Supplement Material: Calibration-based Dual Prototypical Contrastive Learning Approach for Domain Generalization Semantic Segmentation

Muxin Liao
Guangdong Key Laboratory of Intelligent Information Processing
College of Electronics and Information Engineering
Shenzhen University
Shenzhen, China
liaomuxin2020@email.szu.edu.cn

Shishun Tian
Guangdong Key Laboratory of Intelligent Information Processing
College of Electronics and Information Engineering
Shenzhen University
Shenzhen, China
stian@szu.edu.cn

Yuhang Zhang
Guangdong Key Laboratory of Intelligent Information Processing
College of Electronics and Information Engineering
Shenzhen University
Shenzhen, China
zhangyuhang2019@email.szu.edu.cn

Guoguang Hua
Guangdong Key Laboratory of Intelligent Information Processing
College of Electronics and Information Engineering
Shenzhen University
Shenzhen, China
huaguoguang2021@email.szu.edu.cn

Wenbin Zou*
Guangdong Key Laboratory of Intelligent Information Processing
College of Electronics and Information Engineering
Shenzhen University
Shenzhen, China
wzou@szu.edu.cn

Xia Li
Guangdong Key Laboratory of Intelligent Information Processing
College of Electronics and Information Engineering
Shenzhen University
Shenzhen, China
lixia@szu.edu.cn

Table 1: Performance comparison in terms of mIoU (%) between domain generalization methods in the architecture of the ResNet-50 [6]. The best results are marked in bold and the second-best results are underlined.

Methods	Backbone	Mean	Train on G, S, and I		
			→C	→B	→M
IBN [16]	ResNet-50	53.1	54.4	48.9	56.1
MLDG [10]		53.1	54.8	48.5	55.9
ISW [4]		53.5	54.7	49.0	56.9
TSMLDG [32]		50.7	53.0	46.4	52.8
PinMem [8]		<u>55.0</u>	<u>56.6</u>	<u>50.2</u>	<u>58.3</u>
Ours		**59.5**	**61.1**	**54.8**	**62.5**

1 EXPERIMENTS

1.1 Datasets

Our approach is evaluated on five standard single-source benchmarks and a standard multi-source benchmark. Five standard single-source benchmarks contain "G→{S, C, M, B}", "S→{G, C, M, B}", "C→{G, S, M, B}", "B→{G, S, C, M}", and "M→{G, S, C, B}". The two standard multi-source benchmarks are "{G, S}→{C, B, M}" and "{G, S, I}→{C, B, M}". The "G" and "S" mean GTA5 [20] and SYNTHIA [21] datasets which are two synthetic datasets. The "C", "B", "M", and "I" mean the Cityscapes [5], BDD [31], Mapillary [14], and IDD [26] datasets which are three real-world datasets.

*Corresponding author.

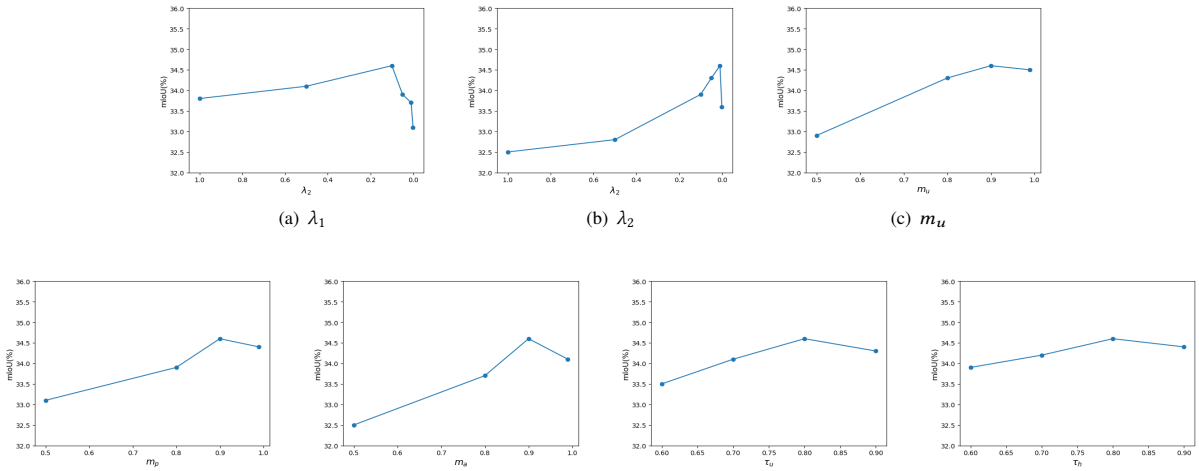
Muxin Liao and Shishun Tian contributed equally to this work.

