JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

中国科学技术大学
University of Science and Technology of China

Style Projected Clustering for Domain Generalized Semantic Segmentation

Wei Huang[1*] Chang Chen[2†] Yong Li[2] Jiacheng Li[1]
Cheng Li[2] Fenglong Song[2] Youliang Yan[2] Zhiwei Xiong[1]
[1]University of Science and Technology of China [2]Huawei Noah's Ark Lab
{weih527,jclee}@mail.ustc.edu.cn, zwxiong@ustc.edu.cn,
{chenchang25,liyong156,licheng89,songfenglong,yanyouliang}@huawei.com

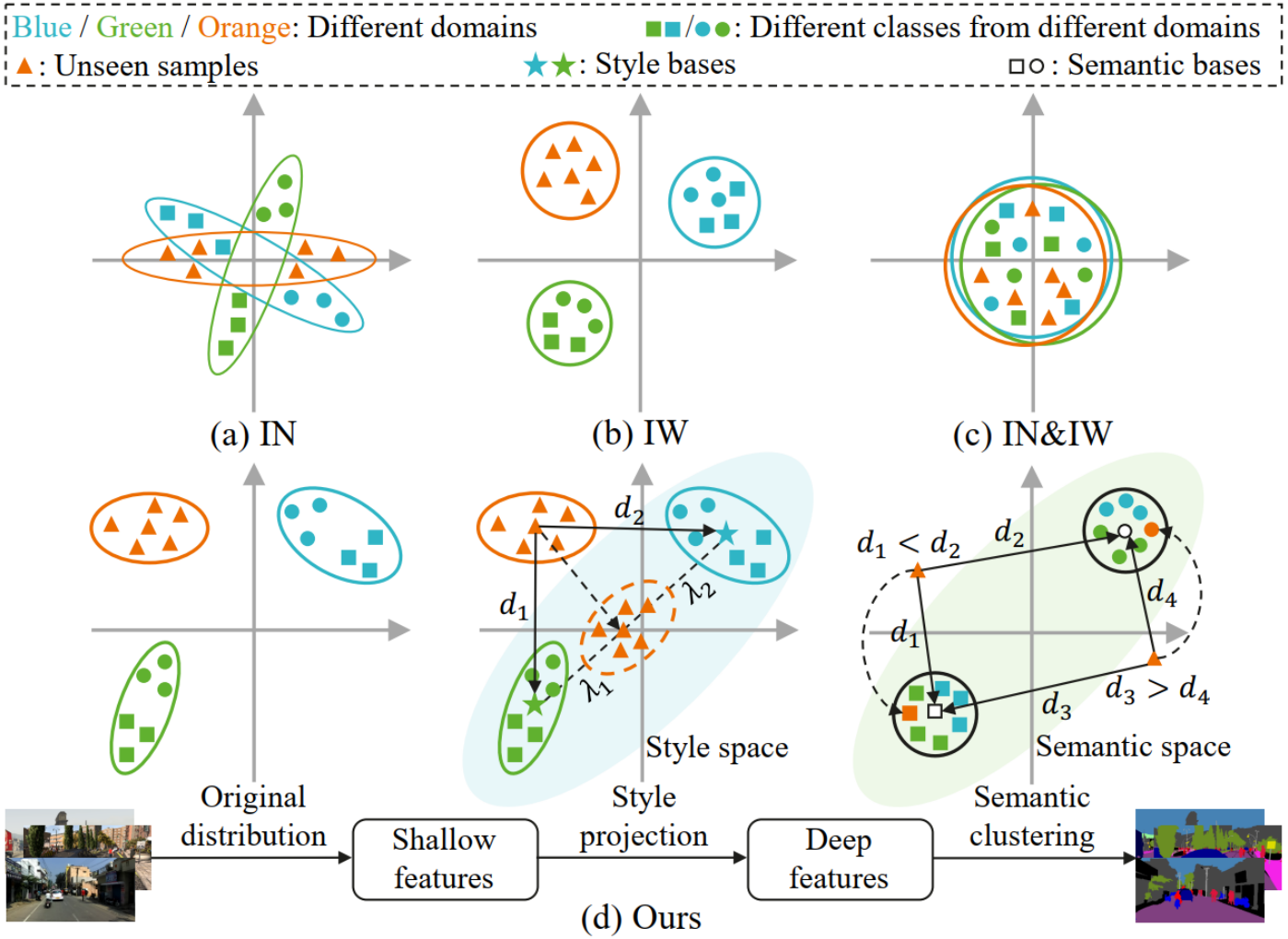
Presentation Date	Poster Number	Tag	Submission
June 20, 2023	291	TUE-AM-291	4546

