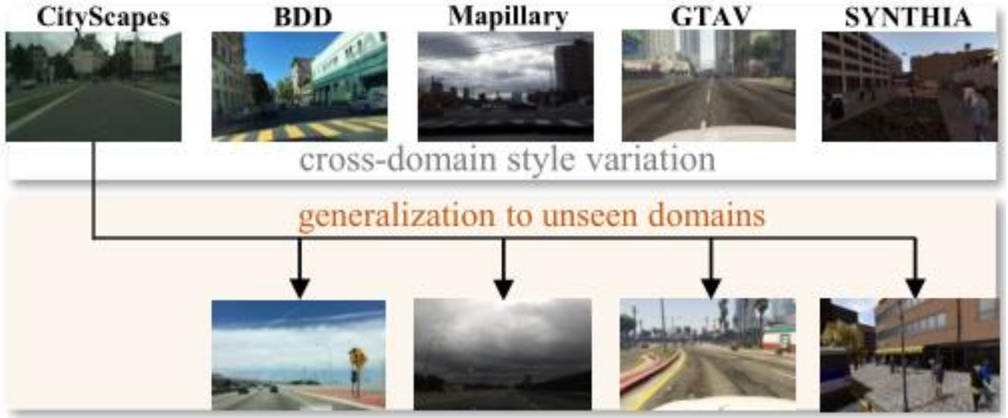
Learning Frequency-Adapted Vision Foundation Model for Domain Generalized Semantic Segmentation

Qi Bi, Jingjun Yi, Hao Zheng, Haolan Zhan, Yawen Huang, Wei Ji, Yuexiang Li, Yefeng Zheng

1. Westlake University, China 2. Jarvis Research Center, Tencent Youtu Lab, China

Problem Statement

- Domain-generalized semantic segmentation infers robust pixel-wise semantic predictions on arbitrary unseen target domains when a segmentation model is trained on the source domain.
- The key challenge lies in the stability of the scene content, while the domain gap is caused by the style variation.

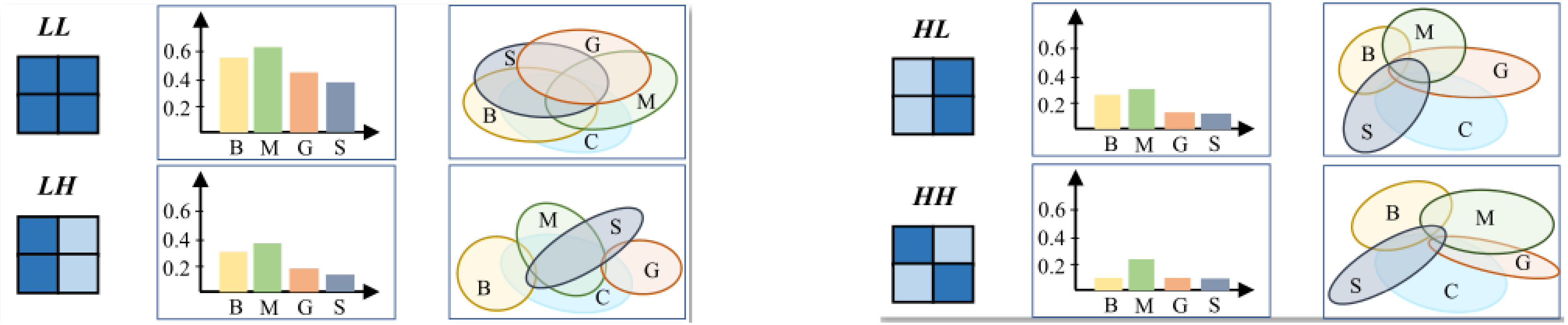


Problem Statement

- Analysis of frozen VFM features after Haar wavelet transform:

-- Low-frequency component exhibits a higher correlation & smaller domain gap

-- High-frequency components exhibit a lower correlation & larger domain gap



What's New?

