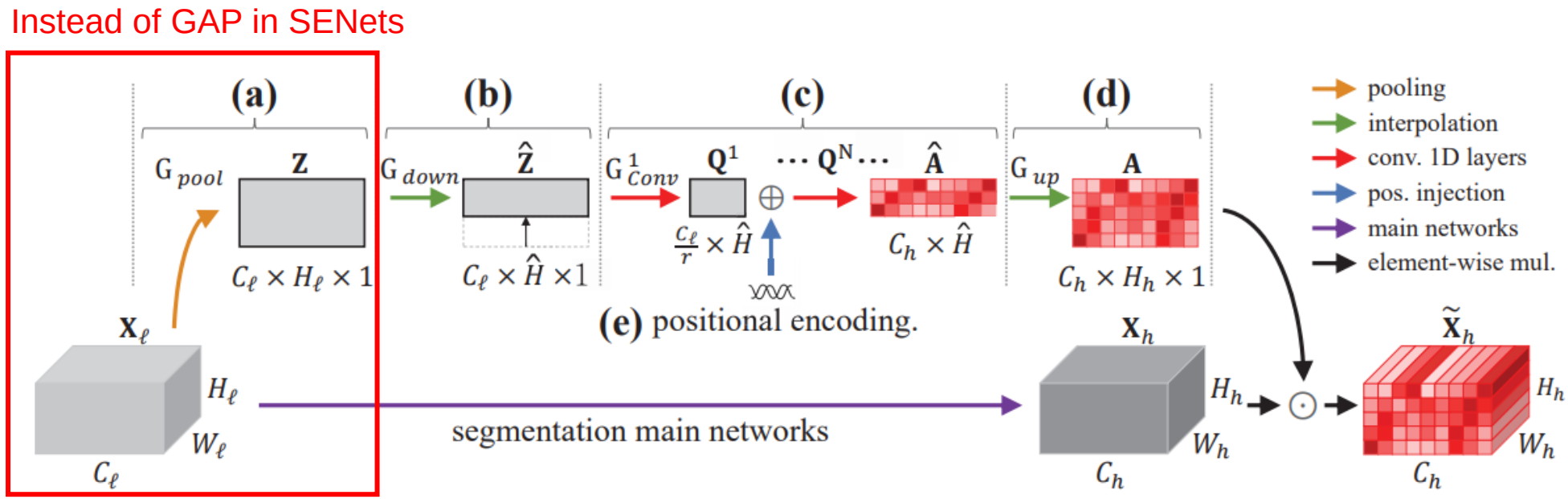

Figure 2: Architecture of our proposed HANet.

Method (HANet)

- Interpolation for coarse attention (Fig. 2(b,d)).
  - Not all the rows of matrix Z are necessary for computing an effective attention map.