VIDAR
Visual Information
Discovery And Recovery

HUAWEI

Summary

➢ Domain generalized semantic segmentation

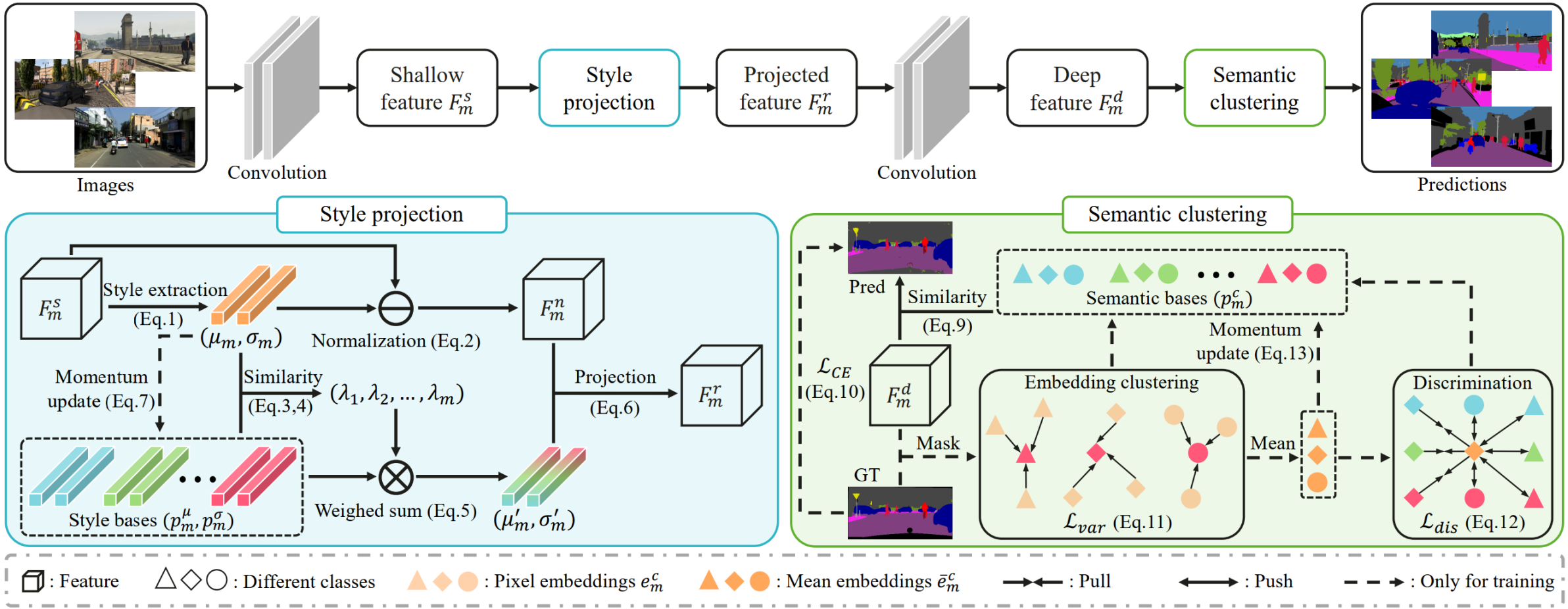


➢ Instance normalization/whitening (IN/IW) regularize image features from different domains to a canonical space (a-c).

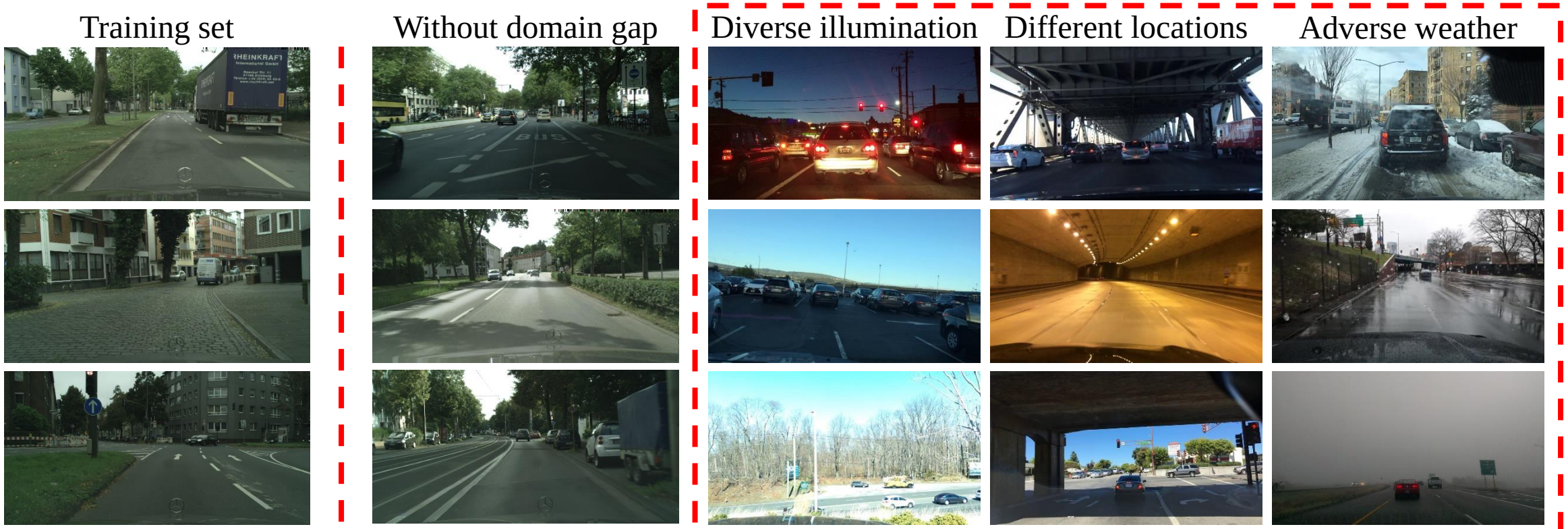
➢ Our method builds style and semantic representation spaces based on the data from known domains (d).

