Supplementary Material of Exploring Probabilistic Modeling Beyond Domain Generalization for Semantic Segmentation

I-Hsiang Chen[1,2*,†] Hua-En Chang[1*] Wei-Ting Chen[3] Jenq-Neng Hwang[2] Sy-Yen Kuo[1,4]

[1]National Taiwan University [2]University of Washington [3]Microsoft [4]Chang Gung University

1. Derivations of ELBO

Given the following predictive function:

$$p_{\theta,\phi}(y_t|x_t) = \int p_\theta(y_t|x_t, z)p_\phi(z|x_t)\, dz \tag{1}$$

where z is a latent variable, $p_\phi(z|x_t)$ is the prior, and $p_\theta(y_t|x_t, z)$ is the segmentation model. Our objective is to maximize $\log p_{\theta,\phi}(y_t|x_t)$. For computational tractability, we introduce a variational posterior $q_\varphi(z|x_t, x_s)$ to approximate the true posterior $p(z|x_t, x_s)$.

First, we can rewrite the objective function by introducing $q_\varphi(z|x_t, x_s)$:

$$\begin{aligned}
&\log p_{\theta,\phi}(y_t|x_t) \\
&= \log \int p_\theta(y_t|x_t, z)p_\phi(z|x_t)\, dz \\
&= \log \int p_\theta(y_t|x_t, z)\frac{p_\phi(z|x_t)}{q_\varphi(z|x_t, x_s)}q_\varphi(z|x_t, x_s)\, dz
\end{aligned} \tag{2}$$

Next, we apply Jensen's Inequality [8], noting that the log function is concave and obeys:

$$\log \mathbb{E}[X] \geq \mathbb{E}[\log X]$$

Thus, we can bring the logarithm outside the integral to form a lower bound:

$$\begin{aligned}
&\log p_{\theta,\phi}(y_t|x_t) \\
&\geq \int q_\varphi(z|x_t, x_s) \log \left(p_\theta(y_t|x_t, z)\frac{p_\phi(z|x_t)}{q_\varphi(z|x_t, x_s)} \right) dz
\end{aligned} \tag{3}$$

We then decompose the logarithmic term inside the integrand:

$$\begin{aligned}
&\log \left(p_\theta(y_t|x_t, z)\frac{p_\phi(z|x_t)}{q_\varphi(z|x_t, x_s)} \right) \\
&= \log p_\theta(y_t|x_t, z) + \log \frac{p_\phi(z|x_t)}{q_\varphi(z|x_t, x_s)}
\end{aligned} \tag{4}$$

Hence, the lower bound becomes:

$$\begin{aligned}
\log p_{\theta,\phi}(y_t|x_t) \geq &\int q_\varphi(z|x_t, x_s) \log p_\theta(y_t|x_t, z)\, dz \\
&- \int q_\varphi(z|x_t, x_s) \log \frac{q_\varphi(z|x_t, x_s)}{p_\phi(z|x_t)}\, dz
\end{aligned} \tag{5}$$

Therefore, we obtain the final form of the ELBO:

$$\begin{aligned}
&\log p_{\theta,\phi}(y_t|x_t) \geq \mathbb{E}_{q_\varphi(z|x_t,x_s)} \left[\log p_\theta(y_t|x_t, z)\right] \\
&- \mathbb{KL}\left[q_\varphi(z|x_t, x_s)||p_\phi(z|x_t)\right].
\end{aligned} \tag{6}$$

The ELBO demonstrates that maximizing the left-hand side is equivalent to jointly maximizing the expected log likelihood and minimizing the KL divergence between the variational posterior and the prior.

2. More Experiments

2.1. Robustness to Image Corruptions

We evaluate PDAF on Cityscapes-C [10, 17], which simulates real-world degradation by introducing 16 corruption types across four categories: blur (motion, defocus, glass, Gaussian), noise (Gaussian, impulse, shot, speckle), digital distortions (brightness, contrast, saturation, JPEG), and adverse weather (snow, spattering, fog, frost). We adopt Mask2Former [3] (with Swin-T and Swin-L [12]) and train on Cityscapes [5]. As shown in Table 1, PDAF outperforms previous methods across all corruption types, demonstrating that our method can adaptively compensate for various degradation factors by leveraging the Latent Domain Prior (LDP) as a structured guidance.

