MRFP: Learning Generalizable Semantic Segmentation from Sim-2-Real with Multi-Resolution Feature Perturbation

Sumanth Udupa[† 1], Prajwal Gurunath[† 1], Aniruddh Sikdar[† 2], Suresh Sundaram[1]
[1]Department of Aerospace Engineering, Indian Institute of Science, Bengaluru, India
[2]Robert Bosch Centre for Cyber-Physical Systems, Indian Institute of Science, Bengaluru, India

{sumanthudupa, prajwalg, aniruddhss, vssuresh} @iisc.ac.in

A. Supplementary Material

This supplementary section provides additional details which could not be added in the main paper due to space constrains, including: (1) additional implementation details, (2) additional quantitative experiments, (3) qualitative results, (4) further ablation study, (5) details about computational complexity, and (6) comparison to state of the art domain generalization techniques.

A.1. Additional Implementation Details

Each of our models are trained for 40K iterations for both single-domain and multi-domain (G+S) generalization settings. The dataset partitioning (training, validation and test splits) strictly follow [1]. The re-implementations of other methods also follow the same dataset partitioning. To ensure fair-comparison, [7] is also validated on the the Synthia dataset partitioning of [1][*]. The model saved at the last iteration is used for all our analyses for out-of-domain and in-domain performance. The probabilities p of all our perturbations (HRFP, HRFP+ and NP+) are empirically set to 0.5 where each probability is independent of another. In short, the perturbations occur on independent probabilities. For the style perturbation technique, the noise variance is drawn from a Gaussian distribution of mean 1 and standard deviation on 0.75. The batch size was kept as 16 for all the experiments that used GTAV dataset as the source-domain data. In the case of Synthia dataset being the source domain data, the batch size was kept as 8. All experiments with multi-domain generalization setting was conducted with the batch size being 8. All models are trained and tested on Nvidia RTX A6000 GPU.

A.2. Quantitative Results

Table 1 presents the out-of-domain (OOD) performance of MRFP when trained on the Synthia dataset. MRFP improves on the baseline and outperforms other state-of-the-art DG methods by approximately 2% and 1% respectively showcasing the wide applicability of the technique for sim-2-real DG. Table 2 and Table 3 show the source domain performance on GTAV(G) and Synthia (S) simulated datasets respectively. Although a performance drop of 1.75% on average is observed in source domain performance compared to the baseline, it is not a major concern, as the overall generalization ability on all real-world datasets is improved significantly. Table 4 displays the wide applicability of the proposed MRFP technique as a plug-and-play module as it improves the OOD performance of a different segmentation network(in this case U-Net [9] with ResNet50 as the encoder backbone), without any additional tuning of the hyperparameters by 2.29%.

Models (S)	B	C	M	G	Avg
Baseline	20.83	27.95	24.06	29.81	25.66
IBN-Net [6] †	21.12	27.09	22.94	27.38	24.63
ISW [1] †	22.66	**29.92**	25.44	28.30	26.58
WildNet* † [5]	20.76	27.54	24.65	29.73	25.67
MRFP (Ours)	**25.68**	27.89	**25.98**	**30.50**	**27.51**

Table 1. Performance comparison of domain generalization methods using ResNet50 backbone, in terms of mIoU. Models are trained on S → {B, C, M, G }. †denotes re-implementation of the method. * indicates that the external dataset(i.e., ImageNet) used in WildNet is replaced with the source dataset for fair comparison. The best result is highlighted, and the second best result is underlined.

[†] Equal contribution of authors.

[*] RobustNet data partitioning: http://github.com/shachoi/RobustNet/blob/main/split_data/synthia_split_val.txt

Models (G)	G
Baseline	76.57
ISW [1]	72.10
MRFP (Ours)	74.28
MRFP+ (Ours)	74.85

Table 2. Source-domain performance of models trained on the GTAV dataset with Resnet50 backbone .

Models (S)	S
Baseline	78.37
ISW [1]	77.48
MRFP (Ours)	75.87
MRFP+ (Ours)	74.05

Table 3. Source-domain performance of models trained on the Synthia dataset with Resnet50 backbone .

Model(GTAV)	B	C	S	M	Avg
U-Net [9]	25.32	28.49	19.48	30.80	26.02
U-Net with MRFP	**28.87**	**28.79**	**19.60**	**36.01**	**28.31**

Table 4. Performance comparison of domain generalization methods in terms of mIoU (%) using ResNet-50 backbone. Models are trained on G → {B, C, S, M }.