Summary

➢ Domain generalized semantic segmentation

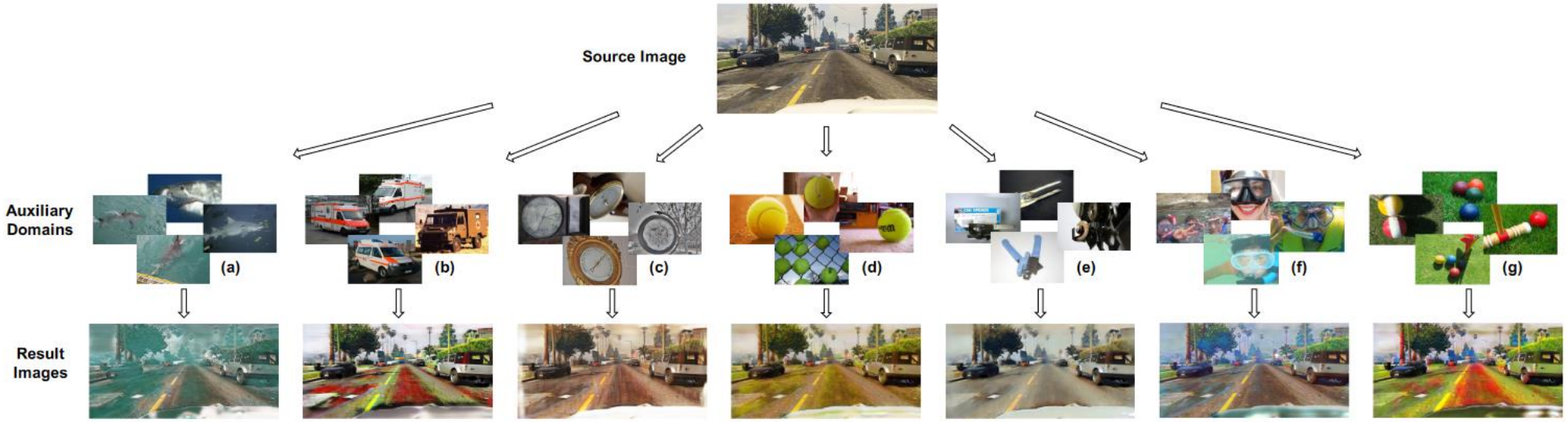


- **Fully supervised methods:** Limited generalizability in different real urban-scenes
- **Domain adaptation methods:** Only work for one specific real urban-scene or dataset
- **Domain generalization methods:** Improve the robustness of DNNs to arbitrary unseen scenarios



➢ Style augmentation

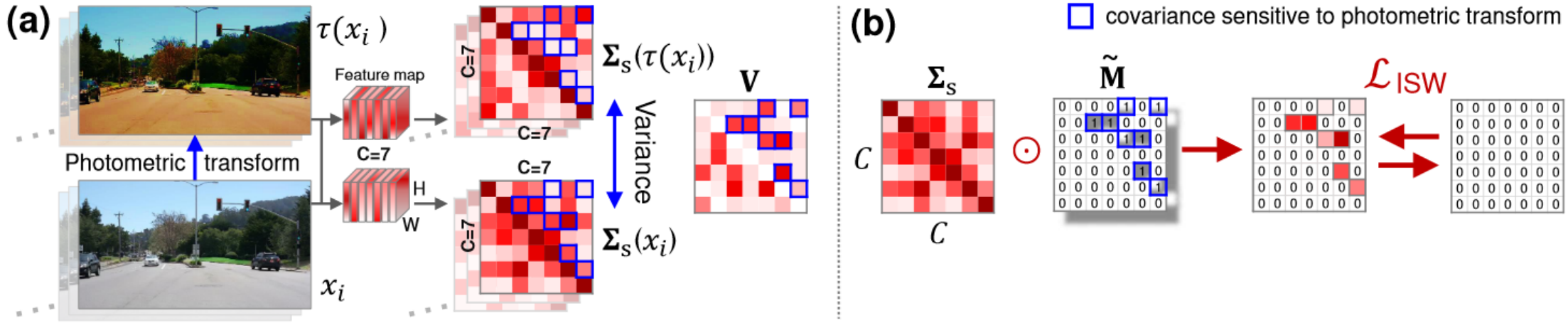
- Method: Using style transfer algorithms to transfer the style of natural images to the training datasets, enriching the style of the training datasets
- Disadvantage: There are still distribution discrepancies between the migrated style and the real scene, and it still cannot cover all the real scenes



Xiangyu Yue, et al. Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization without Accessing Target Domain Data. ICCV, 2019

➢ Instance normalization/whitening

- Method: Using instance normalization or whitening operations (removing the interrelationships between feature channels) to eliminate specific style information of images
- Disadvantage: There is difficult to perfectly decouple the style and content information, and the content information is often also eliminated simultaneously