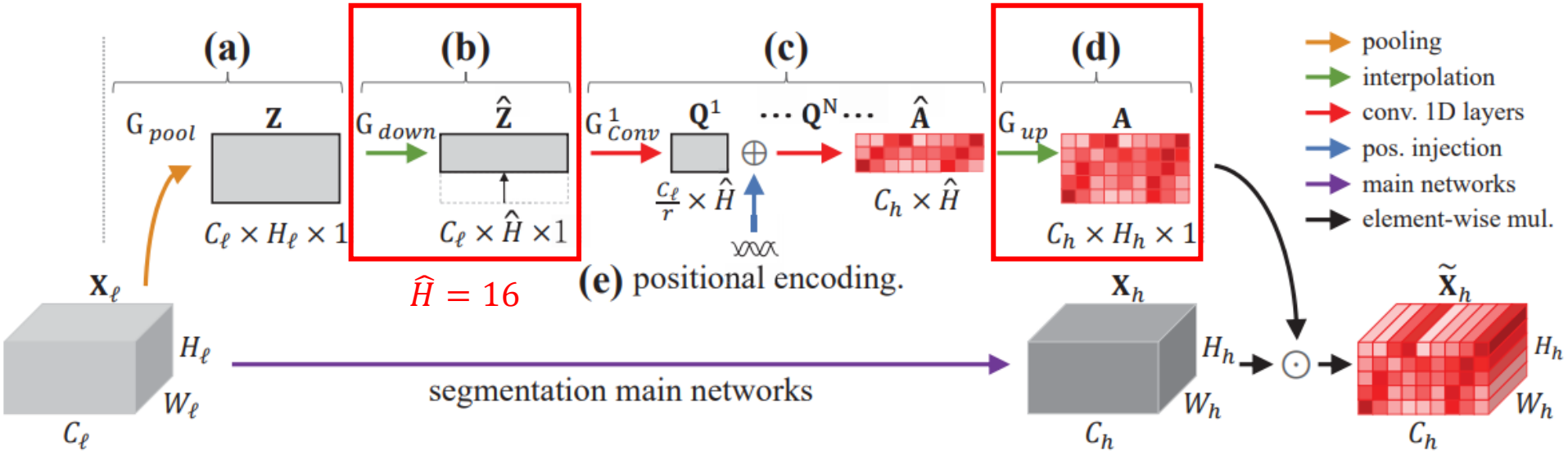

Figure 2: Architecture of our proposed HANet.

Method (HANet)

- Computation of height-driven attention map (Fig. 2(c)).
  - Adopt convolutional layers to let the relationship between adjacent rows be considered while estimating the attention map

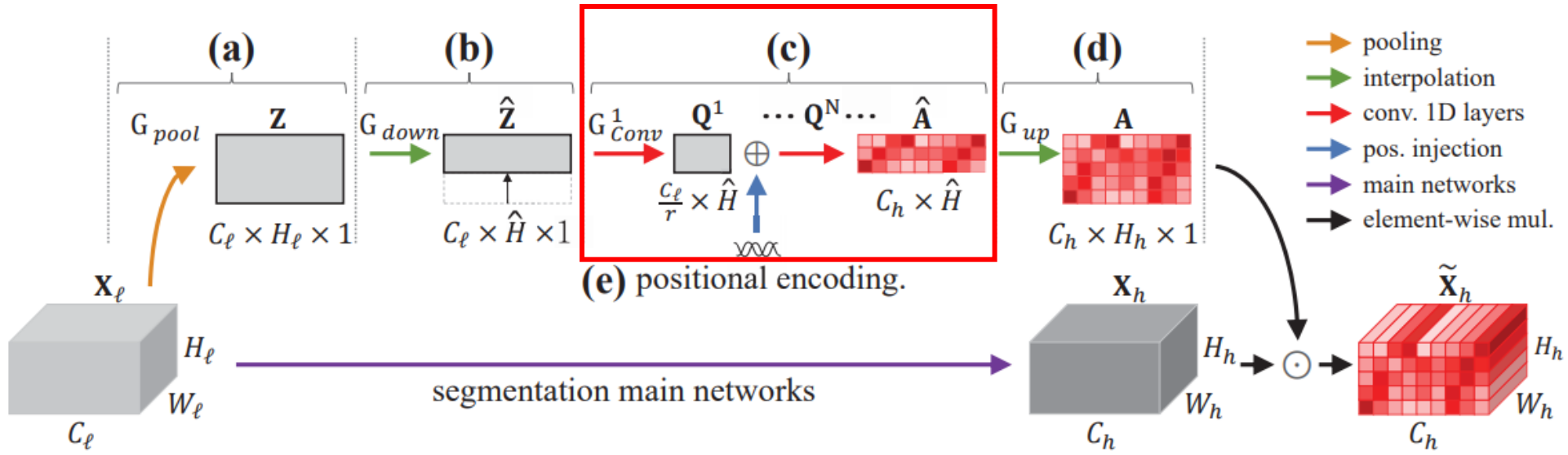

Figure 2: Architecture of our proposed HANet.