1.2 Implementation Details

In our experiments, the ISW [4] is adopted as baseline. The ResNet-50 [6], ShuffleNetV2 [12], and MobileNetV2 [22] are utilized in DeepLabV3+ [3] as the segmentation network. We follow the data augmentations of previous works, including ISW [4], SAN [17], WildNet [9], PinMemory [8], and SHADE [33]. Specifically, color jittering (brightness of 0.4, contrast of 0.4, saturation of 0.4, and hue of 0.1), Gaussian blur, random cropping, random horizontal flipping, and random scaling with the range of $[0.5, 2.0]$ are used in our approach. The input images of five datasets are cropped to the resolution of 768×768. The mean Intersection-Over-Union value ($mIoU = \frac{TP}{TP+FP+FN}$) is utilized as the metric of evaluation, where TP, FP, and FN are denoted as the predicted pixels numbers of true positive, false positive, and false negative. The Stochastic Gradient Decent (SGD) optimizer with an initial learning rate of $1e-2$ and a momentum of 0.9 is leveraged to optimize our backbone network. In the training stages, the learning rate is adjusted by the power of 0.9 according to the polynomial learning rate scheduler and the maximum number of iterations is set to 40k steps. Moreover, the weight coefficients m_p, m_a, m_u, τ_u, τ_h, λ_1, and λ_2 are set as 0.9, 0.9, 0.9, 0.8, 0.8, 0.1, and 0.01 for all experiments.

1.3 Comparison with Multi-source Domain Generalization Methods

Our proposed approach is trained on three source domains to compare with recent methods, including IBN [16], MLDG [10], ISW [4], TSMLDG [32], and PinMem [8], for further verifying the effectiveness of our proposed approach. As shown in Table 1, our proposed approach respectively achieves an improvement of 4.5% in average mIoU over the PinMem [8]. Thus, with richer source domains during the training process, our proposed approach can better



(a) λ_1 (b) λ_2 (c) m_u

(d) m_p (e) m_a (f) τ_a (g) τ_h

Figure 1: Quantitative comparisons for the hyperparameters, including λ_1, λ_2, m_u, m_p, m_a, τ_a, and τ_h. All experiments are based on the ShuffleNetV2 [12] in the G→{C, B, M, S} setting. We show the average performance of four datasets for all cases.