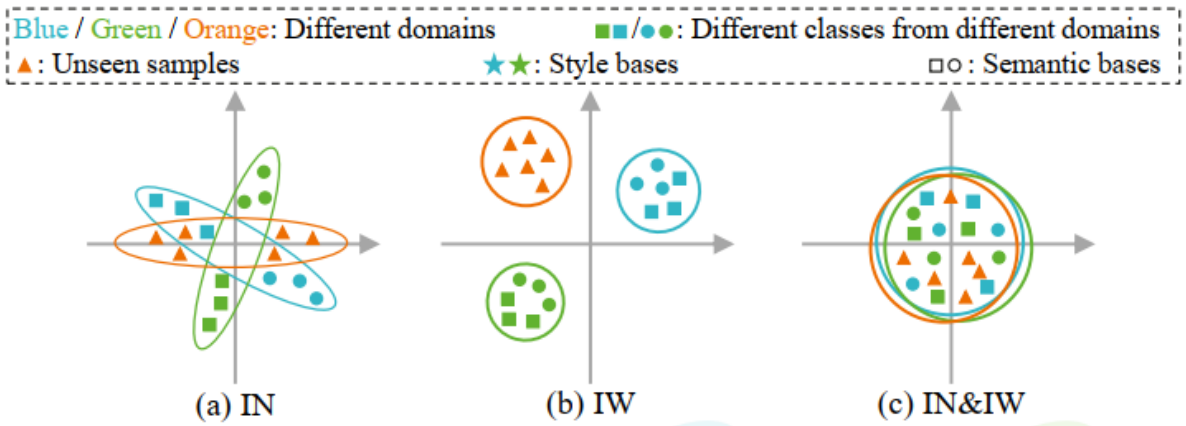


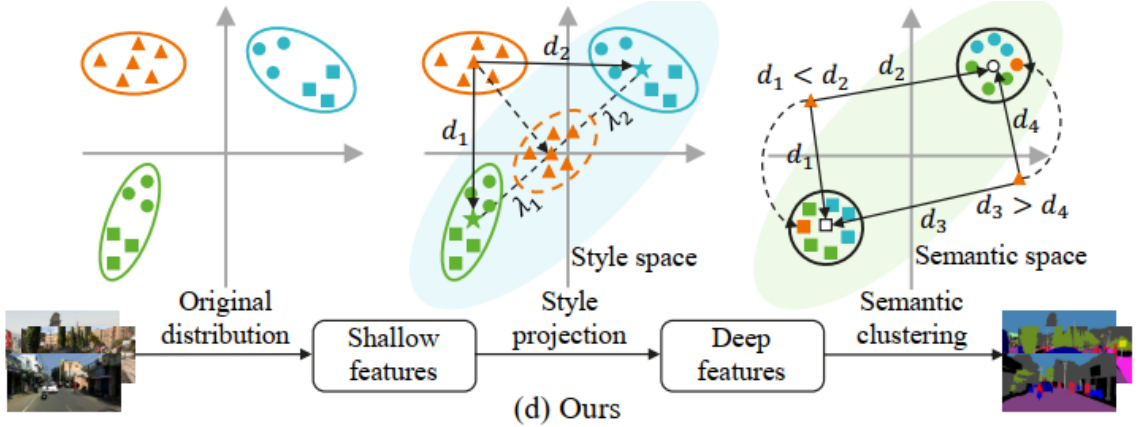
Sungha Choi, et al. RobustNet: Improving Domain Generalization in Urban-Scene Segmentation via Instance Selective Whitening, CVPR, 2021

➢ Motivations

- Using the existing training datasets to represent the image style in the unknown scene
  - ➢ No need to expand the style of training datasets
  - ➢ Preserved the style information of the training data

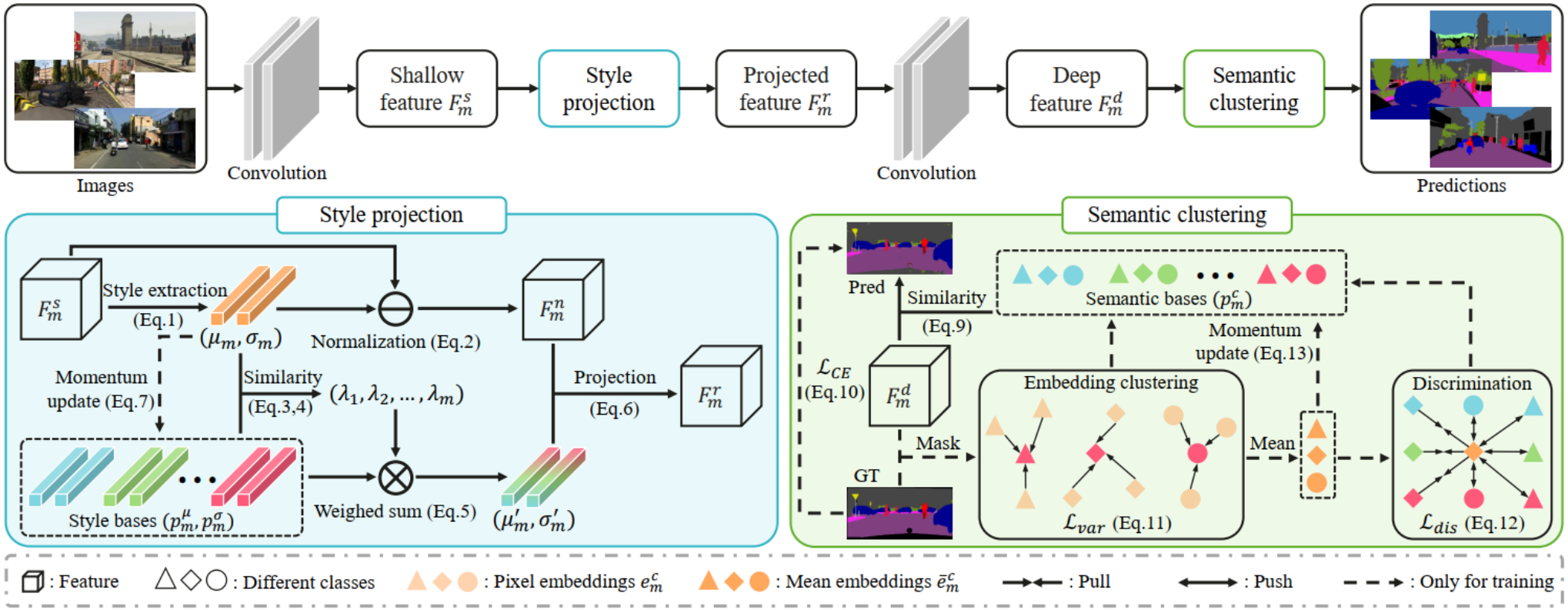
- Using clustering operations to achieve semantic classification of pixels
  - ➢ Preserved the semantic (content) information of the training data
  - ➢ Clustering has better generalization than the learnable classifier



