HGFormer: Hierarchical Grouping Transformer for Domain Generalized Semantic Segmentation

Jian Ding[1,2], Nan Xue[1], Gui-Song Xia[1], Bernt Schiele[2], Dengxin Dai[2]

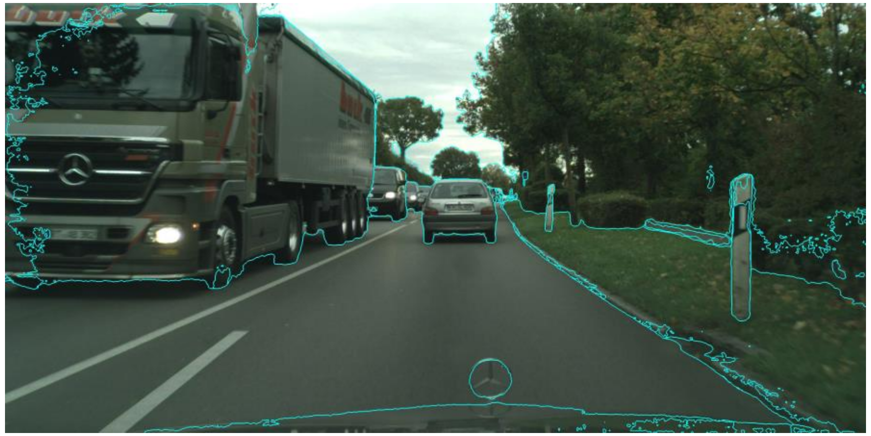
[1]Wuhan University

[2]Max Planck Institute for Informatics, Saarland Informatics Campus

Quick preview

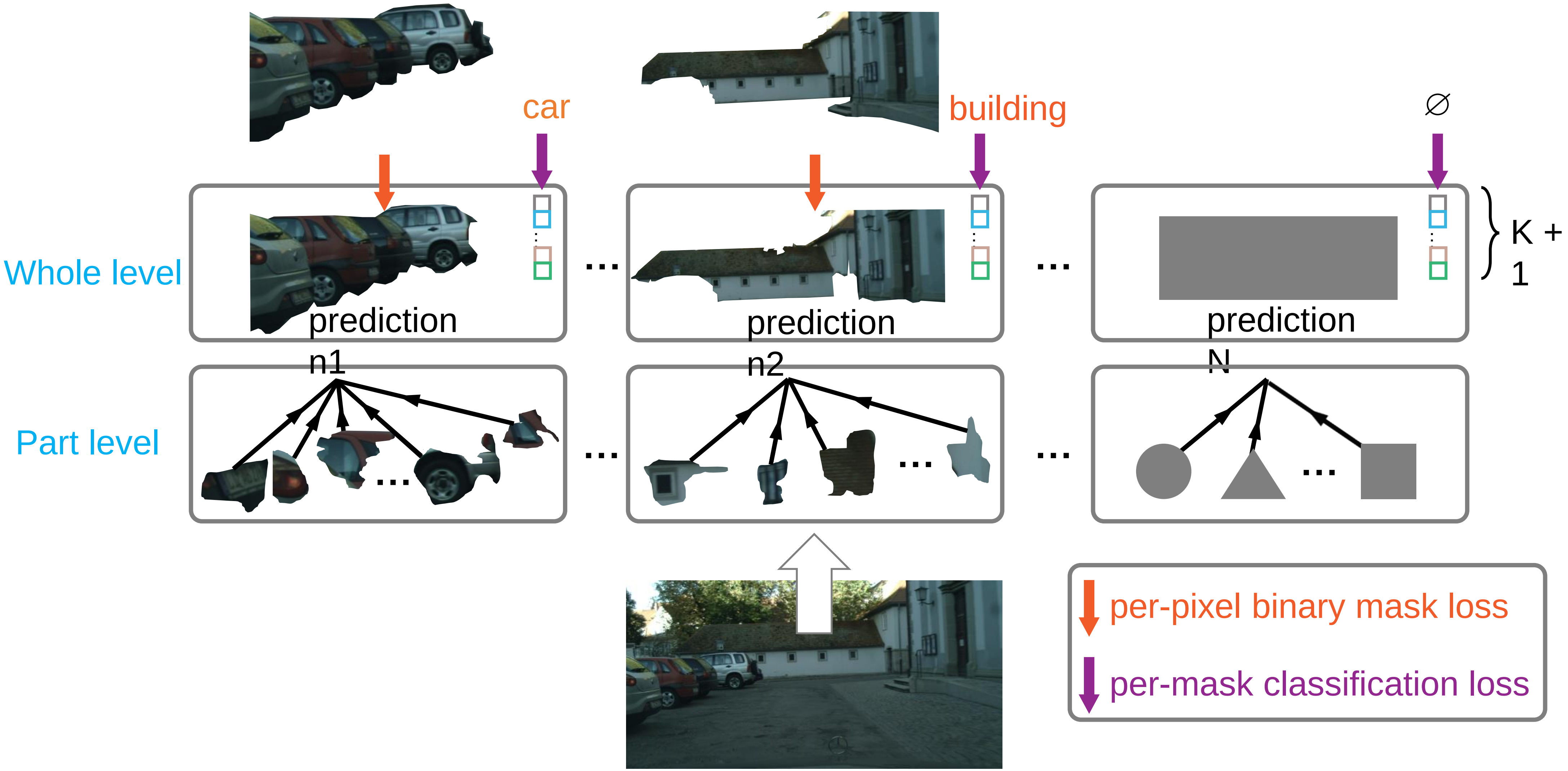
- **Domain generalization setting:** generalize a segmentation model from a **source domain** to a different **target domain** without fine-tuning
- We study the domain generalized segmentation from the perspective of **segmentation formulation**
  - Intuitively, classification on large units (**masks**) should be more robust than classification on small units (**pixels**)
  - The process of grouping pixels into whole-level masks directly form pixels is challenging under distribution shifts

