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

Abstract

Deep neural networks have shown exemplary performance on semantic scene understanding tasks on source domains, but due to the absence of style diversity during training, enhancing performance on unseen target domains using only single source domain data remains a challenging task. Generation of simulated data is a feasible alternative to retrieving large style-diverse real-world datasets as it is a cumbersome and budget-intensive process. However, the large domain-specific inconsistencies between simulated and real-world data pose a significant generalization challenge in semantic segmentation. In this work, to alleviate this problem, we propose a novel Multi-Resolution Feature Perturbation (MRFP) technique to randomize domain-specific fine-grained features and perturb style of coarse features. Our experimental results on various urban-scene segmentation datasets clearly indicate that, along with the perturbation of style-information, perturbation of fine-feature components is paramount to learn domain invariant robust feature maps for semantic segmentation models. MRFP is a simple and computationally efficient, transferable module with no additional learnable parameters or objective functions, that helps state-of-the-art deep neural networks to learn robust domain invariant features for simulation-to-real semantic segmentation. Code is available at https://github.com/airl-iisc/MRFP.

1. Introduction

Semantic segmentation is a fundamental computer vision task, with diverse downstream applications, such as autonomous driving [52], robot navigation [30, 38], medical image analysis [1], land cover classification [49], and building detection [4, 37]. Producing synthetic data for training deep neural networks (DNNs) is a cost-effective and straightforward alternative compared to the myriad of challenges associated with collecting real-world data. However, models trained on synthetic data leads to the domain shift problem [2, 6, 27, 34, 42], resulting in a drastic drop in performance. In safety-critical applications such as autonomous driving, different illuminations, adverse weather conditions and the domain shift from synthetic data [32, 33] cause a significant drop in the performance of DNNs when tested on real-world datasets. There are two primary challenges with training models on synthetic datasets: 1) It is not feasible to synthetically generate all potential unseen domains. 2) Models trained on synthetic data do not generalize to real-world scenarios due to the domain gap that persists when deployed in real-world situations.

In this paper, we propose a novel feature perturbation technique referred to as Multi-Resolution Feature Perturbation (MRFP) to address the domain gap between the source and target domains, especially in a Sim-2-Real Single Domain Generalization (SDG) setting. We hypothesise that there exists a latent space consisting of domain-agnostic features that are: 1) independent of style information such as illumination and color characterized as low frequency (LF) components; and 2) fine-grained local texture information characterized as high-frequency (HF) components. The objective of MRFP is to selectively perturb domain-variant LF and HF encoder features, aiming to improve the generalizability of DNNs. MRFP achieves this through two divergent receptive field branches tasked to focus on HF fine-grained features and LF semantic information. While previous works have focused their efforts on SDG [12, 18, 44, 46, 47, 56] and some have tackled the Sim-2-Real [7, 8, 45] setting, there is a dearth of literature on the Sim-2-Real problem with an image frequency perspective. To address the domain gap, existing approaches involve enhancing the domain variance of the available source domain data and enriching its representation. This is achieved through methods such as introducing adversarial perturba-

[†]Equal contribution of authors.

tion [5, 48] or employing style manipulation techniques [11, 48, 56]. However, most of these techniques involve intricate training procedures and multiple objective functions.

DNNs can focus on a broad spectrum of frequencies, ranging from low to high. Wang *et al*. [45] suggests that convolution-based DNNs tend to grasp LF attributes in the initial training phases, and gradually transition their attention towards HF components that are notably more domain-specific. This phenomenon is illustrated with *vanilla training* in Fig.1. Convolution-based DNNs progressively cover the entire frequency spectrum as marked by '[ ]' in Fig.1 denoting the model focus range. This wide range also includes fine-grained domain-specific information. Huang *et al*. [17] shows that the lowest and the highest spectral bands in the frequency domain capture domain variant features which hinders generalization performance on unseen domains. From our observations of the feature space, it becomes apparent that high-resolution (in a spatial sense) features mostly correlate to fine-grained features (as studied in Section 4). Similarly, we observe that low-resolution (in a spatial sense) features correspond to coarse features. Drawing connections from a spatial to a Fourier perspective, we hypothesize that fine-grained information predominantly corresponds to domain-specific HF features. While coarse features correspond to LF features.