2.2. Visualization of Feature Representation

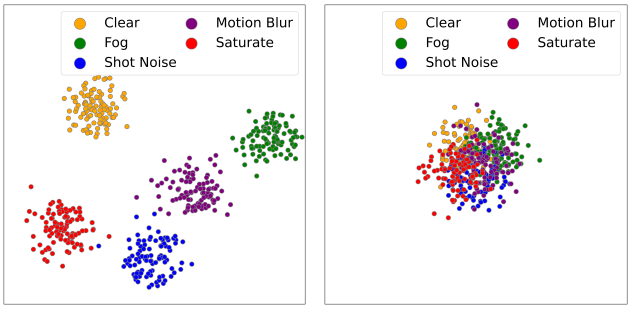
We visualize feature representations under Cityscapes-C [10, 17] to analyze domain shifts induced by degradations. We randomly sample 50 images per degradation and extract features from DeepLabV3Plus [3] (ResNet50 [7]),

* Indicates equal contribution.

† This work was conducted while I-Hsiang Chen was an intern at University of Washington.

Method	backbone	Avg	Blur				Noise				Digital				Weather			
			Motion	Defoc	Glass	Gauss	Gauss	Impul	Shot	Speck	Bright	Contr	Satur	JPEG	Snow	Spatt	Fog	Frost
Mask2Former [3]	Swin-T	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 [6]		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
PDAF		49.88	56.1	57.7	47.8	55.7	23.2	24.1	30.4	55.1	77.6	62.0	74.5	33.2	30.1	63.8	73.2	33.6
Mask2Former [3]	Swin-L	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 [6]		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
PDAF		62.5	64.9	68.3	62.9	66.2	35.6	41.1	45.0	68.3	82.3	73.4	79.3	54.6	56.4	76.6	78.1	46.6

Table 1. **Comparison with existing methods evaluated on synthetic Cityscapes-C [10, 17].**



(a) DeepLabV3Plus (b) PDAF

Figure 1. **Visualization of feature representation based on PDAF and its segmentation backbone (DeepLabV3Plus [3]).**

Figure 2. **Visualization of LDPs under different domains.**

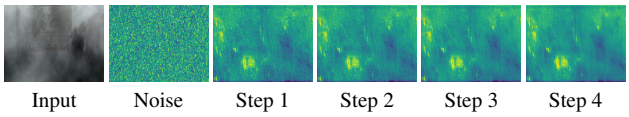


Figure 3. **Visualization of dynamic changing maps.**

trained on Cityscapes [5]. As shown in Figure 1, degradations cause significant domain shifts, forming distinct clusters in feature space. In contrast, PDAF effectively reduces domain shift by leveraging LDP, leading to a more compact feature distribution. These results indicate that PDAF improves feature alignment, enhancing segmentation robustness against challenging degradations.

2.3. Analysis of LDP

Figure 2 compares LDP maps under fog and snow conditions from the same image, demonstrating distinct spatial patterns associated with domain shifts. Furthermore, we apply four synthetic degradations to 50 Cityscapes [5] test images, extract their LDPs, and visualize them with t-SNE. The resulting clusters are clearly separated by degradation type, confirming that LDPs effectively encode

Spatial Resolution	1/64	1/32	1/16	1/8	1/4
Avg. of mIoU	48.83	49.10	49.41	49.14	48.90

Table 2. **Effect of LDP Spatial Scale**

$(\lambda_{\text{task}}, \lambda_{\text{sc}}, \lambda_{\text{prior}})$	Setting	B	M	G	S	Avg.
(0.5, 0.5, 0.1)	Adjust Prior	52.99	61.07	48.11	30.58	48.19
(0.5, 0.5, 10.0)		52.85	60.75	47.86	30.40	47.97
(0.5, 0.5, 1.0)	Balance	53.50	62.93	50.54	30.68	49.41
(5.0, 0.5, 1.0)	Adjust Task	51.98	58.45	46.70	29.33	46.62
(0.5, 5.0, 1.0)		52.75	59.45	47.37	29.81	47.35

Table 3. **Ablation study on the coefficient of loss function.** The experiments are conducted using DeepLabV3Plus [2] with ResNet50 [7], trained on the Cityscapes [5] and evaluated on BDD-100K (B) [18], Mapillary (M) [13], GTAV (G) [14] and SYNTHIA (S) [15].

domain-specific variations. As shown in Table 2, we assess LDPs at multiple spatial scales and observe that the 1/16 resolution, matching the backbone's feature map dimensions, achieves the best performance. Note that we follow the same experimental settings as in the ablation study of the main paper.