Classification on pixels



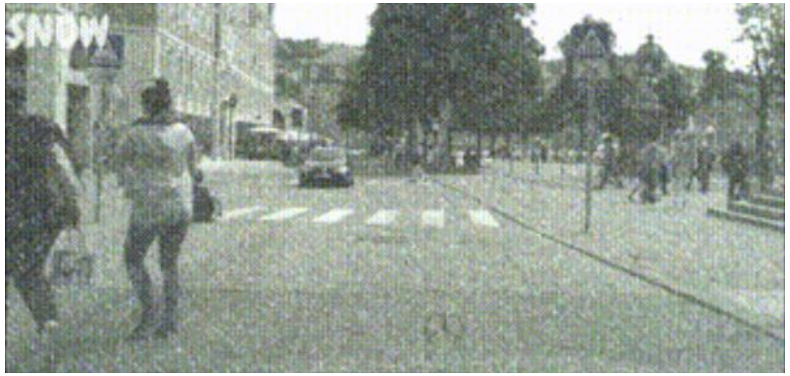
Classification on masks

Quick preview



Task: Domain Generalization in Segmentation

- **Domain generalization setting:** train a segmentation model on a **source domain**, and directly test it on a different **target domain** without fine-tuning



Cityscapes -> Cityscapes-C generalization (normal to synthetic corruptions)

Cityscapes -> ACDC generalization (normal to real adverse conditions)

Existing Methods and Our Motivation

Domain randomization

① DGPC [Xiangyu et al., ICCV 2019]

② GTR [Duo et al., TIP 2021]

Normalization

① SAN [Duo et al., CVPR 2022]

② IBN-Net [Xingang et al., ECCV 2018]

Transformer

① Segformer [Choi et al., CVPR 2021]

② FAN [Daquan et al., ICML 2022]

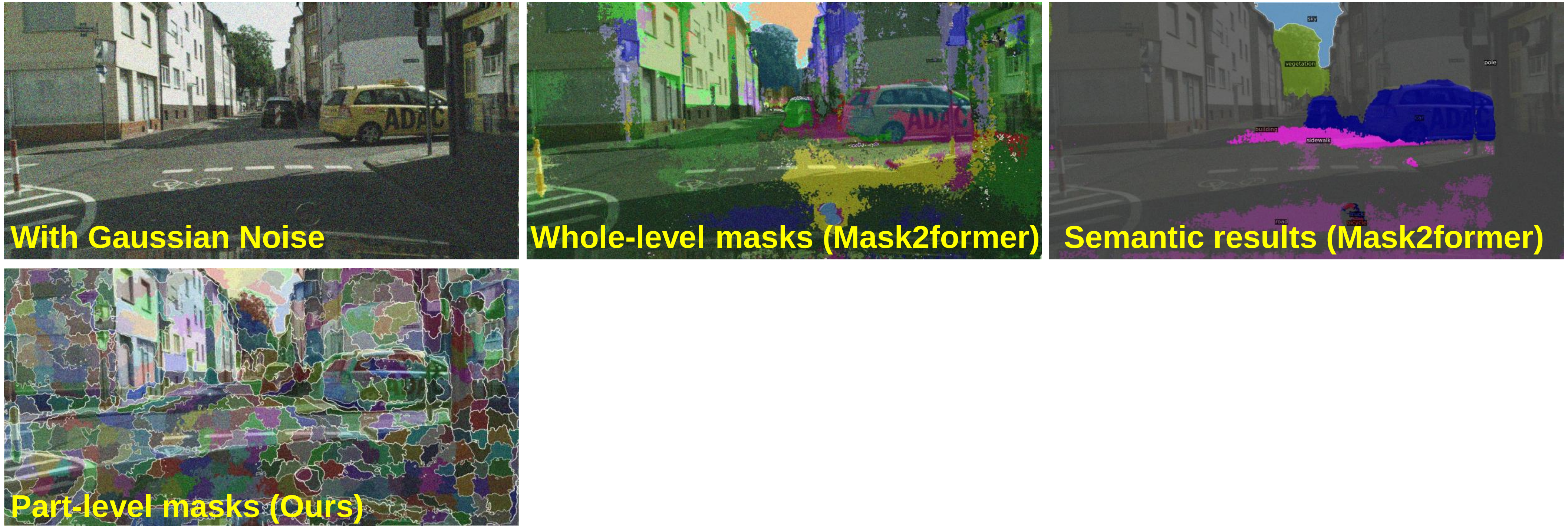
Our motivation

- Vision Transformer has been shown to be more robust than traditional CNNs, and attention in transformers can be explained as a kind of **visual grouping**
- Can we explicitly introduce the grouping process into segmentation decoder to improve the robustness?

The Problem of Existing Grouping Based Semantic Segmentation

- If we already have grouped the pixels into masks correctly, we can make reliable classification, since the masks allow to aggregate features over large image regions
- The process to group pixels directly into (class agnostic) ***whole-level*** masks is not robust under distribution shifts