- We propose a Frequency-Adapted learning scheme, dubbed FADA, to fine-tune VFMs for domain-generalized semantic segmentation.
- The proposed FADA, aided by the Haar wavelet guidance to mine the style-invariant property of VFM, is versatile to a variety of VFMs.
- Experimentally, the proposed FADA significantly outperforms the state-of-the-art DGSS methods, and yields an improvement up to 2.9% mIoU over the contemporary REIN.

Methodology

- What's Haar wavelet?

Definition 1. Haar Scaling Function. *Given an input signal x, the Haar scaling function is mathematically defined as*

$$\phi(t) = \begin{cases} 1 & 0 \le t < 1 \\ 0 & \text{otherwise} \end{cases}. \quad (1)$$

Given the space of all functions of the form $\sum_{k \in \mathbb{Z}} a_k \phi(x-k)$ as V_0, where $k \in \mathbb{Z}$ is an arbitrary integer, and $a_k \in \mathbb{R}$. As each element of V_0 is zero outside a bounded set, such a function $a_k \phi(x-k)$ has *finite or compact support*.

Definition 2. Basis of the Step Function Space. *Given an arbitrary nonnegative integer $j \in \mathbb{Z}_0^+$, Let V_j denote the step function space at the level j, which is spanned by the set*

$$\{\cdots, \phi(2^j x + 1), \phi(2^j x), \phi(2^j x - 1), \cdots\}. \quad (2)$$

Definition 3. Haar Wavelet Function. *The Haar wavelet is the function $\psi(x) = \phi(2x) - \phi(2x-1)$.*

Haar Wavelet Transformation Haar wavelet pooling (Porwik and Lisowska 2004) enables the separation from the low-frequency component to high-component. It has four kernels, namely, LL^{T}, LH^{T}, HL^{T}, HH^{T}, given by

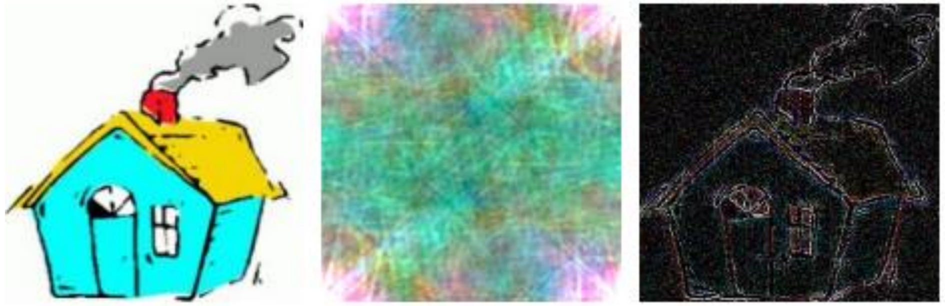
$$L^{\mathrm{T}} = \frac{1}{\sqrt{2}}[1 \quad 1], H^{\mathrm{T}} = \frac{1}{\sqrt{2}}[-1 \quad 1]. \quad (2)$$

Level-1 Haar Wavelet

Methodology

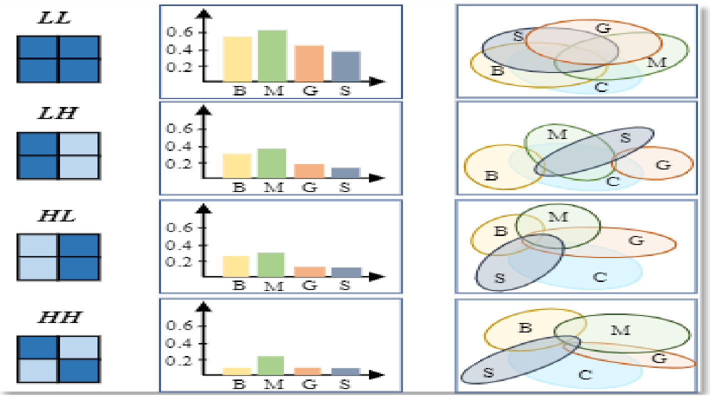
- Why Haar Wavelet benefits DGSS?