Method (HANet)

- The attention map A indicates which channels are critical at each row.
- There may exist multiple informative features at each row in the intermediate layer.
- To allow these multiple features and labels, a sigmoid function is used

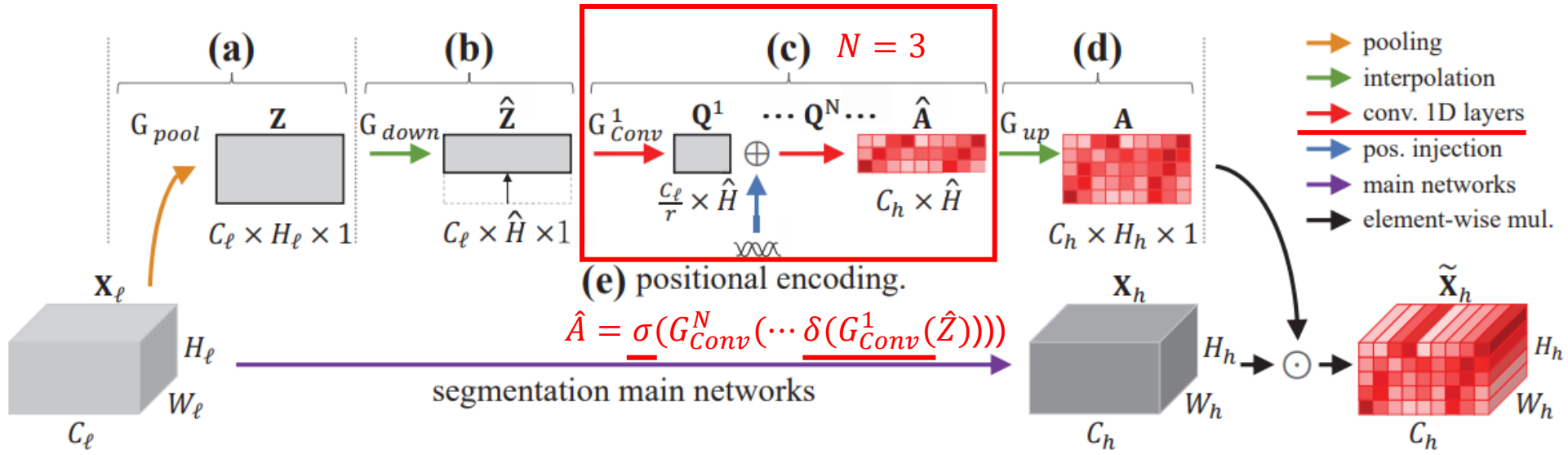

Figure 2: Architecture of our proposed HANet.

Method (HANet)

- Incorporating positional encoding (Fig. 2(e))
  - When humans recognize a driving scene, they have prior knowledge on the vertical position of particular objects (e.g., road and sky appear in the lower and upper part)