MRFP not only contributes to style perturbation but also provides control over perturbation of fine-grained features. It primarily consists of two components, i.e, the High-Resolution Feature Perturbation (HRFP) module, which comprises of a randomly initialized overcomplete (in a spatial sense) autoencoder, and style perturbation with the normalized perturbation technique (NP+) [11] within the feature space. Style perturbation techniques at the image level, however, is limited, deterministic and sacrifices source domain performance because of its potential to adversely affect image content [11]. Although NP+ (a feature level perturbation technique) can generate diverse styles while preserving high content fidelity, it does not perturb domain-specific fine-grained features, thus missing opportunities to further enhance generalizability. Randomizing these features during the training process restricts the model from drawing inherent source domain specific patterns.

RandConv [51] and ProRandConv [8] are image augmentation techniques that introduce variance in features in the form of contrast and texture diversification in the image space. In contrast to ProRandConv, the proposed HRFP module operates in the feature space by extracting fine-grained characteristics via a decreasing receptive field from a randomly initialized overcomplete autoencoder, serving as perturbations to prevent domain-variant feature overfitting. The inherent benefit of utilizing overcomplete convolutions lies in their decreasing receptive field, where perturbations do not induce significant semantic distortions. This

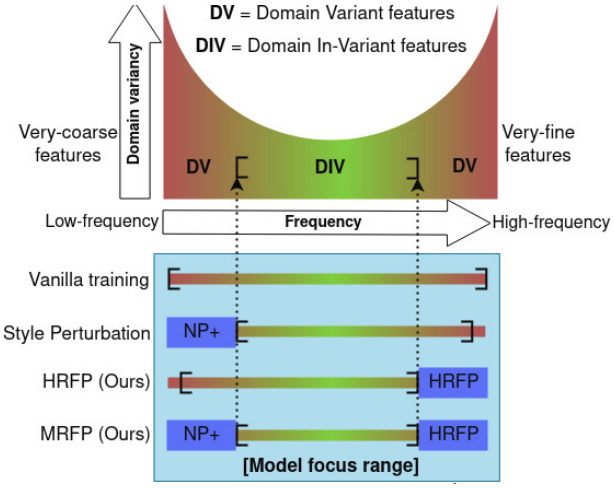

Figure 1. Deep models focus on low-frequency features in the initial stages of vanilla training and shift their focus mainly to domain-variant HF (very-fine) features, covering the entire spectrum. Introducing variability with Style Perturbation (NP+) and High-Resolution Feature Perturbation (HRFP) at both ends of the spectrum, shifts the model's focus to domain in-variant features.

is because the influence of pixels beyond the center of the receptive field is minimal compared to the increased impact observed in undercomplete networks with an increasing receptive field. As shown in Fig.1, the introduction of *style perturbation* with NP+ and *HRFP* helps restrict the model focus range to domain in-variant features.

To the best of our knowledge, MRFP is the first technique that employs decreasing receptive fields of overcomplete networks to focus on fine-grained features and perturb them through randomly initialized weights to enhance generalization performance. To summarize, the overall main contributions are as follows:

- A novel MRFP technique is proposed to introduce perturbations to fine-grained information and infuse varied style information into any baseline segmentation encoder backbone to facilitate the learning of domain-agnostic semantic features.
- The proposed HRFP technique aims to prevent overfitting on source domain, by using a randomly initialized overcomplete autoencoder on the shallow encoder layers and decoder layer of the baseline segmentation model as a feature-space perturbation.
- MRFP is a simple transferable module with no additional learnable parameters or objective functions, which improves the generalizability of deep semantic segmentation models.
- Extensive experiments over seven urban semantic segmentation datasets show that the proposed model achieves superior performance for single and multi-domain generalization tasks in a Sim-2-Real setting.

2. Related Works

Domain Generalization: These techniques aim to improve the generalization ability of models to unseen target domains, without any access to these domains during training. Various methods like domain alignment [13], meta-learning [25], adversarial learning [48], and data augmentation [5, 16] have been proposed to learn domain invariant features. IBN-Net [29] shows significant improvement combining batch and instance norm to learn discriminative features and avoid overfitting on the training data. NP+ [11] has been used to perturb the feature statistics and synthesize diverse domain styles. Whitening transform has shown to eliminate style information when applied to each instance [26], but may remove domain invariant content at the same time. An instance selective whitening loss was proposed by Robustnet [7] to selectively remove only feature representations that cause domain shifts. WildNet [24], SAN-SAW [31] and Style Projected Clustering [19] are recent DG methods that either use external real-world data (ImageNet) for synthesising styles or have complicated training strategies with multiple objective functions. MRFP on the other hand, does not introduce any learnable parameters during the training phase or use any external data.