Frequency analysis MATTERS to differentiate the content and details in an image



■ Low-frequency: object contour, content, shape,···
content

■ High-frequency: details, illumination, noise, ···
style

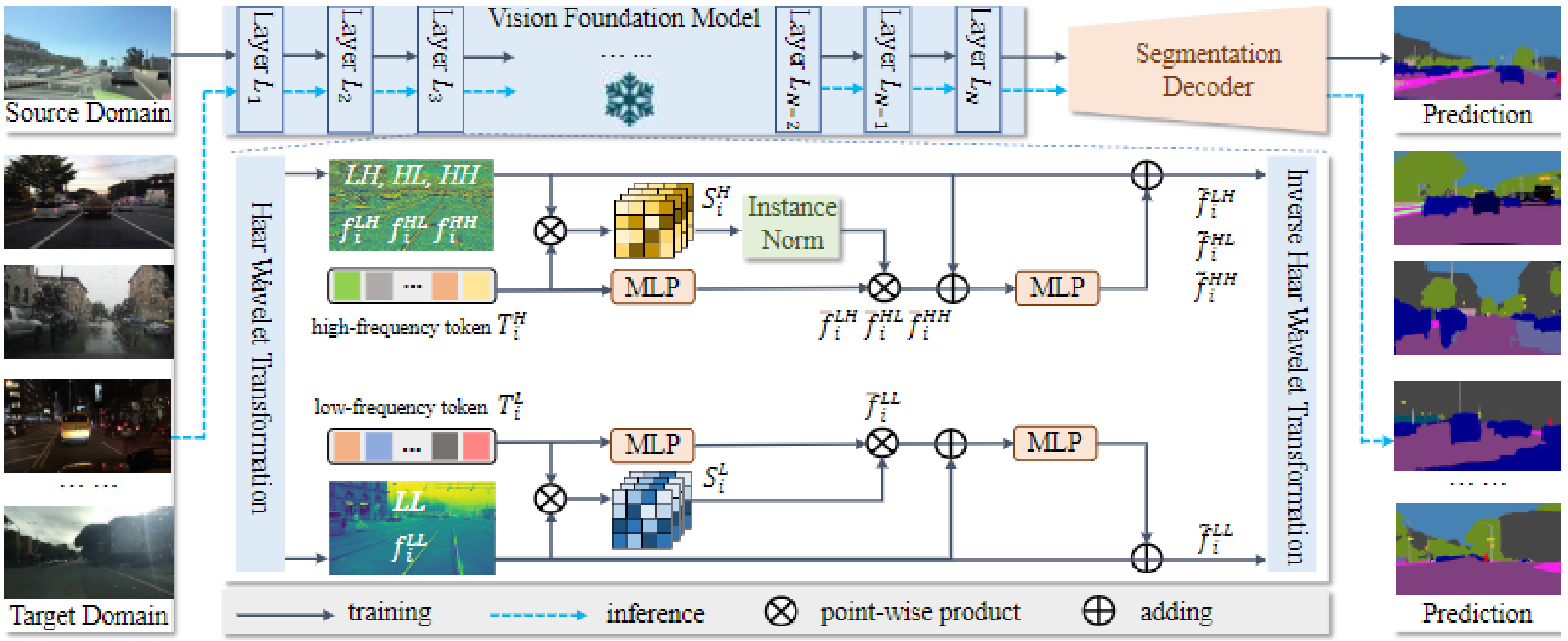


■ LL component of VFM feature for content

■ LH, HL, HH components for style

Methodology

- Framework Overview: Frequency Adapted fine-tuning scheme (FADA)



Three key steps:

- Frequency Decomposition
- Low-Frequency Adaptation
- High-Frequency Adaptation

Instance norm:

for decouple purpose

Baseline adapter:

Wei et. al. Stronger, Fewer, & Superior: Harnessing Vision Foundation Models for Domain Generalized Semantic Segmentation. CVPR 2024