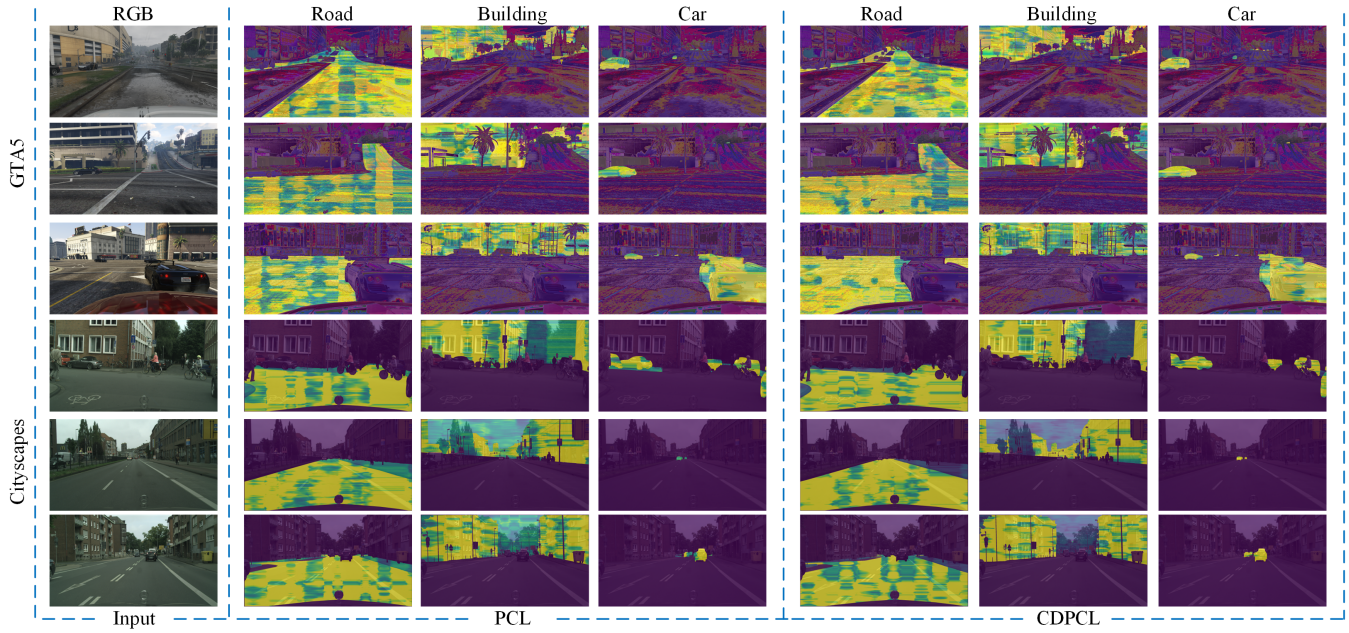


Figure 2: Activation maps visualization comparison of the weight matrix of the learned class-wise features between the PCL and our proposed approach (CDPCL) in the GTA5 to Cityscapes task.

learn class-wise domain-invariant features from the prototypes of different domains to improve the generalization ability for semantic segmentation.

1.4 Performance Dependency of Hyper-parameters

We further investigate the sensitivity of our approach to the hyper-parameters $\lambda_1, \lambda_2, m_u, m_p, m_a, \tau_a$, and τ_h. For these hyper-parameters, we use a grid search over $m_u, m_p, m_a \in \{0.5, 0.8, 0.9, 0.99\}$, $\tau_a, \tau_h \in \{0.6, 0.7, 0.8, 0.9\}$, and $\lambda_1, \lambda_2 \in \{1.0, 0.5, 0.1, 0.05, 0.01, 0.001\}$. The results are shown in Figure 1. We set the weight m_p, m_a, m_u, τ_u,

Table 2: The similarity of different classes between the learned class-wise features and the source domain prototypes. The first row represents the learned class-wise features and the first column represents the source domain prototypes.

	PCL					CDPCL			
Class	Road	Sidewalk	Building	Car	Class	Road	Sidewalk	Building	Car
Road	0.8598	0.5131	0.4979	0.5368	Road	0.9182	0.4531	0.4157	0.3148
Sidewalk	0.5990	0.8502	0.5972	0.4974	Sidewalk	0.5130	0.8904	0.4691	0.4211
Building	0.5011	0.3995	0.8511	0.3019	Building	0.4853	0.3154	0.9003	0.2117
Car	0.4996	0.5018	0.5810	0.9051	Car	0.3548	0.4183	0.4519	0.9331

Table 3: The similarity of different classes between the learned class-wise features and the augmented domain prototypes. The first row represents the learned class-wise features and the first column represents the augmented domain prototypes.