Data Augmentation and Domain Randomization: Data augmentation and domain randomization [39] techniques are used to expand the training data for better generalization. For classification task, frequency based techniques like APR [3] have been proposed, where the main idea is to augment only the amplitude spectrum of an image while keeping the phase spectrum constant. Style randomization is used to expand the coverage and diversify the source domain using normalization layers [20] and random convolutions [51]. Progressive random convolutions [8] employ progressively stacked randomly initialized convolutions with an increasing receptive field, as an image-space style perturbation to introduce diversity in style and contrast. In contrast, the suggested MRFP technique perturbs both low and high frequencies in the feature space, incorporating both style perturbation and HF perturbation.

Overcomplete Networks: Previous studies [21, 37, 41] have suggested that overcomplete representations have the ability to pick up on the finer details in the input image while also being robust to noise compared to their undercomplete counterparts in the semantic segmentation task. It is shown that these fine features are paramount for high source domain accuracy [43]. However, the generalization ability of overcomplete models has not been explored. When subjected to diverse domains, learning these fine features can have a detrimental impact on out-of-domain performance. MRFP however, employs these fine features as perturbations to prohibit the model from overfitting on HF domain-specific features.

3. Methodology

3.1. Problem Formulation

Domain generalization aims to learn a domain-independent model, to train on only a source domain S and test on unseen target domains $T = \{T_1, T_2...T_n\}$. Let the source domain training dataset be denoted as $S = \{x_n, y_n\}_{n=1}^{N_S}$ where x_n is the n^{th} image, y_n is its corresponding pixel-wise label, and N_s is the total samples in the source domain S. The focus of this work is the semantic segmentation task, with the assumption of a common label space for source and unseen target domains. For multi-domain generalization case, where multiple source domains exist, $S = \{S_1, S_2...S_n\}$ are used during training, and for each training iteration, samples are selected randomly from multiple source domains as input. The objective function to train the model using empirical risk minimization [40] is given as:

$$\arg\min_{\theta} \frac{1}{N_S} \sum_{n=1}^{N_S} l(f_\theta(x_n), y_n) \tag{1}$$

where $f_\theta(\cdot)$ is the semantic segmentation network that outputs pixel-wise category predictions, θ represents the learnable parameters in the network, and $l(\cdot)$ is the cross-entropy loss function to measure error. To make the segmentation model $f_\theta(\cdot)$ generalizable by learning domain invariant features in both single and multi-domain settings, MRFP technique is proposed.

3.2. Preliminary: Overcomplete Representations

The proposed HRFP module is a randomly initialized overcomplete auto-encoder. In this sub-section a brief overview of overcomplete representations is presented. Let *F1* and *F2* be the feature maps of input image *I*. Assume that the initial Receptive Field (RF) of the convolutional filter is $k \times k$ on the image. In undercomplete auto-encoders, due to the max pooling operation, the spatial dimensionality of *F1* is halved causing an increase in the receptive field of *F1, F2* and so on. Eq. 2 is the generalized RF equation for the i^{th} layer.

$$R.F.\ w.r.t\ I = 2^{2(i-1)} \times k \times k \tag{2}$$

However, with overcomplete architectures spatial dimensionality of features F1, F2 and so on increase. For example, an upsampling operation of coefficient 2 that replaces the max pooling operation causes a decrease in the receptive field as generalized in Eq. 3.

$$R.F.\ w.r.t\ I = (1/2)^{2(i-1)} \times k \times k \tag{3}$$

Due to their decreasing receptive fields overcomplete architectures focus on meaningful fine-grained information [21].