Experiment

- Comparison with State-of-the-art

Method	Proc. & Year	Trained on GTA5 (G)				Trained on SYNTHIA (S)				Trained on Cityscapes (C)			
		→C	→B	→M	→S	→C	→B	→M	→G	→B	→M	→G	→S
ResNet based:													
IBN [44]	ECCV2018	33.85	32.30	37.75	27.90	32.04	30.57	32.16	26.90	48.56	57.04	45.06	26.14
IW [45]	CVPR2019	29.91	27.48	29.71	27.61	28.16	27.12	26.31	26.51	48.49	55.82	44.87	26.10
Iternorm [27]	CVPR2019	31.81	32.70	33.88	27.07	-	-	-	-	49.23	56.26	45.73	25.98
DRPC [63]	ICCV2019	37.42	32.14	34.12	28.06	35.65	31.53	32.74	28.75	49.86	56.34	45.62	26.58
ISW [13]	CVPR2021	36.58	35.20	40.33	28.30	35.83	31.62	30.84	27.68	50.73	58.64	45.00	26.20
GTR [46]	TIP2021	37.53	33.75	34.52	28.17	36.84	32.02	32.89	28.02	50.75	57.16	45.79	26.47
DIRL [60]	AAAI2022	41.04	39.15	41.60	-	-	-	-	-	51.80	-	46.52	26.50
SHADE [66]	ECCV2022	44.65	39.28	43.34	-	-	-	-	-	50.95	60.67	48.61	27.62
SAW [47]	CVPR2022	39.75	37.34	41.86	30.79	38.92	35.24	34.52	29.16	52.95	59.81	47.28	28.32
WildNet [33]	CVPR2022	44.62	38.42	46.09	31.34	-	-	-	-	50.94	58.79	47.01	27.95
AdvStyle [67]	NeurIPS2022	39.62	35.54	37.00	-	37.59	27.45	31.76	-	-	-	-	-
SPC [28]	CVPR2023	44.10	40.46	45.51	-	-	-	-	-	-	-	-	-
BlindNet [2]	CVPR2024	45.72	41.32	47.08	31.39	-	-	-	-	51.84	60.18	47.97	28.51
Mask2Former:													
HGFormer* [18]	CVPR2023	-	-	-	-	-	-	-	-	53.4	66.9	51.3	33.6
CMFormer [7]	AAAI2024	55.31	49.91	60.09	43.80	44.59	33.44	43.25	40.65	59.27	71.10	58.11	40.43
VFM based:													
DIDEX* [42]	WACV2024	62.0	54.3	63.0	-	-	-	-	-	-	-	-	-
REIN* [58]	CVPR2024	66.4	60.4	66.1	48.86†	48.59†	44.42†	48.64†	46.97†	63.54†	74.03†	62.41†	48.56†
FADA (Ours)	-	**68.23**	**61.94**	**68.09**	**50.36**	**50.04**	**45.83**	**49.86**	**48.26**	**65.12**	**75.86**	**63.78**	**49.75**
		↑1.83	↑1.54	↑1.99	↑1.50	↑1.45	↑1.41	↑1.22	↑1.29	↑1.58	↑1.83	↑1.37	↑1.19

Experiment

- Ablation Studies

Table 2: Ablation studies on each component of the proposed FADA. LL, LH, HL and HH denote the f_i^{LL}, f_i^{LH}, f_i^{HL} and f_i^{HH} components, respectively. ✓ refers to that fine-tuning is implemented. Evaluation metric is mIoU in %.

Frequency Components				Trained on CityScapes (C)				Trained on SYNTHIA (S)			
LL	LH	HL	HH	→ B	→ M	→ G	→ S	→ C	→ B	→ M	→ G
✘	✘	✘	✘	62.43	73.05	61.29	47.61	48.03	43.27	47.85	46.02
✓	✘	✘	✘	63.85	74.16	62.04	48.68	48.79	44.81	48.96	47.35
✓	✓	✘	✘	64.04	74.89	62.95	48.92	49.18	45.07	49.13	48.07
✓	✓	✓	✘	64.69	75.16	63.20	49.35	49.62	45.37	49.50	48.16
✓	✓	✓	✓	**65.12**	**75.86**	**63.78**	**49.75**	**50.04**	**45.83**	**49.86**	**48.26**

Table 3: Ablation studies of the rank r on generalization performance. Evaluation metric is mIoU in %.