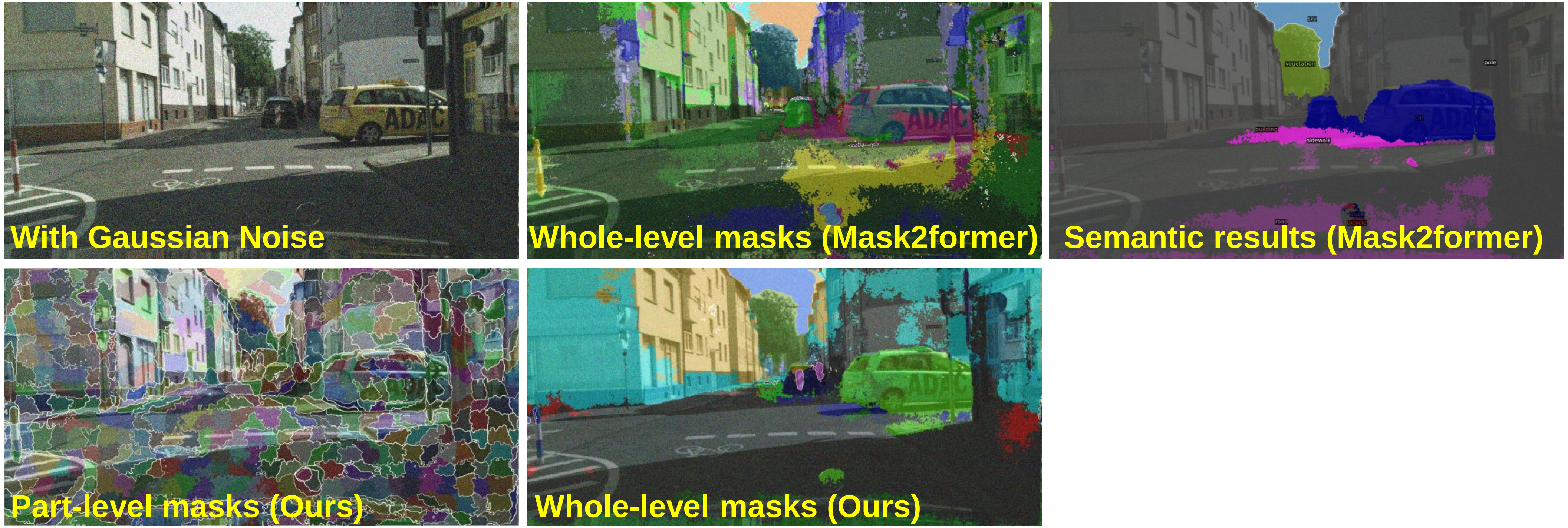
The Problem of Existing Grouping Based Semantic Segmentation



Our solution: hierarchical grouping

- ☐ We first group pixels into ***part-level*** masks
- ☐ Then we group part-level masks into whole-level masks
- ☐ Then we make classifications on both part-level and whole-level masks

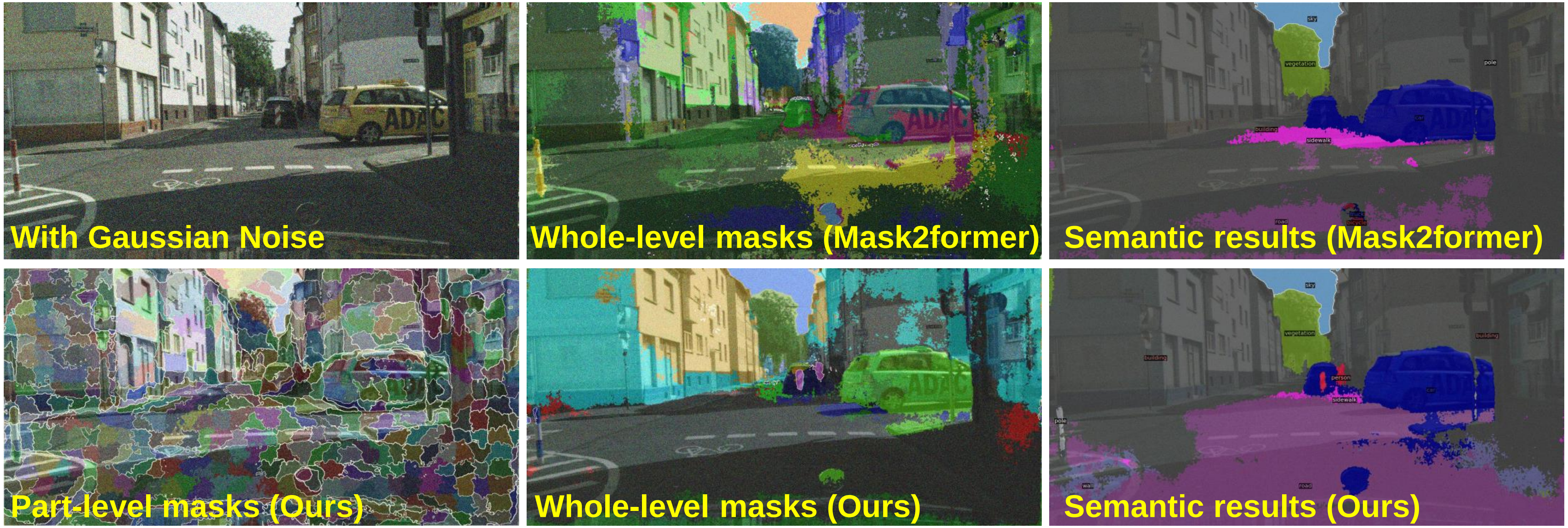
The Problem of Existing Grouping Based Semantic Segmentation



Our solution: hierarchical grouping

- ☐ We first group pixels into ***part-level*** masks
- ☐ Then we group part-level masks into whole-level masks
- ☐ Then we make classifications on both part-level and whole-level masks