2.4. Dynamic Changing Maps of DPE

DPE employs an accelerated diffusion-based optimization [16] for efficient and rapid inference. As shown in the Figure 3, even a single step is sufficient to capture domain-specific patterns. As the number of steps increases, the estimated LDP becomes progressively clearer.

2.5. Analysis of Loss Coefficient

The loss coefficients $(\lambda_{\text{task}}, \lambda_{\text{sc}}, \lambda_{\text{prior}})$ adjusts the objective function by balancing task-specific term, semantic consistency and prior regularization. Based on experimental results, we set the values to $(0.5, 0.5, 1.0)$. As shown in Table 3, increasing a specific λ value enhances focus on the corresponding loss term but distorts the balance between regularization and segmentation, leading to suboptimal domain generalization performance.

2.6. Impact of Pseudo-target Quality

We further leverage DGInStyle [9] for pseudo-target augmentation. The average mIoU improves from 42.85 to

Method	Params (M)	GFLOPS	Time(ms)
DeepLabV3Plus [2]	45.08	278.82	10.01
PDAF	45.43	280.74	11.52

Table 4. **Computational analysis of PDAF and its segmentation network (DeepLabV3Plus [2]) during inference.** Inference time is averaged over 400 trials.

Modeling Variation	Avg. of mIoU
w/o Constraint	46.80
Standard Laplace	47.23
Standard Gaussian	47.79
Learned Gaussian (Ours)	49.41

Table 5. **Comparison of Different Probabilistic Modeling Strategies**

44.24, confirming that higher-fidelity pseudo-targets enhance LDP estimation. In our original experiments, we used only photometric augmentation to ensure fair comparisons with baseline methods.

2.7. Computational Efficiency

To investigate the computational efficiency, we adopt DeepLabV3Plus [2] with a ResNet-50 [7] backbone and conduct our analysis on an NVIDIA Tesla V100 GPU using images at 2048 $\times$ 1024 resolution. As shown in Table 4, PDAF introduces only a marginal increase in parameters and computational overhead while maintaining comparable inference efficiency. Remarkably, our additional modules require only 1.51 ms, about 13.1% the total, indicating that overall deployment speed is largely determined by the chosen backbone. Moreover, the Latent Prior Extractor (LPE), the Domain Compensation Module (DCM) and the Diffusion Prior Estimator (DPE) contribute 0.15M, 0.01M and 0.34M parameters, respectively. Although the LPE and an additional pre-trained segmentation network are utilized during training, these modules can be removed during inference to maintain minimal overhead. In the DPE, each denoising step for estimating LDP requires 0.26 ms and 0.142 GFLOPs, and we adopt 4 steps by default. Notably, since PDAF estimates the LDP rather than an entire feature representation, allowing it remains efficient while adapting to arbitrary target domains.

2.8. Analysis of Probabilistic Modeling

To analyze the impact of prior assumptions, we compare four types of probabilistic modeling: no prior constraint, a standard Laplace prior, a standard Gaussian prior and a learned Gaussian prior. As shown in Table 5, the learned Gaussian prior achieves the highest overall performance.

2.9. Discussion of Failure Cases

Despite PDAF's ability to mitigate domain shift through latent domain prior modeling, it still yields limited gains under severe class imbalance, similar to other DGSS methods. For instance, in the Cityscapes [5] training set, the "terrain" and "train" categories account for only 0.83% and 0.11% of annotated pixels. On the BDD dataset, their absolute mIoU gains over the backbone (DeepLabV3Plus [2]) are only 1.71 and 0.27, respectively, compared with an average per-class improvement of 8.54. This indicates that even with PDAF, the backbone struggles to improve performance on severely underrepresented classes.