Method	Trained on Cityscapes (C)			
	→ B	→ M	→ G	→ S
4	64.21	74.96	62.79	48.68
8	64.73	75.18	63.06	49.03
16	65.12	**75.86**	**63.78**	**49.75**
32	**65.28**	75.34	63.56	49.42
64	64.85	75.12	62.38	49.64

Table 4: Generalization ability test of the proposed FADA on different VFM models. One decimal result is reported and compared following prior references.

Backbone	Fine-tune Method	Trainable Params*	mIoU			
			Citys	BDD	Map	Avg.
CLIP [62]	Full	304.15M	51.3	47.6	54.3	51.1
	Freeze	0.00M	53.7	48.7	55.0	52.4
	REIN [69]	2.99M	57.1	54.7	60.5	57.4
	FADA	11.65M	**58.7**	**55.8**	**62.1**	**58.9**
SAM [39]	Full	632.18M	57.6	51.7	61.5	56.9
	Freeze	0.00M	57.0	47.1	58.4	54.2
	REIN [69]	4.51M	59.6	52.0	62.1	57.9
	FADA	16.59M	**61.0**	**53.2**	**63.4**	**60.0**
EVA02 [20]	Full	304.24M	62.1	56.2	64.6	60.9
	Freeze	0.00M	56.5	53.6	58.6	56.2
	REIN [69]	2.99M	65.3	60.5	64.9	63.6
	FADA	11.65M	**66.7**	**61.9**	**66.1**	**64.9**
DINOV2 [53]	Full	304.20M	63.7	57.4	64.2	61.7
	Freeze	0.00M	63.3	56.1	63.9	61.1
	REIN [69]	2.99M	66.4	60.4	66.1	64.3
	FADA	11.65M	**68.2**	**62.0**	**68.1**	**66.1**

Table 5: Generalization performance comparison on the four adverse condition domains from ACDC dataset [65]. CityScapes as the source domain. Top three results are highlighted as **best**, second and third, respectively.

Method	Trained on Cityscapes (C)				
	→ Fog	→ Night	→ Rain	→ Snow	mean
ResNet Based:					
IBN [55]	63.8	21.2	50.4	49.6	43.7
Iternorm [30]	63.3	23.8	50.1	49.9	45.3
IW [56]	62.4	21.8	52.4	47.6	46.6
ISW [14]	64.3	24.3	56.0	49.8	48.1
Transformer Based:					
ISSA [45]	67.5	33.2	55.9	53.2	52.5
HGFormer [19]	69.9	52.7	72.0	68.6	67.2
Mask2Former [13]	73.4	37.1	63.6	62.5	58.0
CMFormer [8]	77.8	33.7	67.6	64.3	60.9
VFM based:					
REIN† [69]	79.5	55.9	72.5	70.6	69.6
Ours	**80.2**	**57.4**	**75.0**	**73.5**	**71.5**
	↑0.7	↑1.5	↑2.5	↑2.9	↑1.9

Experiment

- Cross-domain Visualization

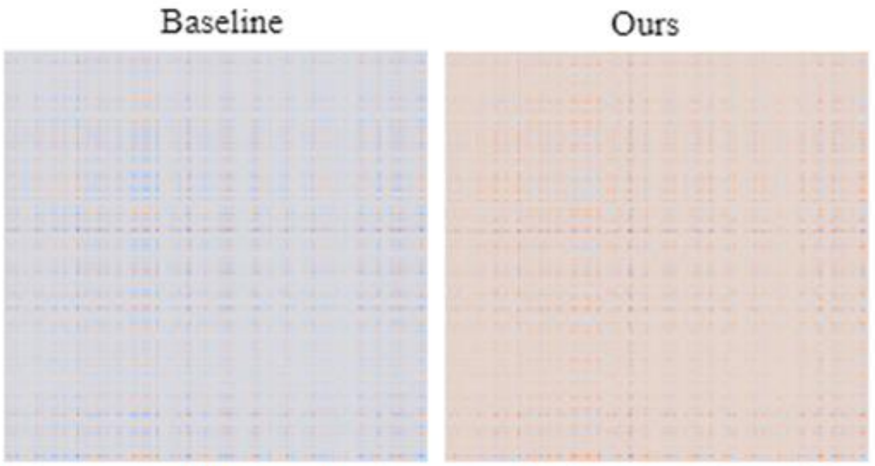


Figure 3: Channel-wise correlation matrix of the last layer VFM feature between source domain (C) and unseen domain (B). The brighter a cell is, the higher response.