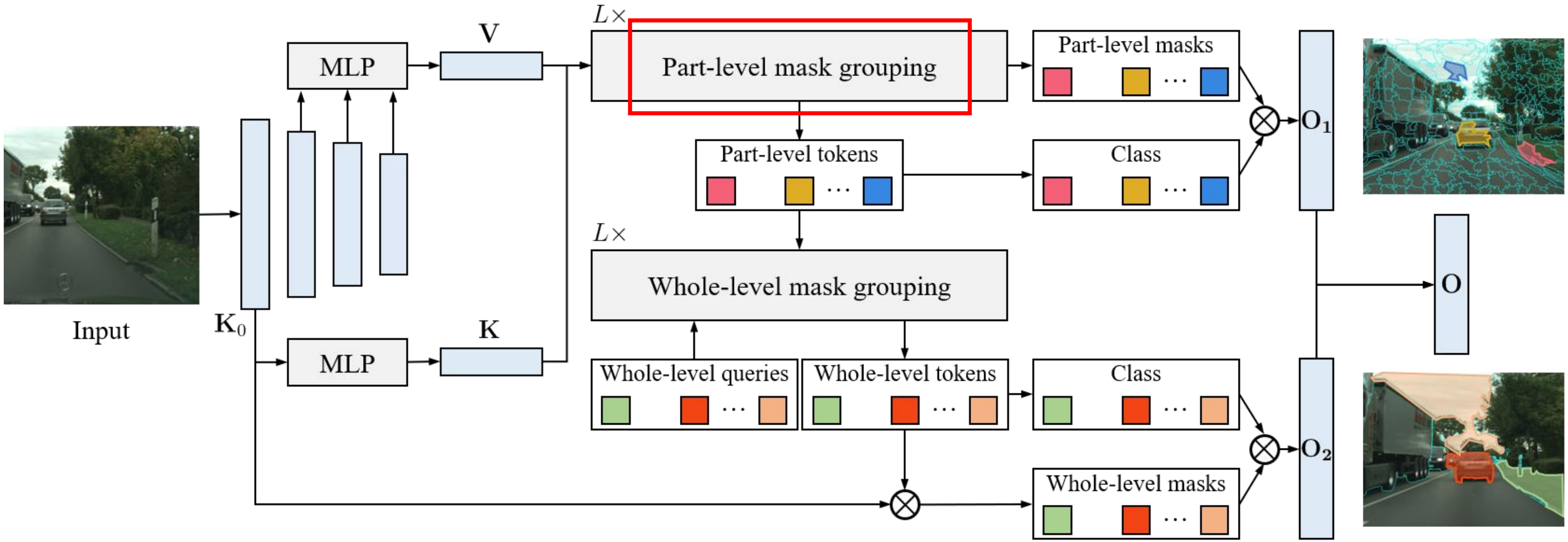
The Problem of Existing Grouping Based Semantic Segmentation



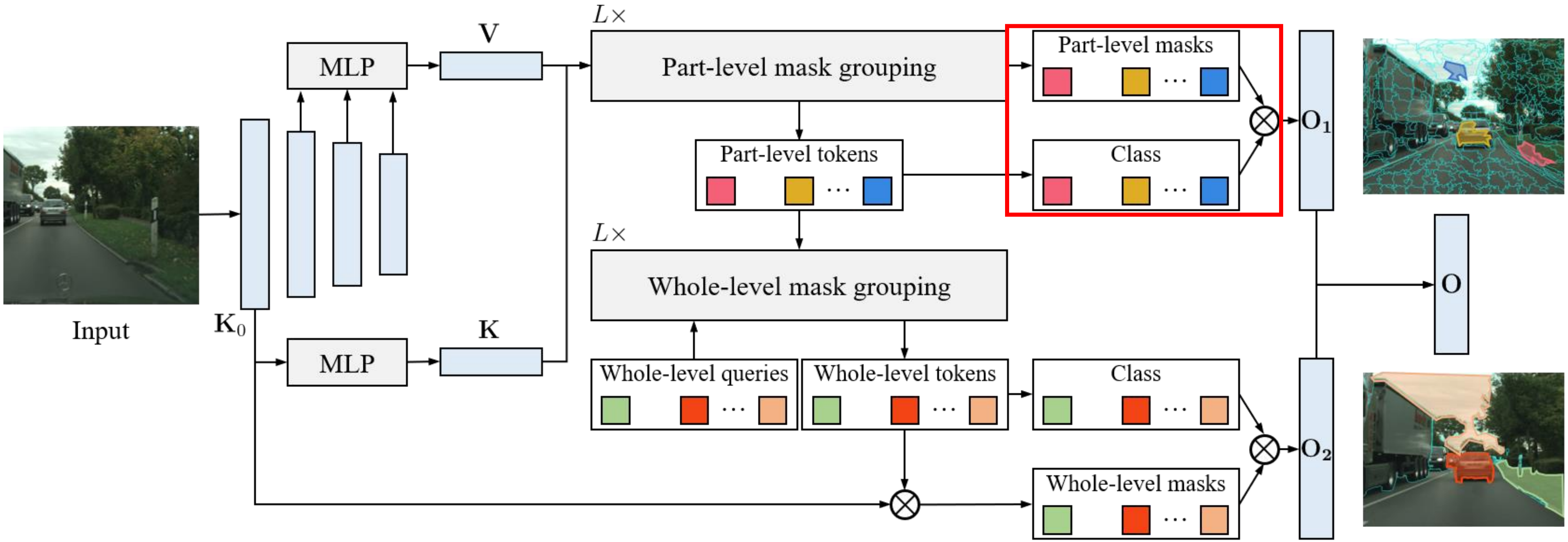
Our solution: hierarchical grouping

- ☐ We first group pixels into ***part-level*** masks
- ☐ Then we group part-level masks into whole-level masks
- ☐ Then we make classifications on both part-level and whole-level masks