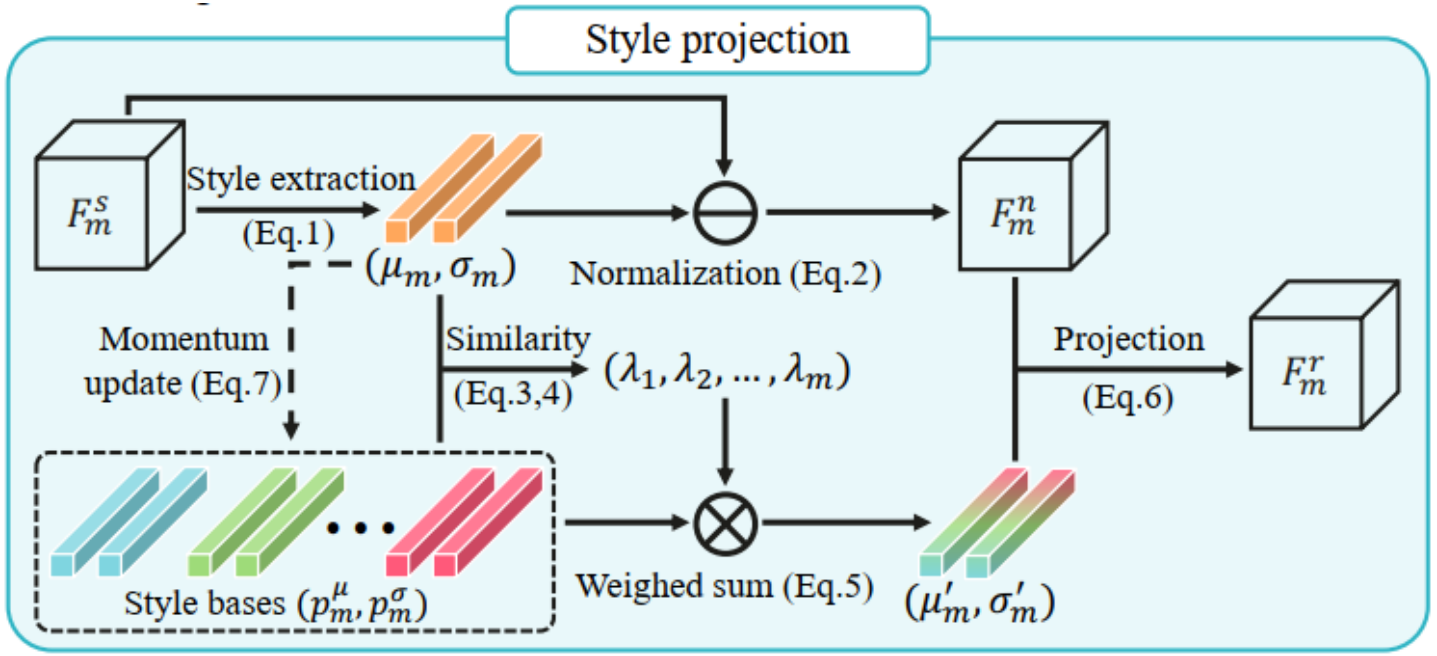


➢ Overall framework

- ➢ Style projection
- ➢ Semantic clustering



➢ Style projection

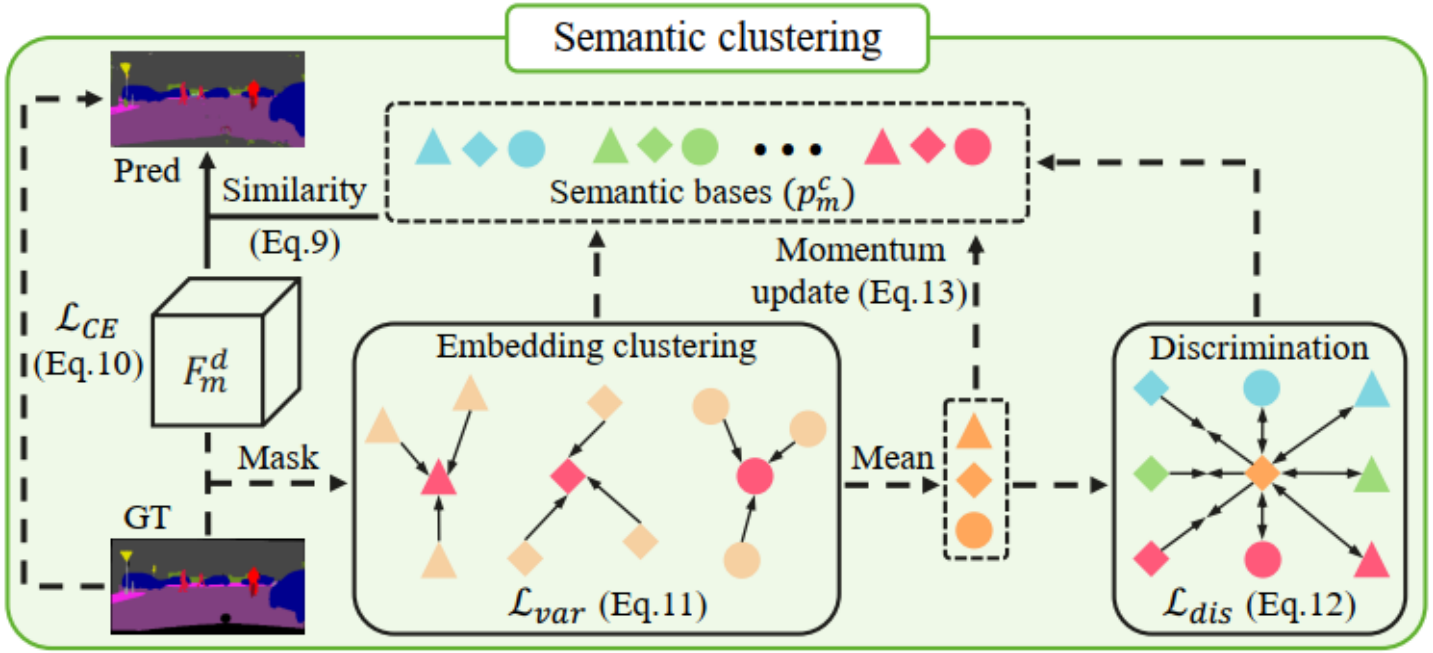


- Extract the style information of the input image, i.e., the mean and variance of the shallow features
- Calculate the similarity between the input image style and the style bases
- Inject the weighted combination of style bases into normalized features

$$d_m = ||\mu_m - p_m^\mu||_2^2 + ({\sigma_m}^2 + {p_m^\sigma}^2 - 2\sigma_m p_m^\sigma), \quad \lambda_m = \frac{exp(1/(1+d_m))}{\sum_{m=1}^{M} exp(1/(1+d_m))},$$

$$\mu_m' = \sum_{m=1}^{M} \lambda_m p_m^\mu, \ \sigma_m' = \sum_{m=1}^{M} \lambda_m p_m^\sigma. \qquad F_m^r = \sigma_m' F_m^n + \mu_m'.$$

➢ Semantic clustering



- Calculate the similarity between each pixel vector and semantic bases
- Select the nearest semantic base as the category of each pixel
- During the training process, three kinds of losses are adopted to supervise the training of the network

$$\mathcal{L}_{CE} = -\frac{1}{H_d W_d}\sum_{h=1}^{H_d}\sum_{w=1}^{W_d}\sum_{c=1}^{C} y_m log(v(c|e)), \quad \mathcal{L}_{var} = \frac{1}{MC}\sum_{m=1}^{M}\sum_{c=1}^{C}(1 - e_m^c p_m^c)^2.$$

$$\mathcal{L}_{dis} = \frac{1}{M}\sum_{p_{c+}} -log\frac{exp(\bar{e}_m^c p_{c+}/\tau)}{exp(\bar{e}_m^c p_{c+}/\tau) + \sum_{p_{c-}} exp(\bar{e}_m^c p_{c-}/\tau)},$$

➢ Datasets

Datasets	Syn/Real	Site	Resolution	Training Num	Test Num