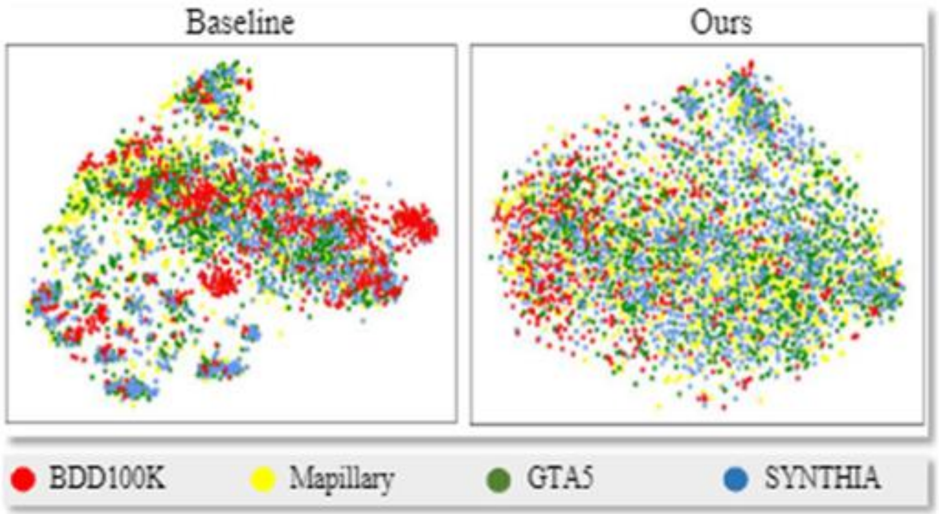
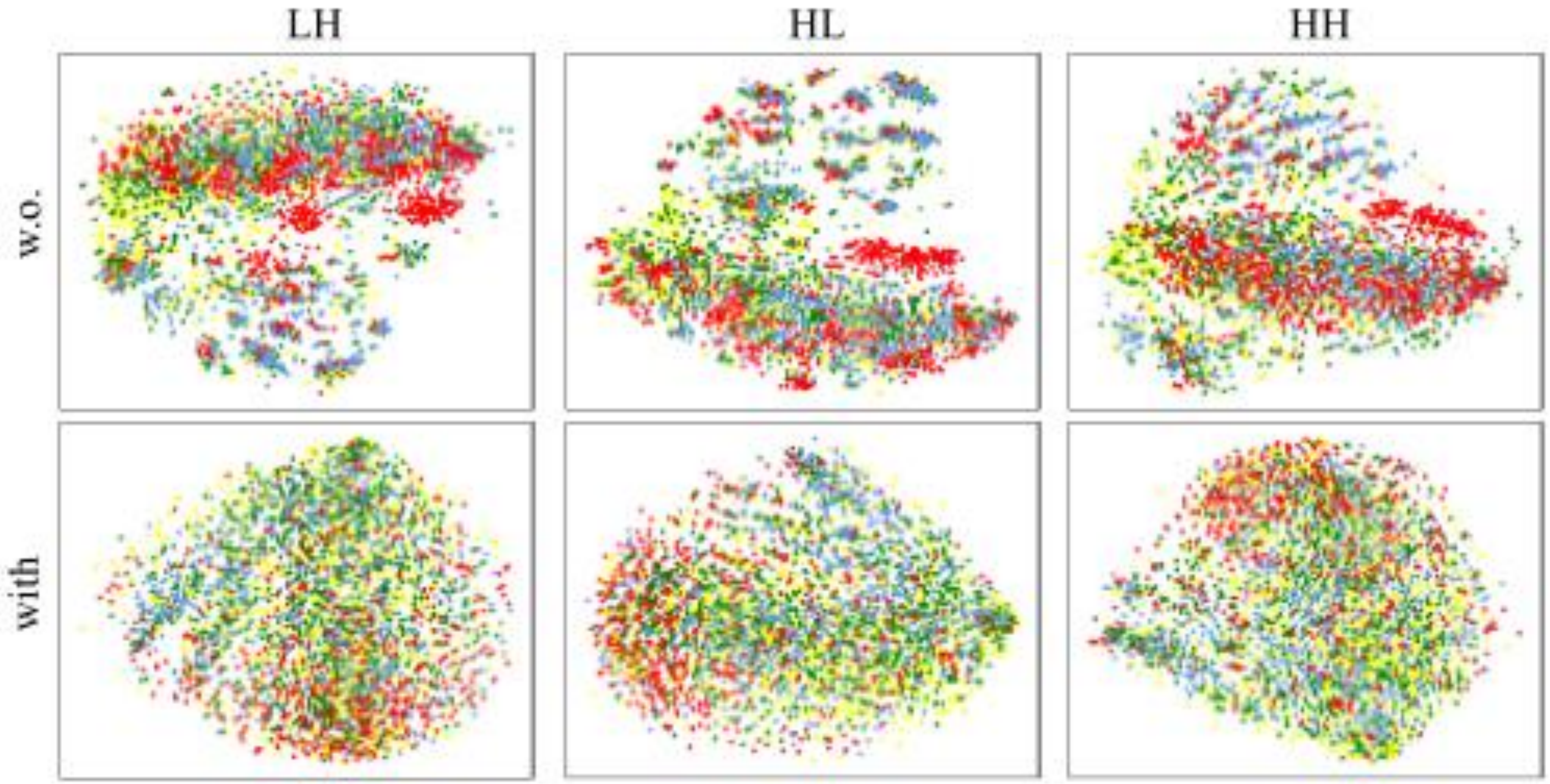