HGFormer: Hierarchical Grouping Transformer

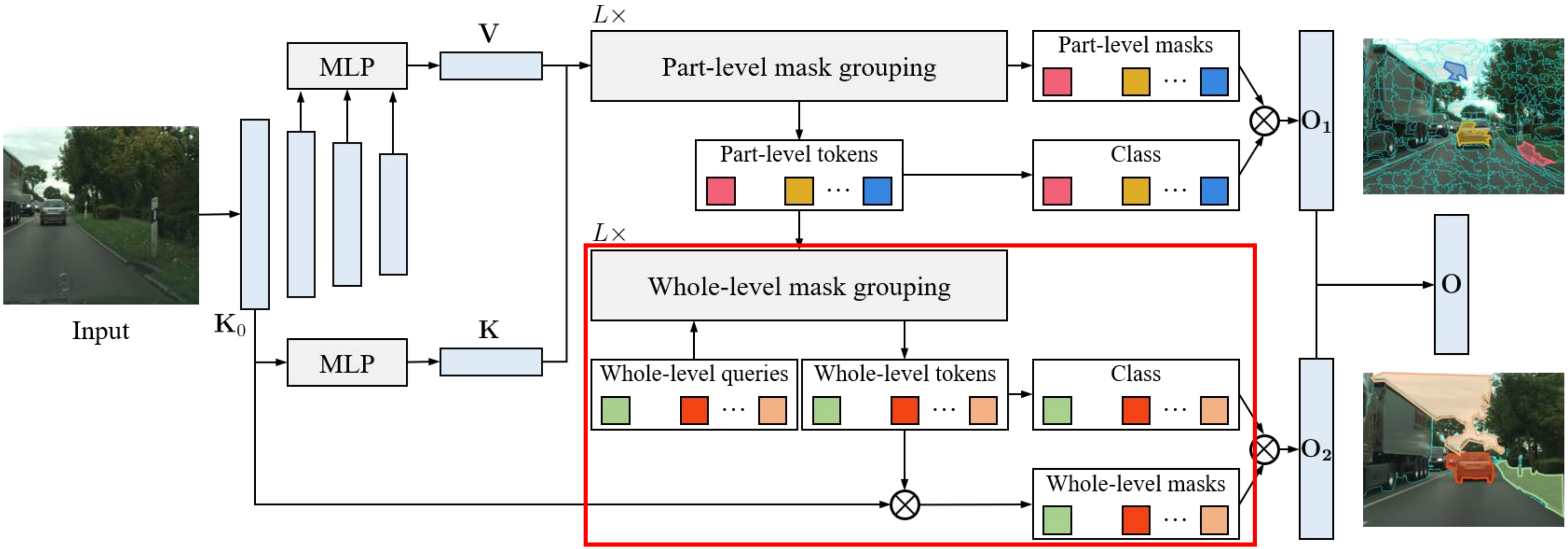


HGFormer: Hierarchical Grouping Transformer

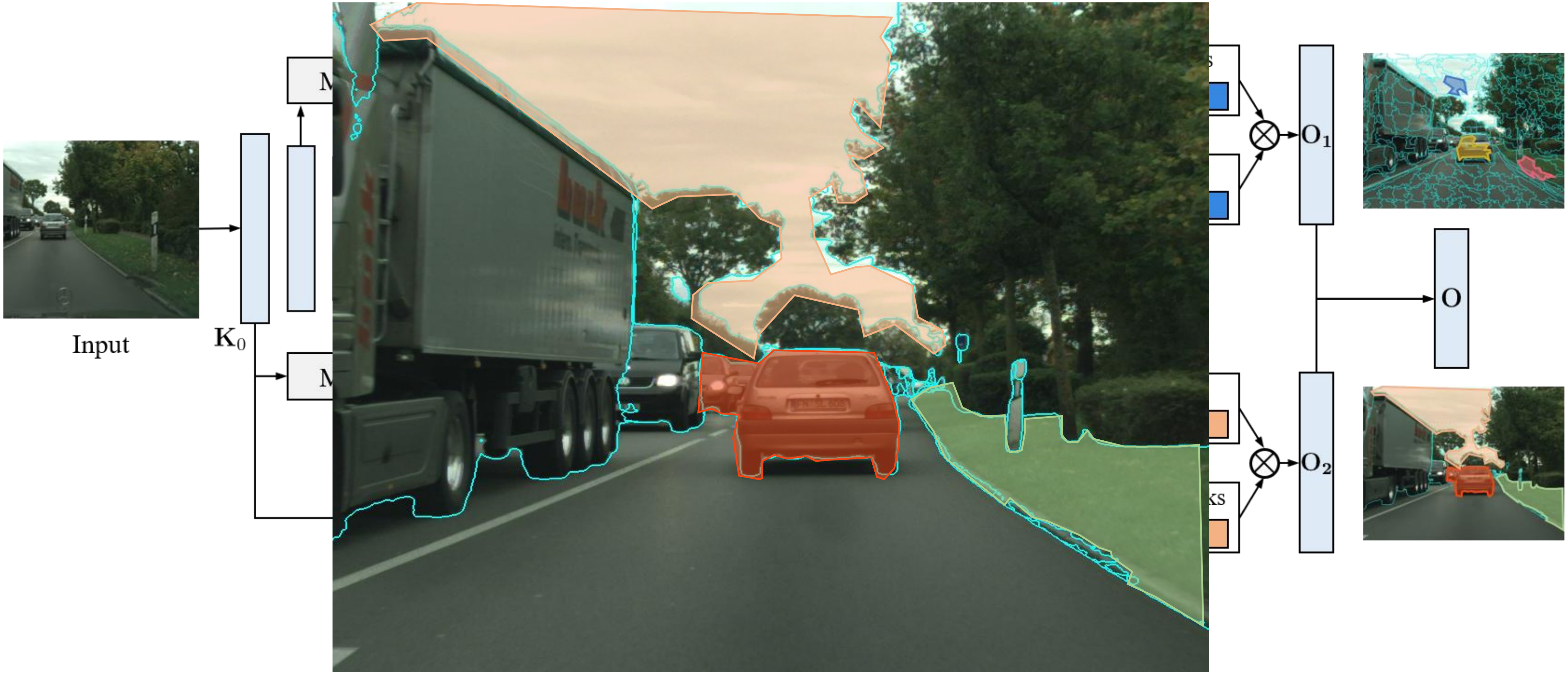


HGFormer: Hierarchical Grouping Transformer

HGFormer: Hierarchical Grouping Transformer



HGFormer: Hierarchical Grouping Transformer



HGFormer: Hierarchical Grouping Transformer

Algorithm 1 Part-level grouping

Require: *Pixel* feature map $\mathbf{K} \in \mathbb{R}^{(H\times W)\times d}$, classification feature map $\mathbf{V} \in \mathbb{R}^{(H\times W)\times d}$

1: Initialize the cluster *center* features $\mathbf{Q^1} \in \mathbb{R}^{N_p\times d}$ by down sampling $\mathbf{K}$