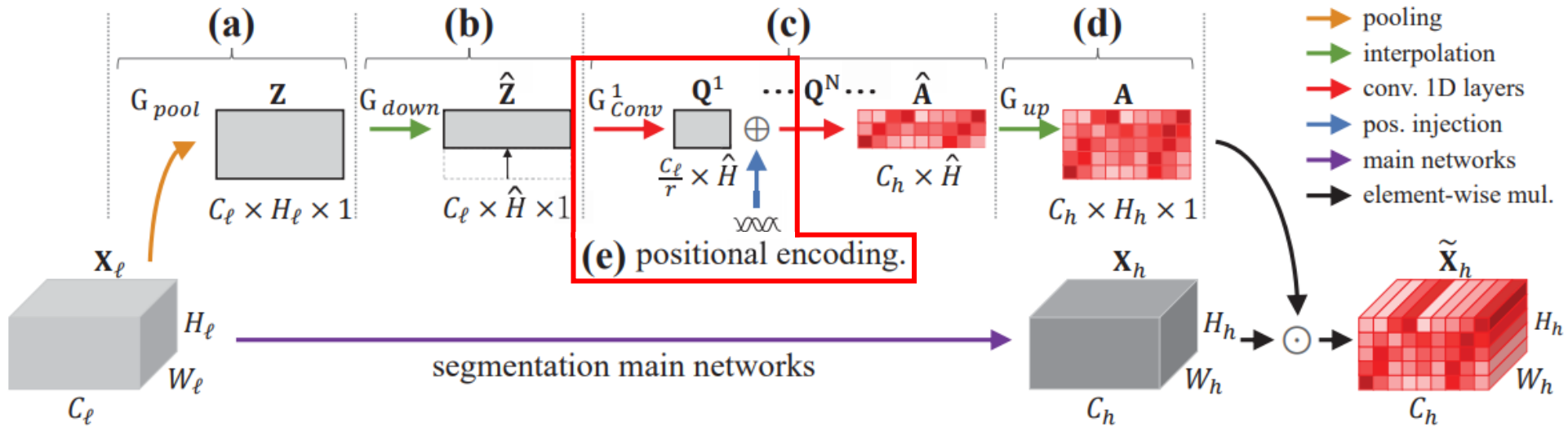

Figure 2: Architecture of our proposed HANet.

Apply on Segmentation Networks

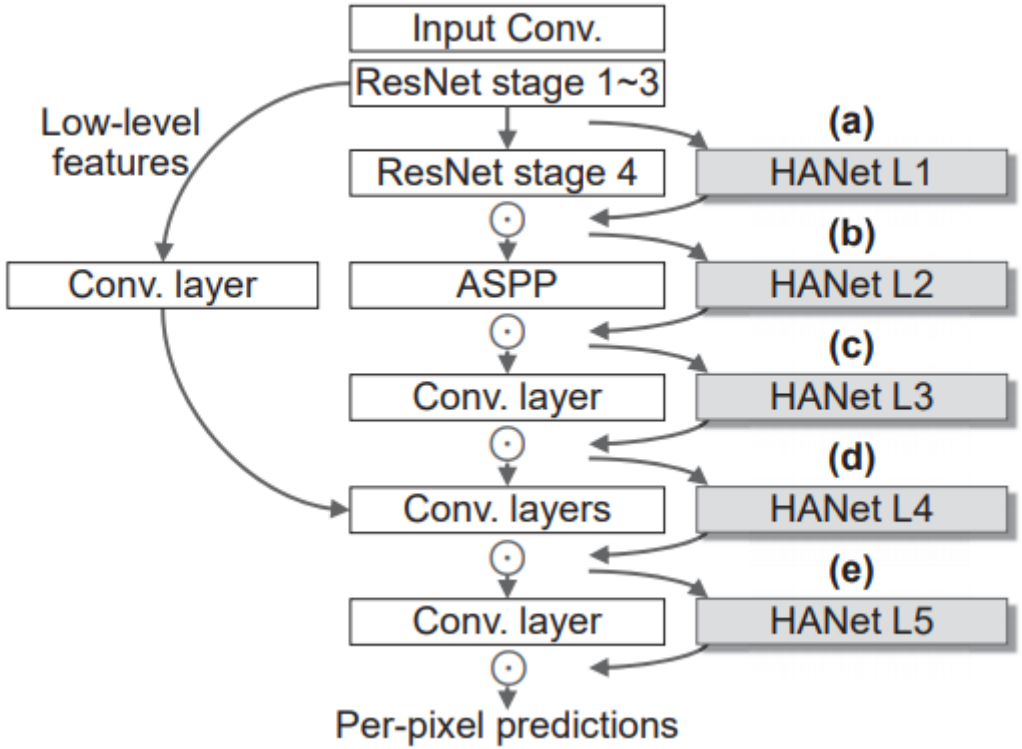

Figure 3: Semantic segmentation networks incorporating HANet in five different layer