Figure 4: t-SNE visualization. Feature embedding is extracted from the last VFM layer. Left: baseline; Right: ours.

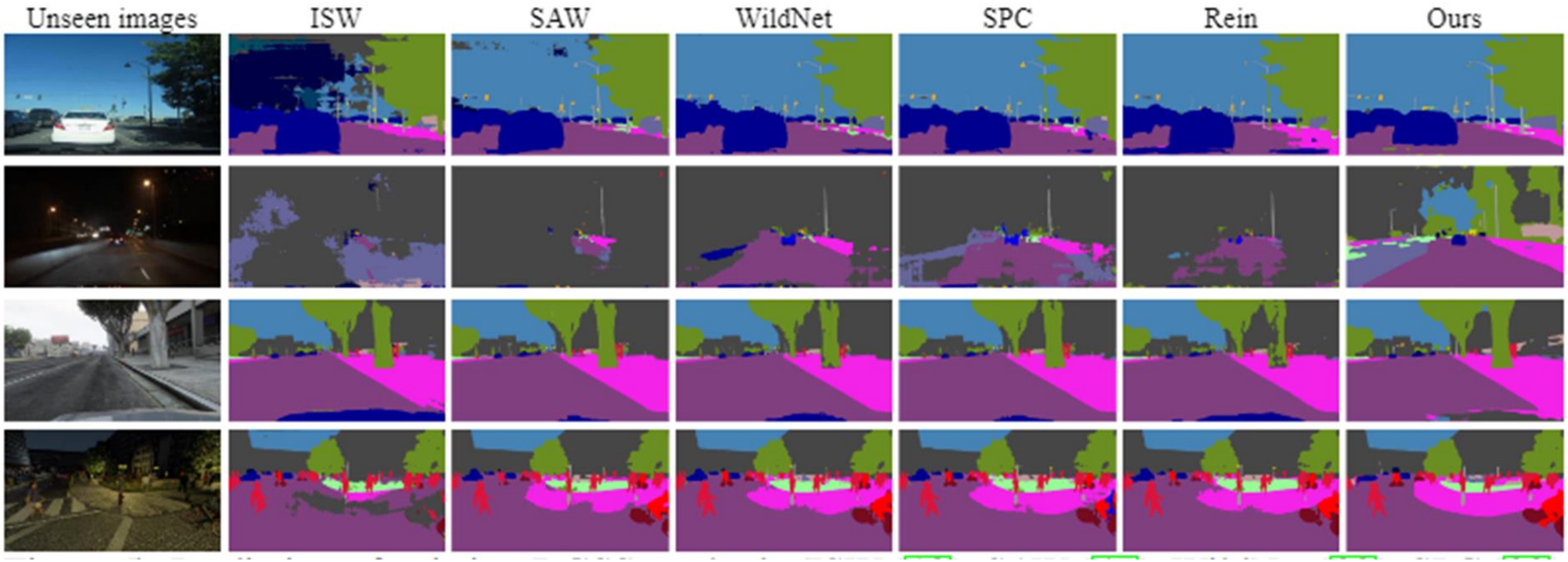
Experiment