A.3. Qualitative Results

Figure 1 shows the qualitative results of MRFP benchmarked against other state-of-the-art domain generalization methods. The results indicate that MRFP outperforms other methods in unseen real-world datasets, especially in adverse weather conditions as seen in Foggy Cityscapes.

A.4. Ablation study

Choice of network design: In designing the HRFP module, we chose a four layer encoder-decoder structure guided by Grad-CAM visualizations. Empirically, incorporating 4+ layers mostly captures redundant fine features leading to performance saturation with only a marginal +0.3% OOD performance gain, albeit with increased GPU memory usage during training. Conversely, 2 or 3 layers result in -1.5% OOD performance. In the HRFP module, the number of channels in each layer remains consistent, except for the layers added to the base network, ensuring compatibility with the summation operation.

Table 5 shows the relative performance of three different up-scaling factors for the bilinear interpolation of the (High Resolution Feature Perturbation) HRFP block. The proposed method HRFP is empirically set to have a overall scale increase of 2 in its latent space. In other words, the spatial resolution, over the course of the first four encoder layers, is ultimately increased to double the spatial resolution of the input to the HRFP block. This can be termed as overall scale-factor increase. To evaluate the memory consumption to accuracy trade-offs, we present additional results on overall scale factors of 1.5 and 2.5 respectively in Table 5. It is observed that while a higher overall scale-factor of 2.5 provides a slight improvement in the OOD performance, the memory requirement during training is substantially higher. An up-scaling factor of 1.5 provides a slight drop in the OOD performance while the memory requirements tend to be on the lower end. Overall, the sensitivity to the up-scaling factor seems to be minimal, but in the proposed method we adopt an overall scale factor of 2 as it helps us decrease the memory usage compared to 2.5 and add the HRFP+ decoder perturbation which yields significant OOD performance improvements.

Table 6 shows the benefit of using the proposed MRFP/MRFP+ for domain generalization. In Table 6, L-MRFP is learnable MRFP which refers to the HRFP block of MRFP being learnable and not randomized every iteration. RGN refers to random Gaussian noise being added to features as feature perturbations instead of the proposed MRFP/MRFP+ module to the baseline model (DeepLabv3+). It is seen that feature perturbation is helpful for robustness to domain shifts as L-MRFP performs inferior to the proposed RGN and the proposed MRFP/MRFP+. MRFP/MRFP+ out-of-domain(OOD) performance surpasses L-MRFP and RGN by approximately 4%. This suggests that adding random noise to the features as perturbations does increase the OOD performance but has the issue of redundancy, and distorting the semantics of the domain-invariant features which can negatively impact OOD performance thus leaving room for improvement. To fill this gap, the proposed MRFP/MRFP+ technique predominantly perturbs domain-specific features while minimizing the semantic distortion that can occur to domain-variant features by random perturbation.

The proposed HRFP block and the NP+ style perturbation technique is introduced in the initial layers as they predominantly contain style information and fine-grained domain-specific features. To restrict these domain-specific features from propagating and influencing the later layers of the network, we add the HRFP/HRFP+ block with style perturbation techniques at the initial layers. This setting was empirically found to be the best.

A.5. Computational Complexity

Table 7 shows the computational complexity comparison of the proposed MRFP technique with other domain generalization methods. Since MRFP is only used for training, and only the baseline segmentation model DeepLabv3+ is used for inference, the number of parameters and the computational cost (GMACs) remains the same as the baseline model. Although the parameters of SAN-SAW [7] is considerably less compared to MRFP and the baseline model