	PCL					CDPCL			
Class	Road	Sidewalk	Building	Car	Class	Road	Sidewalk	Building	Car
Road	0.7998	0.6131	0.3377	0.4957	Road	0.8512	0.4937	0.2903	0.3304
Sidewalk	0.6982	0.7573	0.3984	0.5965	Sidewalk	0.5701	0.8210	0.2965	0.3899
Building	0.4005	0.3002	0.8007	0.3012	Building	0.2418	0.2419	0.8411	0.1944
Car	0.5190	0.4921	0.2910	0.8793	Car	0.3941	0.3823	0.2740	0.8999

τ_h, λ_1, and λ_2 as 0.9, 0.9, 0.9, 0.8, 0.8, 0.1, and 0.01 to achieve the best performance. We fix hyperparameters and train our model on all settings.

Table 4: The discrepancy between the learned class-wise features and the prototypes of the source domain.

Class	PCL	CDPCL
Road	0.0406	0.0337
Sidewalk	0.0489	0.0504
Building	0.0516	0.0493
Car	0.0331	0.0289

Table 5: The discrepancy between the learned class-wise features and the prototypes of the augmented domain.

Class	PCL	CDPCL
Road	0.0571	0.0458
Sidewalk	0.0494	0.0383
Building	0.0508	0.0462
Car	0.0414	0.0364

1.5 The Analysis of The Distribution Discrepancy Change

We analyze the distribution discrepancy change between the learned class-wise features and the different domains before and after applying the proposed approach from two aspects, including qualitative and quantitative experiments.

Qualitative Results. We visualize activation maps of the weight matrix of the learned class-wise features between the conventional prototypical contrastive learning (PCL) and the proposed approach (CDPCL), which are shown in Figure 2. In the visualization of the PCL, some other classes are activated in the unseen domain (Cityscapes). For example, the classes of "vegetation" and "sky" are activated in the activation maps of the "building". These results demonstrate that there is still domain discrepancy between the learned class-wise features and the prototypes of the unseen domain (Cityscapes). Compared with the visualization of the PCL, the visualization of our proposed approach (CDPCL) achieves a significant improvement. It demonstrates that the difference between the learned class-wise features and the prototypes of the unseen domain (Cityscapes) is reduced.

Quantitative Results. We respectively use the cosine similarity and the Manhattan distance to measure the discrepancy changes between the learned class-wise features and the different domains before and after applying the proposed approach. The discrepancy changes of the "road", "sidewalk", "building", and "car" classes are shown in Table 2, Table 3, Table 4, and Table 5.

First, we compare the discrepancy measured by the cosine similarity. The discrepancy between the learned class-wise features and the source domains is shown in Table 2. The discrepancy between the learned class-wise features and the augmented domains is shown in Table 3. From Table 2 and Table 3, compared with the conventional PCL, the similarity of different classes is reduced while the similarity of the same class is significantly increased by using the proposed approach (CDPCL).

Second, we compare the discrepancy measured by the Manhattan distance. The discrepancy between the learned class-wise features and the source domains is shown in Table 4. The discrepancy between the learned class-wise features and the augmented domains is shown in Table 5. From Table 4, compared with the conventional PCL, the discrepancy between the learned class-wise features and the prototypes of the source domain is reduced in the classes of "road", "building", and "car" by using the proposed approach (CDPCL). From Table 5, compared with the conventional PCL, the discrepancy between the learned class-wise features and the prototypes of the augmented domain is reduced in the classes of "road", "sidewalk", "building", and "car" by using the CDPCL. In addition, although the discrepancy of the "sidewalk" class between the learned class-wise