GTAV	Synthetic	-	1914 x 1052	12403	6382
Synthia	Synthetic	-	1280 x 760	6580	2820
IDD	Real-world	Indian	1678 x 968	6993	981
Cityscapes	Real-world	German	2048 x 1024	2975	500
BDD100K	Real-world	American	1280 x 720	7000	3000
Mapillary	Real-world	Worldwide	1920 x 1080	18000	2000

➢ Settings

- Single source (GTAV)
- Two sources (GTAV + Synthia)
- Three sources (GTAV + Synthia + IDD)

➢ Metric

- mIoU

➢ Quantitative comparisons on three different target domains

Methods	Publication	Cityscapes	BDD100K	Mapillary	Avg.-$\mathcal{T}$	GTAV	Synthia	Avg.-$\mathcal{S}$	Avg.-$\mathcal{A}$
Baseline†	-	35.46	25.09	31.94	30.83	68.48	67.99	68.24	45.79
IBN-Net† [40]	ECCV 2018	35.55	32.18	38.09	35.27	69.72	66.90	68.31	48.49
RobustNet† [5]	CVPR 2021	37.69	34.09	38.49	36.76	68.26	68.77	68.52	49.46
Baseline‡	-	33.42	29.07	32.19	31.56	69.63	63.93	66.78	45.65
MLDG‡ [26]	AAAI 2018	38.84	31.95	35.60	35.46	64.61	51.69	58.15	44.54
PintheMem‡ [20]	CVPR 2022	44.51	38.07	42.70	41.76	65.85	54.49	60.17	49.12
Baseline	-	36.03	28.15	32.61	32.26	69.30	67.61	68.46	46.65
SAN-SAW [42]	CVPR 2022	42.13	37.74	42.91	40.93	63.98	62.58	63.28	49.87
WildNet [25]	CVPR 2022	43.65	39.90	43.28	42.28	68.05	63.98	66.02	51.77
WildNet* [25]	CVPR 2022	39.33	34.76	41.06	38.38	69.70	62.11	65.91	49.39
Ours	-	**46.36**	**43.18**	**48.23**	**45.92**	**72.46**	**74.87**	**73.67**	**57.02**

Source (G+S) → Target (C, B, M)