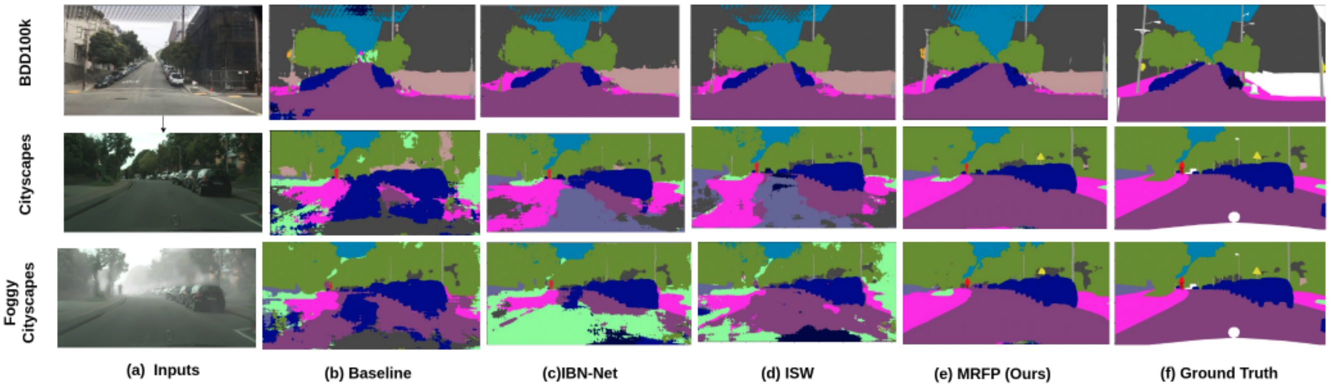

Figure 1. Qualitative comparison with different domain generalization methods, trained on GTAV [8] with the backbone as ResNet50, tested on BDD-100K [11], Cityscapes [3] and Foggy Cityscapes [10]. MRFP consistently generates superior predictions compared to other methods, especially on Cityscapes and Foggy Cityscapes.

Models(GTAV)	B	C	M	Avg
MRFP (OSF=1.5)	38.29	40.27	41.95	40.17
MRFP (OSF=2.5)	39.21	39.84	43.20	40.75
MRFP	38.80	40.25	41.96	40.33
MRFP+	**39.55**	**42.40**	**44.93**	**42.29**

Table 5. Performance comparison of domain generalization methods in terms of mIoU using ResNet50 backbone. Models are trained on G → {B, C, M}. OSF= Overall scale-factor increase.

Model(GTAV)	B	C	S	M	Avg
Baseline	31.44	34.66	25.84	32.93	31.21
L-MRFP	33.38	38.21	25.40	38.04	33.75
RGN	34.70	36.84	25.36	41.81	34.67
MRFP(Ours)	38.80	40.25	27.37	41.96	37.09
MRFP+(Ours)	**39.55**	**42.20**	**30.22**	**44.93**	**39.27**

Table 6. Performance comparison of domain generalization methods in terms of mIoU using ResNet50 backbone. Models are trained on G → {B, C, S, M }. RGN=Random Gaussian Noise, L-MRFP=Learnable MRFP.

Models	# of Params	GMACs
Baseline	40.35M	554.31
IBN-Net [6]	45.08M	555.64
ISW [1]	45.08M	555.56
SAN-SAW [7]	25.63M	843.72
WildNet [5]	45.21M	554.32
MRFP	40.35M	554.31

Table 7. Inference computation cost comparisons of MRFP and other contemporary methods.

DeepLabv3+ model, its computational cost and training time is extremely high.

A.6. Comparison With Alternative Domain Generalization Techniques

The domain generalization performance of the proposed MRFP/MRFP+ technique is compared with numerous state-of-the-art DG methods. The augmentations used for training and re-implementation are restricted to standard augmentations, and all settings are consistent with ISW [1]. We do not perform a performance comparison with Pro-RandConv [2] on the Foggy Cityscapes dataset [10] because its open-source code is not available. We also do not conduct a performance comparison between MRFP and [4], because of mainly two reasons: (1) in addition to using data augmentations used in [1], extra strong style augmentations are used to enhance the style information of urban-scene images, using Automold road augmentation library[†], and (2) open source code is not available for re-implementation without the extra strong style augmentations used. In the future, we intend to use these augmentations for style variations on source domains.

References

[1] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021. 1, 2, 3

[2] Seokeon Choi, Debasmit Das, Sungha Choi, Seunghan Yang, Hyunsin Park, and Sungrack Yun. Progressive random con-

[†] Automold Road augmentation library : https://github.com/UjjwalSaxena/Automold--Road-Augmentation-Library

volutions for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10312–10322, 2023. 3

[3] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 3

[4] Wei Huang, Chang Chen, Yong Li, Jiacheng Li, Cheng Li, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Style projected clustering for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3061–3071, 2023. 3

[5] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9936–9946, 2022. 1, 3

[6] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 464–479, 2018. 1, 3

[7] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-Aware Domain Generalized Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2595, New Orleans, LA, USA, 2022. IEEE. 1, 2, 3

[8] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 3

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1, 2

[10] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 3

[11] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 3