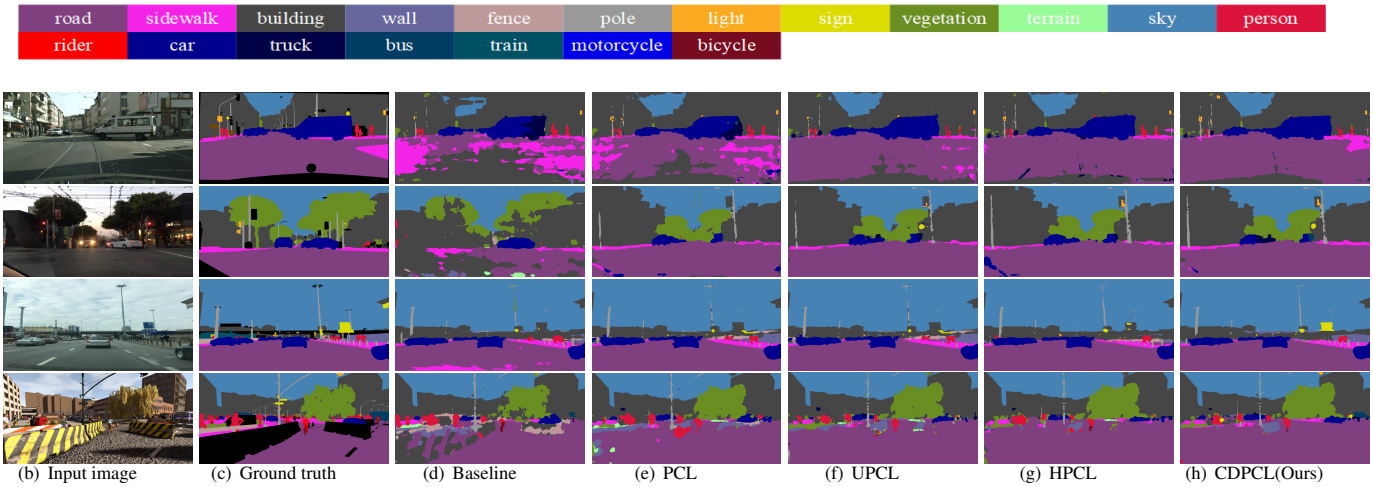


Figure 3: Visualization comparison of ablation experiments in the G→{C, B, M, S} task. The visualization results from four datasets (including Cityscapes, BDD, Mapillary, and SYNTHIA) are respectively shown in four rows.

features and the prototypes of the source domain is increased, the discrepancy of the "sidewalk" class between the learned class-wise features and the prototypes of the augmented domain is significantly decreased. We argue that this phenomenon is caused by the big gap between the "sidewalk" class prototypes of the source and augmented domains, which means the "sidewalk" class prototypes of the source domain are uncertain. Thus, a small weight is assigned to the "sidewalk" class prototypes of the source domain.

In conclusion, from these qualitative and quantitative experiments, the proposed approach can reduce the discrepancy between the learned class-wise features and the prototypes of different domains.

1.6 More Qualitative Results

Figure 3 show qualitative comparison results of the ablation study. As shown in Figure 3, compared with the results of the baseline and the PCL, the classes of "road", "sidewalk", "sign", and "car" are segmented more accurately in the UPCL, the HPCL, and the CDPCL. These results demonstrate that our proposed approach can better learn class-wise domain-invariant features by reducing the domain discrepancy between the learned class-wise features and the prototypes of different domains.

1.7 Comparison with Unsupervised Domain Adaptation Methods.

In order to know whether our proposed approach is up to the performance standard of unsupervised domain adaptation semantic segmentation (UDASS) methods, our proposed approach is compared with current UDASS methods, including AdaptSeg [24], DPR [25], ADVENT [27], DADA [28], CLAN [11], IntraDA [15], FDA [30], PixMatch [13], BiMal [23], and ProCA [7], on the GTA5→Cityscapes domain adaptation settings as shown in Table 6.The UDASS methods have inherent performance superiority over DGSS methods since UDASS methods can access the unlabeled target domain during the training process. For a fair comparison, all methods without using the self-training strategy adopt DeepLabV2 [2] with ResNet-101 [6] are compared. As shown in Table 6, the performance of our approach outperforms many UDASS methods, except for SOTA methods ProCA [7] and PixMatch [13]. The performance of our proposed approach only respectively degrades 0.7% and 0.2% in terms of mIoU than these two methods. It demonstrates that our proposed approach is even on par with the state-of-the-art UDASS methods.

Table 6: Comparison results between ours and Domain Adaptation methods on GTA5→Cityscapes. DA and DG denote Domain Adaption and Domain Generalization respectively.