2.10. Qualitative Evaluation

We provide an additional qualitative comparison with other CNN-based methods that adopt DeepLabV3Plus [2] with ResNet50 [7]. All of these methods are trained on GTAV (G) [14] and evaluated on BDD-100K (B) [18] and Mapillary (M) [13]. As illustrated in Figure 4 and Figure 5, PDAF achieves comprehensive improvements in real-world scenarios, mitigating domain shifts from varying illumination, regional discrepancies and style differences. These results highlight the effectiveness of modeling LDP, providing structured guidance to perform robust feature alignment across diverse target environments.

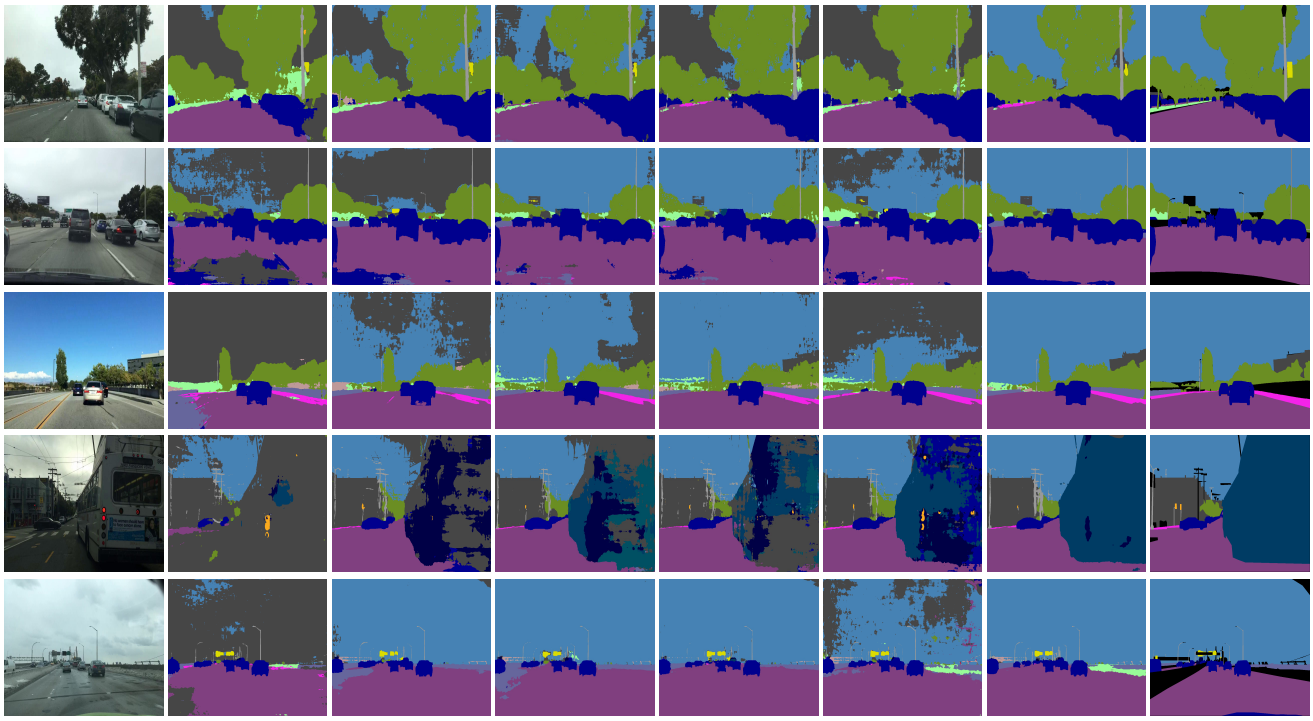


Figure 4. **Qualitative comparison with DGSS methods trained on GTAV (G) [14] and evaluated on BDD-100K (B) [18].**