Experiments

Backbone	OS	Models	Params	GFLOPs	mIoU(%)
ShuffleNet V2 (1×) [29]	32	Baseline	12.6M	64.34	70.27
		+HANet	14.9M	64.39	**71.30**
	16	Baseline	12.6M	117.09	70.85
		+HANet	13.7M	117.14	**71.52**
MobileNet V2 [29]	16	Baseline	14.8M	142.74	73.93
		+HANet	16.1M	142.80	**74.96**
	8	Baseline	14.8M	428.70	73.40
		+HANet	15.4M	428.82	**74.70**
ResNet-50 [18]	16	Baseline	45.1M	553.74	76.84
		+HANet	47.6M	553.85	**77.78**
	8	Baseline	45.1M	1460.56	77.76
		+HANet	46.3M	1460.76	**78.71**
ResNet-101 [18]	16	Baseline	64.2M	765.53	77.80
		+HANet	65.4M	765.63	**79.31**
	8	Baseline	64.2M	2137.82	79.25
		+HANet	65.4M	2138.02	**80.29**

Table 2: Comparison of mIoU, the number of model parameters and FLOPs between the baseline and HANet on Cityscapes validation set according to various backbone networks and output stride. Adding HANet to the baseline consistently increase the mIoU with minimal cost increase.

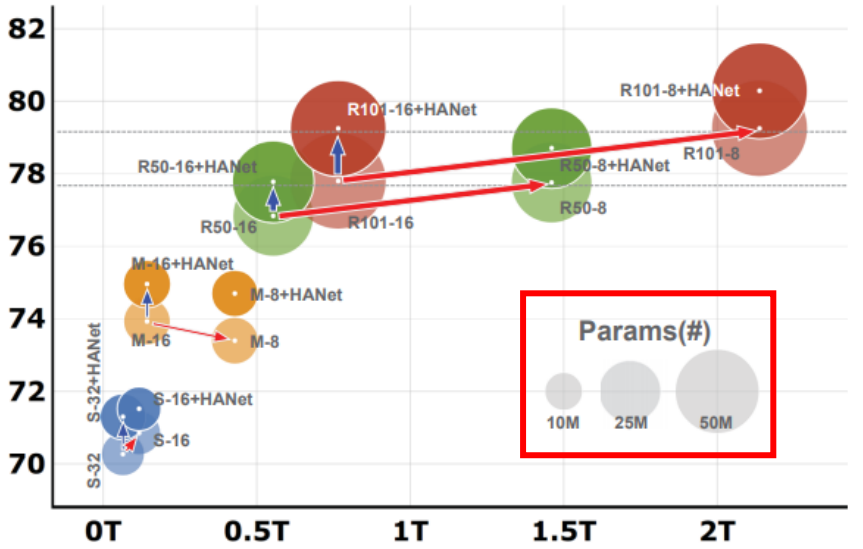

Figure 4: Comparison of the performance and complexity among the baseline and HANet on the various backbone networks. x-axis denotes teraFLOPs and y-axis denotes mIoU. The circle size denotes the number of model parameters. The texts in the colored circle indicate backbone networks, output stride, and whether HANet is adopted to the baseline. S, M, R50, and R101 denote ShuffleNetV2, MobileNetV2, ResNet-50, and -101, respectively. (e.g., S-16: Baseline, ShuffleNetV2, and output stride 16)

Experiments

- The performance significantly rises on the upper and lower regions as in Table 5.

Model	Upper	Mid-upper	Mid-lower	Lower	Entire
Baseline	78.69	76.35	83.16	70.59	81.14
+HANet	80.29	77.09	84.09	73.04	82.05
Increase(%)	**+1.60**	+0.74	+0.93	**+2.45**	+0.91

Table 5: mIoU(%) comparison to baseline on each part of image divided into four horizontal sections. ResNet-101, output stride 8 on Cityscapes validation set.

Qualitative Analysis

- Attention map visualization.
  - HANet assigns a different amount of attention to a different vertical position
  - indicating that the model properly learns structural priors with respect to the height in urban-scene data.