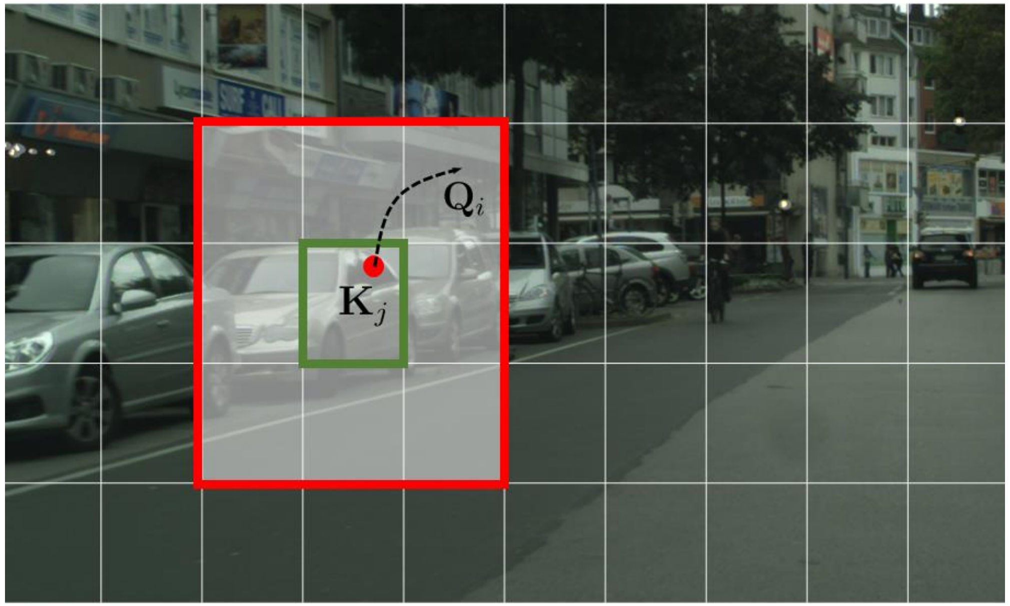
2: **for** $t = 1, \cdots, L$ **do**

3: Compute assignment matrix $\mathbf{A^t}$ by $\mathbf{Q^t}$ and $\mathbf{K}$

4: Update the cluster center features $\mathbf{Q^{t+1}} = \mathbf{A^t} \times \mathbf{K}$

5: Update the part-level tokens $\mathbf{Z^t} = \mathbf{A^t} \times \mathbf{V}$

6: **end for**



Part-level grouping

- A kind of (local) k-means
- Cluster centers are initialized by regular grid c
- Each pixel is only assigned one of its 9 nearby cells

$$\mathbf{D}_{i,j} = \begin{cases} f(\mathbf{Q}_i, \mathbf{K}_j) & \text{if } i \in N_j \\ -\infty & \text{if } i \notin N_j, \end{cases} \tag{2}$$

$$\mathbf{A}_{i,j} = \mathrm{softmax}(\mathbf{D})(i,j) = \frac{\exp(\mathbf{D}_{i,j})}{\sum_{i=1}^{N_p} \exp(\mathbf{D}_{i,j})}, \tag{3}$$

Results

Cityscapes-to-ACDC generalization

Method	backbone	Fog	Night	Rain	Snow	All
RefineNet [31]	R101	46.4	29	52.6	43.3	43.7
DeepLabv2 [10]	R101	33.5	30.1	44.5	40.2	38
DeepLabv3+ [12]	R101	45.7	25	50	42	41.6
DANet [19]	DA101	34.7	19.1	41.5	33.3	33.1
HRNet [54]	HR-w48	38.4	20.6	44.8	35.1	35.3
Mask2former [14]	R50	54.1	36.5	53.1	50.6	49.8
HGFormer (**ours**)	R50	56.5	35.8	57.7	56.2	**53.0**
Mask2former [14]	Swin-T	56.4	39.1	58.9	58.2	54.6
Segformer [60]	B2	59.2	38.9	62.5	58.2	56.2
HGFormer (**ours**)	Swin-T	58.5	43.3	62.0	58.3	**56.7**
Segformer [60]	B5	63.2	47.8	66.4	63.7	62.0
Mask2former [14]	Swin-L	69.1	53.1	68.3	65.2	65.0
HGFormer (**ours**)	Swin-L	69.9	52.7	72.0	68.6	**67.2**

Cityscapes-to-other generalization

Method	backbone	B	M	G	S	Average
IBN [39]	R50	48.6	57.0	45.1	26.1	44.2
SW [40]	R50	48.5	55.8	44.9	26.1	43.8
DRPC [68]	R50	49.9	56.3	45.6	26.6	44.6
GTR [43]	R50	50.8	57.2	45.8	26.5	45.0
ISW [16]	R50	50.7	58.6	45	26.2	45.1
SAN-SAW [42]	R50	53.0	59.8	47.3	28.3	47.1
Mask2former [14]	R50	46.8	61.6	48.0	31.2	46.9
HGFormer (**ours**)	R50	51.5	61.6	50.4	30.1	**48.4**
Mask2former [14]	Swin-T	51.3	65.3	50.6	34	50.3
HGFormer (**ours**)	Swin-T	54.3	66.2	52.0	32.5	**51.2**
Mask2former [14]	Swin-L	60.1	72.2	57.8	42.4	58.1
HGFormer (**ours**)	Swin-L	61.5	72.1	59.4	41.3	**58.6**