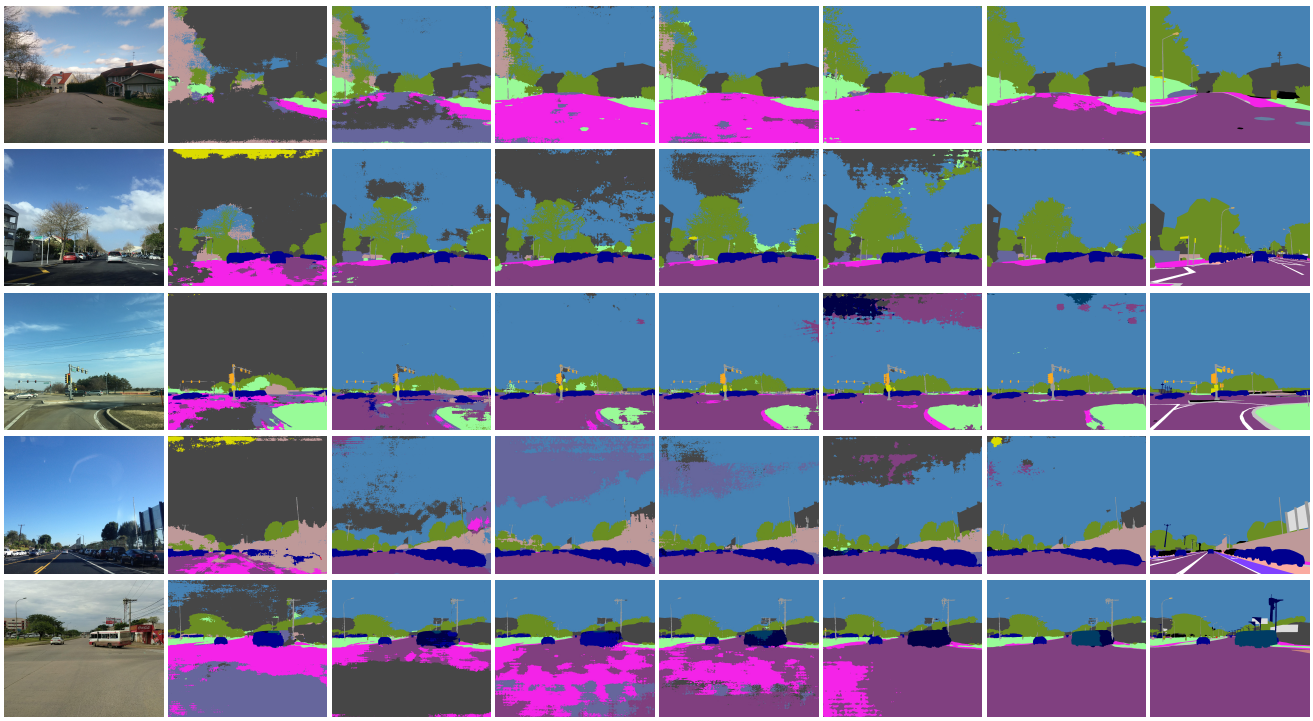


Figure 5. **Qualitative comparison with DGSS methods trained on GTAV (G) [14] and evaluated on Mapillary (M) [13].**

References

[1] Woojin Ahn, Geun Yeong Yang, Hyun Duck Choi, and Myo Taeg Lim. Style blind domain generalized semantic segmentation via covariance alignment and semantic consistence contrastive learning. In *CVPR*, 2024. 4

[2] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 3, 4

[3] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2021. 1, 2

[4] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 2021. 4

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 2, 3

[6] Jian Ding, Nan Xue, Guisong Xia, Bernt Schiele, and Dengxin Dai. Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In *CVPR*, 2023. 2

[7] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2015. 1, 2, 3

[8] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 1906. 1

[9] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *ECCV*, 2024. 2

[10] Christoph Kamann and Carsten Rother. Benchmarking the robustness of semantic segmentation models with respect to common corruptions. *IJCV*, 2019. 1, 2

[11] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *CVPR*, 2022. 4

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[13] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 2, 3, 4

[14] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 2, 3, 4

[15] Germán Ros, Laura Sellart, Joanna Materzynska, David Vázquez, and Antonio M. López. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 2

[16] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[17] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José Manuel Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 1, 2

[18] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2018. 2, 3, 4

[19] Yuyang Zhao, Zhun Zhong, Na Zhao, N. Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *ECCV*, 2022. 4