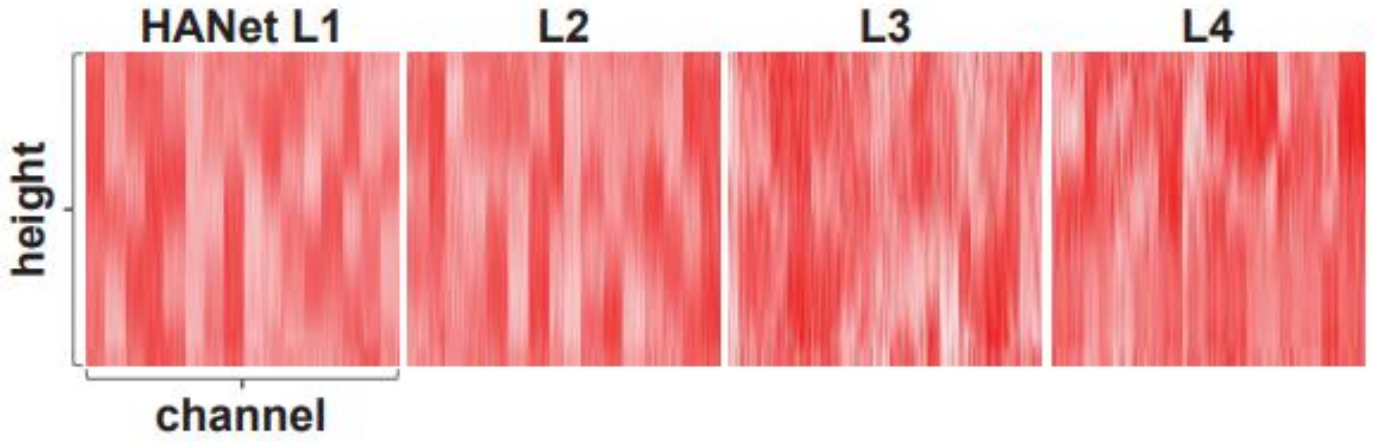

Figure 5: Visualization of attention map from HANet at different layers.
To better visualize, the channels are clustered.

Qualitative Analysis

- The distribution of the attention map from HANet at the last layer, which is following the actual height-wise class distribution obtained from the Cityscapes training images.

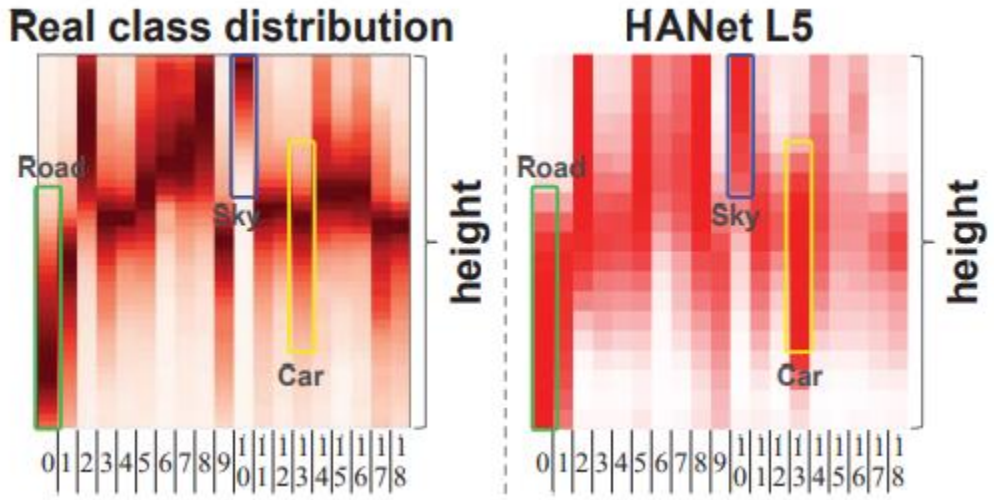

Figure 7: Height-wise class distributions and attention map visualization (L5). The number ranging from 0 to 18 indicates a different class.

References

- Choi, Sungha, Joanne T. Kim, and Jaegul Choo. "Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks." CVPR. 2020.
- Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." CVPR. 2018.
- Fu, Jun, et al. "Dual attention network for scene segmentation." CVPR. 2019.