- Understanding the benefit of instance normalization



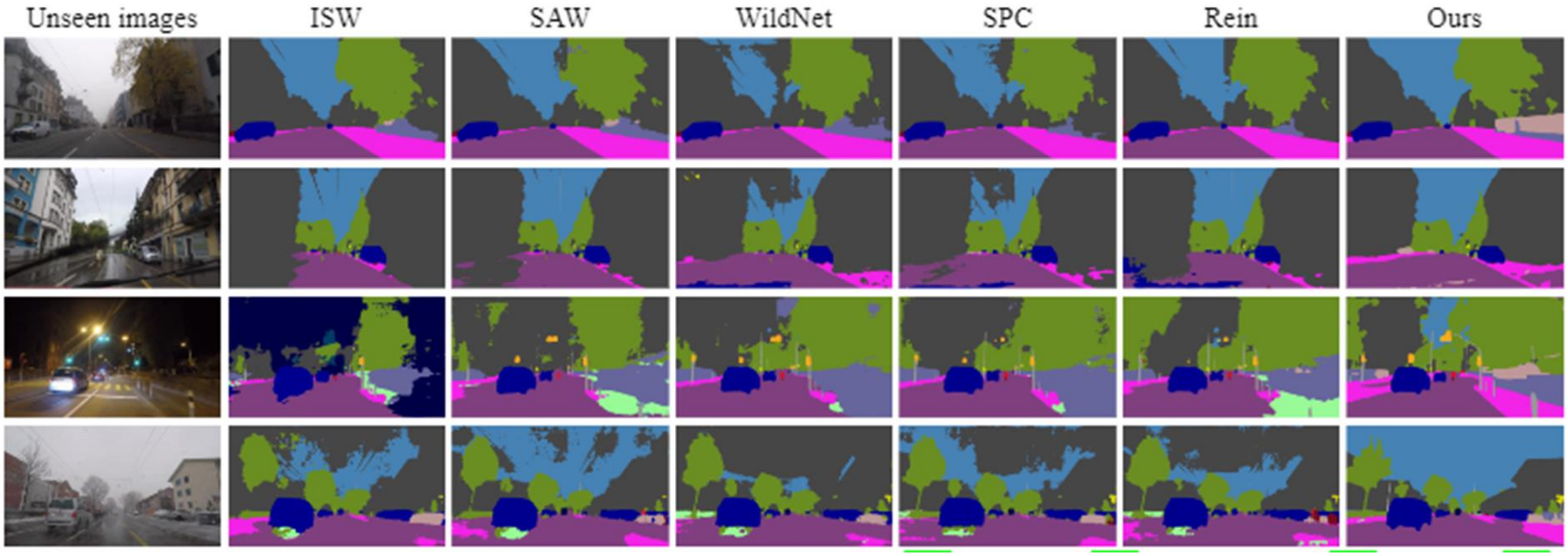
Experiment

- Visual prediction



Experiment

- Visual prediction



Thanks for your attention!