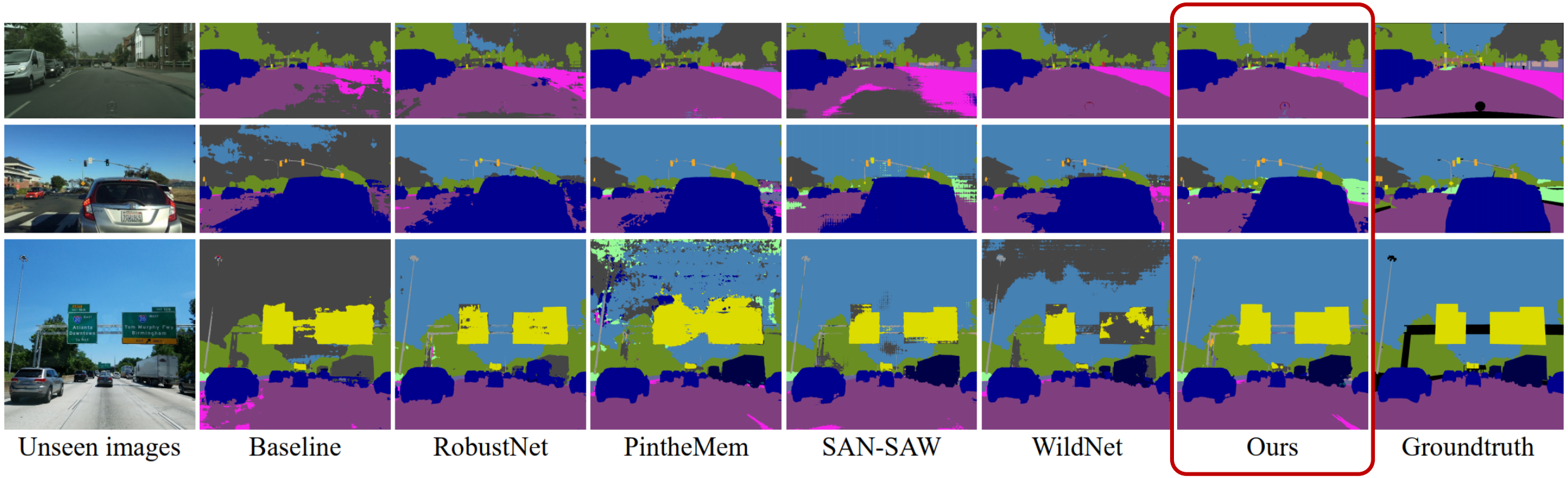
Methods	Cityscapes	BDD100K	Mapillary	Avg.-$\mathcal{T}$	GTAV	Synthia	IDD	Avg.-$\mathcal{S}$	Avg.-$\mathcal{A}$
Baseline‡	52.51	47.47	54.70	51.56	70.31	67.13	71.56	69.67	60.61
IBN-Net‡ [40]	54.39	48.91	56.06	53.12	70.73	63.68	71.02	68.48	60.80
RobustNet‡ [5]	54.70	49.00	56.90	53.53	70.06	66.40	71.02	69.16	61.35
MLDG‡ [26]	54.76	48.52	55.94	53.07	69.53	59.79	67.73	65.68	59.38
PintheMem‡ [20]	56.57	50.18	58.31	55.02	69.99	62.99	67.58	66.85	60.94
Baseline	54.16	46.24	55.57	51.99	68.35	65.12	70.07	67.85	59.92
SAN-SAW [42]	54.89	46.50	56.38	52.59	64.49	64.76	66.37	65.21	58.90
WildNet [25]	55.58	50.31	57.93	54.61	67.65	61.35	70.07	66.36	60.48
WildNet* [25]	53.61	48.92	56.18	52.90	70.98	59.69	64.52	65.06	58.98
Ours	**57.91**	**53.26**	**61.61**	**57.59**	**74.64**	**78.35**	**76.07**	**76.35**	**66.97**

Source (G+S+I) → Target (C, B, M)

Methods	C	B	M	Avg.-$\mathcal{T}$
Baseline	28.95	25.14	28.18	27.42
IBN-Net [40]	33.85	32.30	37.75	34.63
RobustNet [5]	36.58	35.20	40.33	37.37
Baseline	31.60	26.70	29.00	29.10
MLDG [26]	36.70	32.10	32.20	33.67
PintheMem [20]	41.00	34.60	37.40	37.67
Baseline	29.32	25.71	28.33	27.79
SAN-SAW [42]	39.75	37.34	41.86	39.65
Baseline	35.16	29.71	31.29	32.05
WildNet [25]	**44.62**	38.42	**46.09**	43.04
Baseline	32.01	26.04	29.35	29.13
WildNet* [25]	40.10	34.82	39.38	38.10
Ours	44.10	**40.46**	45.51	**43.36**