Cityscapes-to-cityscapes-c generalization

Method	Average	Blur				Noise				Digital				Weather			
		Motion	Defoc	Glass	Gauss	Gauss	Impul	Shot	Speck	Bright	Contr	Satur	JPEG	Snow	Spatt	Fog	Frost
Segformer-B2 [20]	40.4	56.1	56.0	41.5	49.8	2.7	3.0	3.4	21.5	78.3	65.7	74.2	24.9	18.0	53.1	71.1	26.7
Mask2former-Swin-T [2]	41.6	51.5	49.4	38.2	46.2	9.6	9.8	13.5	44.4	74.2	60.0	70.0	23.3	23.7	59.4	65.4	27.3
HGFormer-Swin-T (**ours**)	**43.9**	52.9	53.9	39.0	49.5	12.1	12.3	18.2	46.3	75.0	60.0	71.2	27.2	29.4	60.6	65.0	29.1
Segformer-B5 [20]	49.1	59.9	58.2	51.6	54.0	14.3	16.9	16.4	49.1	80.0	68.6	77.3	40.4	30.3	58.8	74.2	35.7
Mask2former-Swin-L [2]	58.7	63.5	66.6	62.1	62.3	26.2	35.9	33.2	62.9	80.0	72.6	77.3	52.5	50.5	75.3	75.1	43.0
HGFormer-Swin-L (**ours**)	**59.4**	64.1	67.2	61.5	63.6	27.2	35.7	32.9	63.1	79.9	72.9	78.0	53.6	55.4	75.8	75.5	43.2

We compare with (1) previous domain generalization for semantic segmentation methods, and
(2) two representative transformer-based methods: Segformer and Mask2former, which are based on **pixel classification** and **whole-level mask classification**

Ablation Studies

Number of iterations in part-level classification

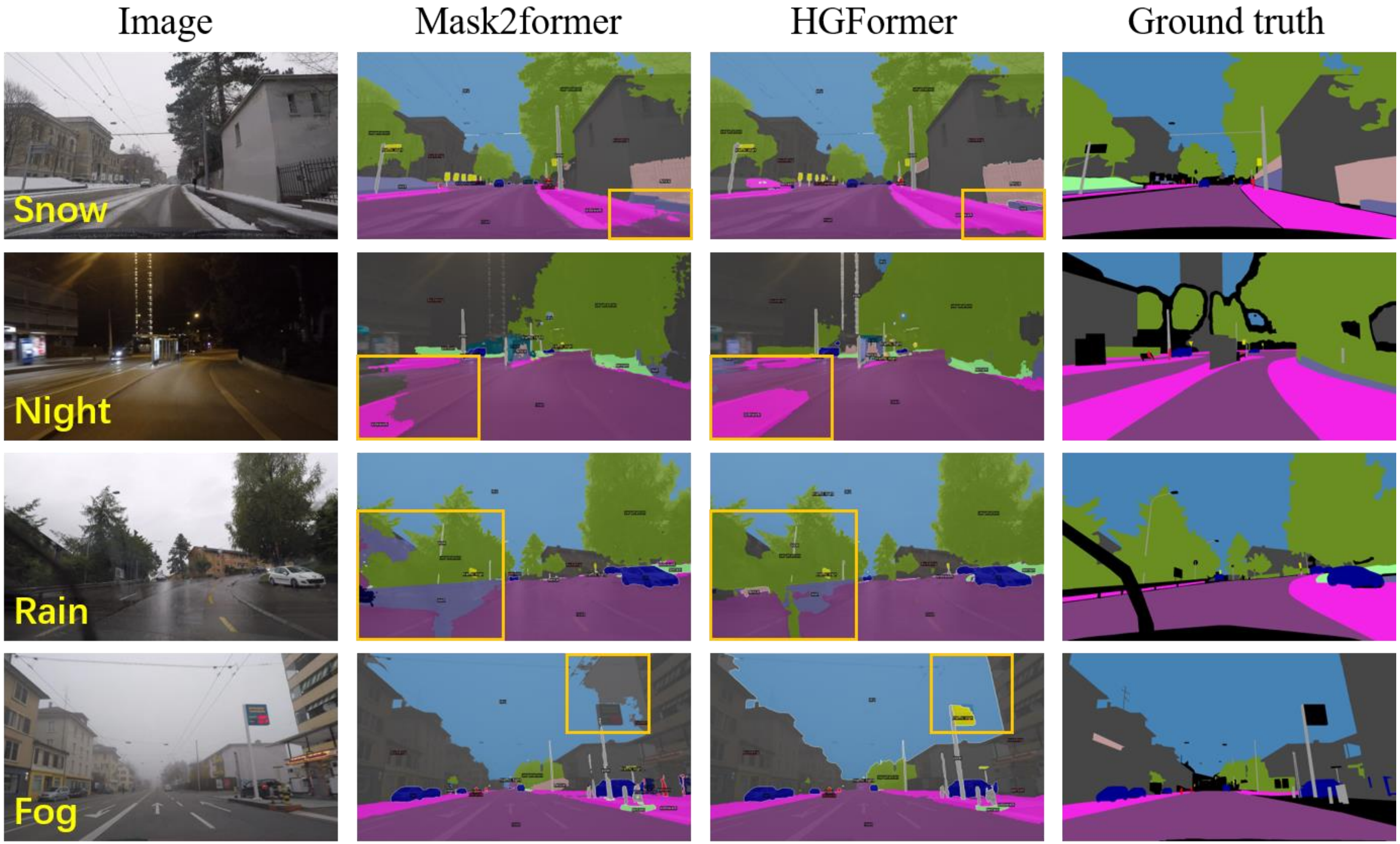
Iter	C	A	G	B	S	M	Avg
1	76.8	56.1	51.3	52.1	32.1	65.8	55.7
2	77.6	56.1	51.4	52.0	32.3	65.9	55.9
3	77.9	56.2	51.8	52.6	32.8	66.2	56.2
4	77.9	56.5	52.0	52.6	32.6	66.3	**56.3**
5	77.8	56.4	51.7	52.6	32.5	66.3	56.2
6	77.4	55.4	50.5	52.2	32.3	65.6	55.6

Comparison of part-level classification and whole-level classification, and their combination