Backbone	Task	Method	Publication	Access Tgt	mIoU(%)
ResNet-101	DA	AdaptSeg	CVPR 2018	√	41.4
		ADVENT	CVPR 2019	√	43.8
		DPR	ICCV 2019	√	46.5
		CLAN	CVPR 2019	√	43.2
		DADA	CVPR 2019	√	47.3
		IntraDA	CVPR 2020	√	46.3
		FDA	CVPR 2020	√	44.6
		PixMatch	CVPR 2021	√	48.3
		BiMal	ICCV 2021	√	47.3
		ProCA	ECCV 2022	√	**48.8**
	DG	Ours		×	48.1

2 RELATED WORK

2.1 Uncertainty-guided Methods for Domain Generalization

In this section, we discuss recent uncertainty-guided methods for domain generalization. Cai et al. [1] propose a Bayesian CNN-based framework to estimate the model uncertainty for guiding an iterative

self-training method. Qiao et al. [19] propose a Bayesian meta-learning framework that aims to increase the capacity of input and label spaces from the source domain by using an uncertainty-guided augmentation strategy. Peng et al. [18] propose an uncertainty-guided domain generalization method to quantify the generalization uncertainty which is used to guide the feature and label augmentation strategies. Xiao et al. [29] propose a probabilistic framework via variational Bayesian inference to learn domain-invariant features by incorporating uncertainty into neural network weights.

Different from these uncertainty-guided domain generalization methods, we propose an uncertainty-guided prototypical contrastive learning to estimate an uncertainty probability matrix for calibrating the weights of the source domain prototypes during the prototypical contrastive learning.

REFERENCES

[1] Minjie Cai, Feng Lu, and Yoichi Sato. 2020. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. 14392–14401.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

[4] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. 2021. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11580–11590.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[7] Zhengkai Jiang, Yuxi Li, Ceyuan Yang, Peng Gao, Yabiao Wang, Ying Tai, and Chengjie Wang. 2022. Prototypical contrast adaptation for domain adaptive semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIV*. Springer, 36–54.

[8] Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. 2022. Pin the Memory: Learning to Generalize Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4350–4360.

[9] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. 2022. WildNet: Learning Domain Generalized Semantic Segmentation from the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9936–9946.

[10] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[11] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2507–2516.

[12] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*. 116–131.

[13] Luke Melas-Kyriazi and Arjun K Manrai. 2021. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12435–12445.

[14] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. 2017. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*. 4990–4999.

[15] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. 2020. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3764–3773.

[16] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. 2018. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 464–479.

[17] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. 2022. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2594–2605.

[18] Xi Peng, Fengchun Qiao, and Long Zhao. 2022. Out-of-domain generalization from a single source: An uncertainty quantification approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[19] Fengchun Qiao and Xi Peng. 2021. Uncertainty-guided model generalization to unseen domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6790–6800.

[20] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for data: Ground truth from computer games. In *European conference on computer vision*. Springer, 102–118.

[21] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3234–3243.

[22] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[23] Thanh-Dat Truong, Chi Nhan Duong, Ngan Le, Son Lam Phung, Chase Rainwater, and Khoa Luu. 2021. Bimal: Bijective maximum likelihood approach to domain adaptation in semantic scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8548–8557.

[24] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. 2018. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7472–7481.

[25] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. 2019. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1456–1465.

[26] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. 2019. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1743–1751.

[27] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2517–2526.

[28] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. 2019. Dada: Depth-aware domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7364–7373.

[29] Zehao Xiao, Jiayi Shen, Xiantong Zhen, Ling Shao, and Cees Snoek. 2021. A bit more bayesian: Domain-invariant learning with uncertainty. In *International Conference on Machine Learning*. PMLR, 11351–11361.

[30] Yanchao Yang and Stefano Soatto. 2020. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4085–4095.

[31] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2636–2645.

[32] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. 2022. Generalizable model-agnostic semantic segmentation via target-specific normalization. *Pattern Recognition* 122 (2022), 108292.

[33] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. 2022. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*. Springer, 535–552.