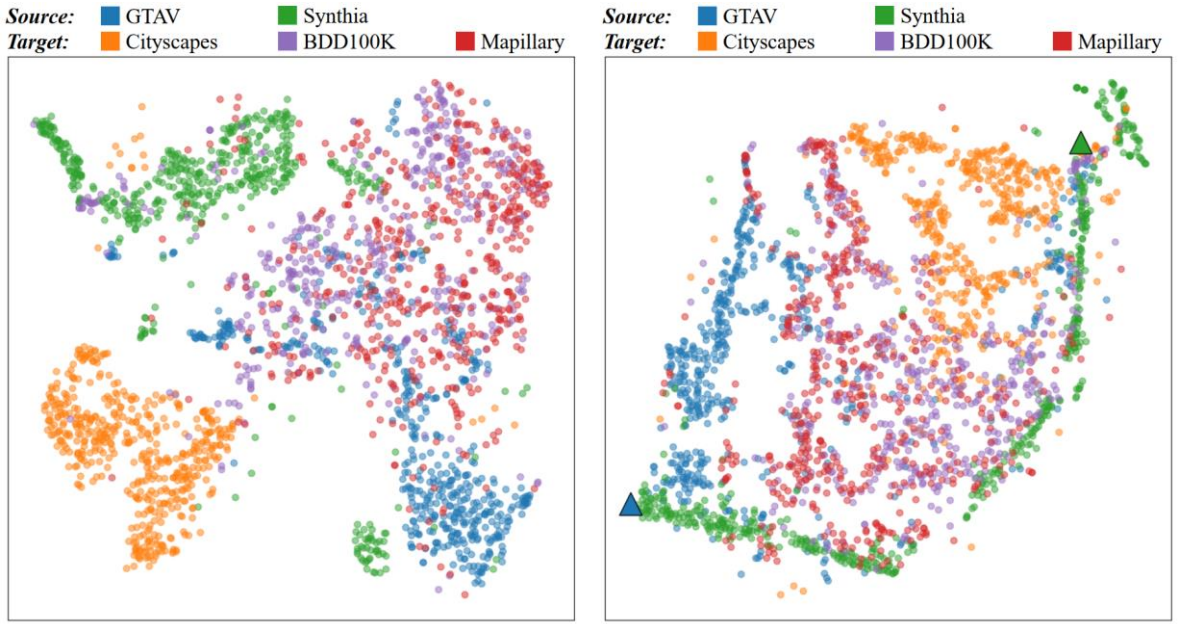
Source (G) → Target (C, B, M)

➢ Qualitative comparisons on three different target domains



Source (G+S) → Target (C, B, M)

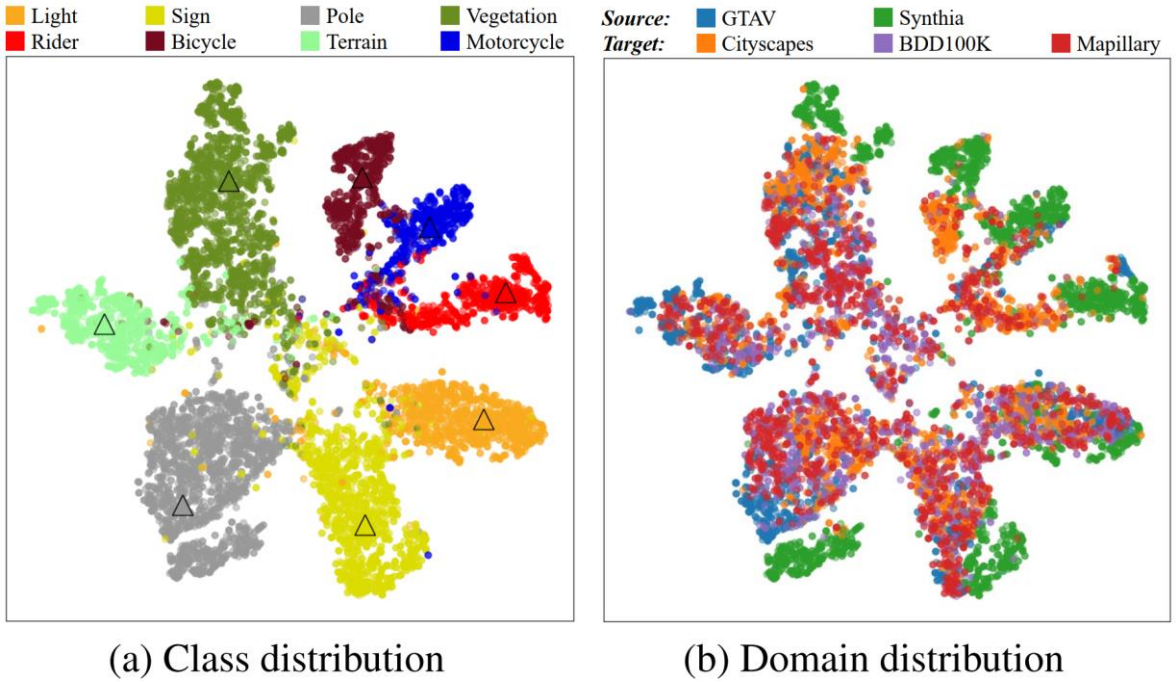
➢ Distribution analysis--Style distribution



(a) Before projection (b) After projection

- Before projection, the style distribution of different domains is well separated before style projection
- After projection, their style distribution is approximately constrained between two style bases

➢ Distribution analysis--Semantic distribution



(a) Class distribution (b) Domain distribution

- Pixel samples belonging to the same class are well clustered while those belonging to different classes are well separated
- These pixel samples from different domains are well clustered according to their classes

- A novel **style projected clustering** method for domain generalized semantic segmentation, which achieves the style and semantic representation of unseen images based on known data

- **Style projection** projects arbitrary unseen styles into the style representation space of source domains

- **Semantic clustering** predicts the class of each pixel by the minimal similarity distance to semantic bases

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

中国科学技术大学
University of Science and Technology of China

Thanks for your listening!

VIDAR
Visual Information
Discovery And Recovery

HUAWEI