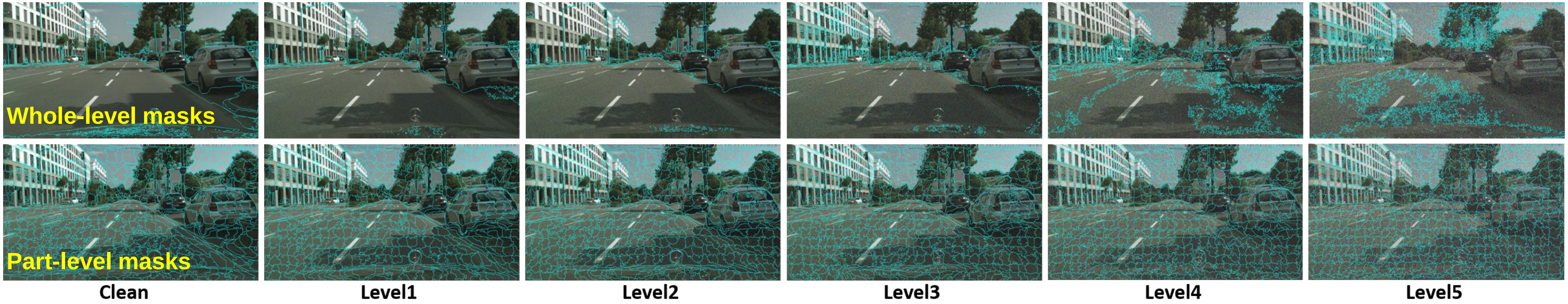
Pixel-level	Whole-level mask	Part-level mask	ACDC (all)	GTAV	BDD	Synthia	Mapillary	Average
✓			54.1	49.5	52.5	32.8	65.4	50.9
	✓		54.5	49.5	51.5	33.8	66.3	51.1
		✓	56.2	51.3	53.1	33.3	66.5	52.1
	✓	✓	56.6	51.3	53.4	33.6	66.9	**52.4**

Visualization results on Cityscapes-C

Visualization results on ACDC



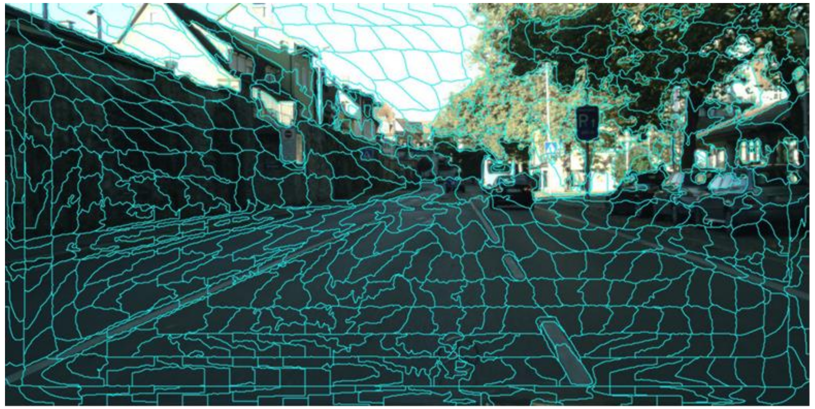
Visualization Analyses

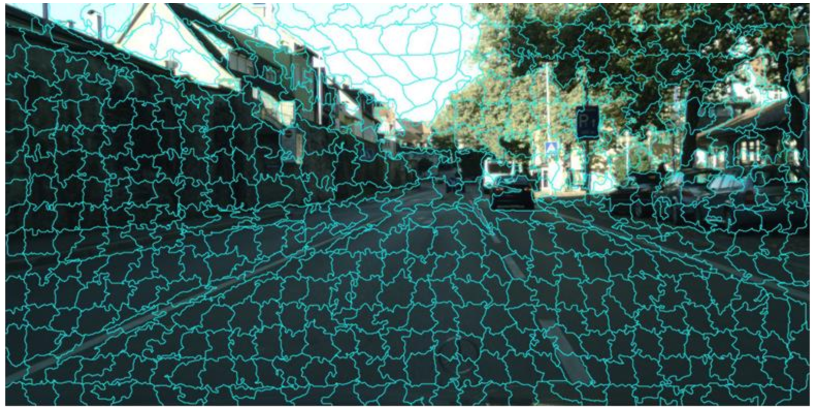


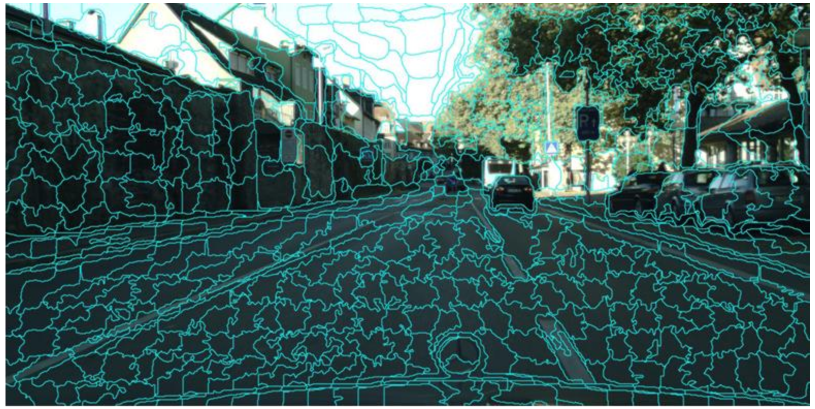
Gaussian noise at different levels

The whole-level masks are not robust as part-level masks.

Visualization Analyses


Randomly initialized


ImageNet pre-trained


Segmentation annotation trained

- Even use the ***randomly initialized*** weights, we can still generate some reasonable part-level masks (super pixels).
- The results also indicate our model has the potential for unsupervised segmentation

Conclusion

- Mask classification is robust, but the process to group pixels into whole-level masks is not robust
- Hierarchical grouping can be used to improve the robustness of segmentation models
- The grouping based segmentation also has the potential for unsupervised segmentation

Code will be available at