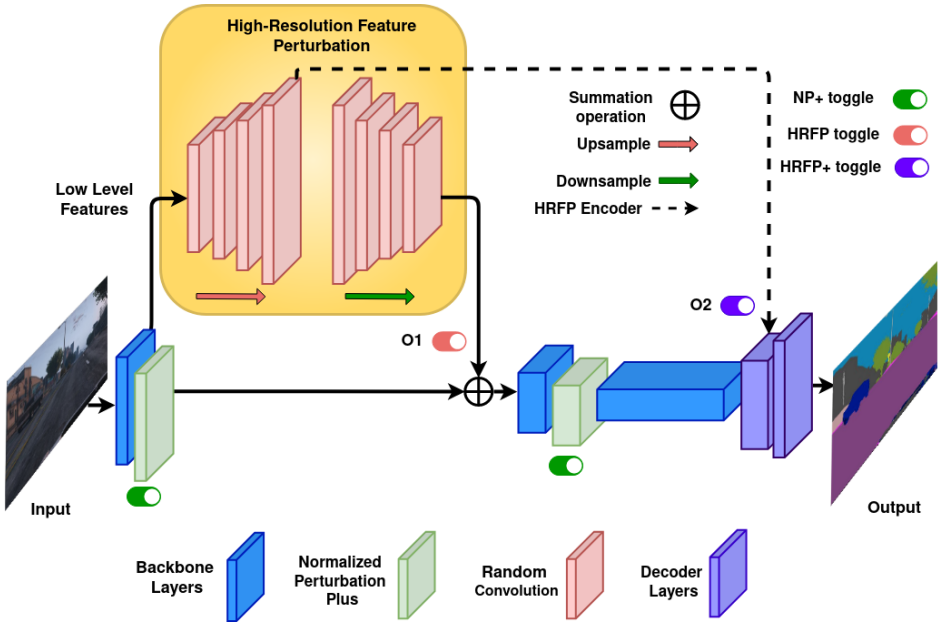

Figure 2. Multi-Resolution Feature Perturbation Technique: Normalized Perturbation (NP+) and High-Resolution Feature Perturbation (HRFP) are randomly incorporated into the training procedure for the baseline segmentation model (DeepLabv3+) , which are represented by the toggles. Dotted line, which is the addition of features to penultimate layer of decoder, is incorporated only in High-Resolution Feature Perturbation Plus (HRFP+) technique. MRFP $\rightarrow$ {HRFP, NP+} and MRFP+ $\rightarrow$ {HRFP, HRFP+, NP+}

3.3. Multi-Resolution Feature Perturbation

An inherent tactic for addressing domain shift in domain generalization approaches involve producing diverse data and integrating it into the training set. In the proposed MRFP technique, in addition to creating diverse styles it aims to perturb the distribution of domain-specific fine-grained features so that the model does not tend to overfit on these fragile features. MRFP technique has two main components namely the high-resolution feature perturbation module and the NP+ module, which are detailed below.

High-Resolution Feature Perturbation (HRFP): DNNs overfitting on fine-grained features are shown to be detrimental to performance when tested on unseen domains [45]. To address this issue, we employ a randomly initialized overcomplete convolutional auto-encoder that transforms input features to a higher dimension (in a spatial sense). With a decreasing receptive field in HRFP, there exists a high focus on fine-grained features. These fine-grained features are perturbed using random convolutional and batch norm layers as shown in Fig. 2. These perturbations are subsequently added to the base network to prevent the model from identifying inherent domain-specific patterns.

HRFP is a plug and play module, and can work with any deep segmentation encoder backbone. The encoder and decoder of HRFP consists of 4 convolutional layers each, where every randomly initialized convolutional layer is followed by a randomly initialized batch normalization layer. The reduction in receptive field as denoted in Eq. 3 occurs as a result of increasing the spatial resolution of the feature maps in each layer consecutively by bilinear interpolation with a scaling factor of approximately 1.2 in the HRFP encoder. The HRFP module is upsampled up to a maximum spatial resolution that is twice the size of its own input. Following this, the final four layers of the HRFP module reduce the spatial dimensionality of the overcomplete latent space back to its original input size, facilitating its smooth integration with the base network. Further details are provided in the supplementary material. The random convolution weights are He-initialized [14] whereas, random batch normalization weights (i.e γ and β) are sampled from a Gaussian distribution $\mathcal{N}(0, \sigma^2)$. The input to the HRFP module originates from the output of the initial stage (stage 0) of the encoder from the backbone layers of the baseline segmentation network DeepLabv3+, as shown in Fig. 2. Since the shallow layers in CNNs preserve style related information through encoding local structures[55], we focus on feature perturbation in these layers, which empirically yielded the best performance. Hence, the output of HRFP is incorporated as a perturbation, added to the output of the same layer within the base network's encoder backbone as shown in O_1 branch in Fig. 2.

HRFP+: In addition, to further encourage the model to focus towards learning robust domain-invariant features, fine-grained perturbation is added to the decoder of the base

network. In HRFP+, the output of the largest upsampled encoder layer of the HRFP block is added to the penultimate layer of the decoder of the baseline segmentation model as shown by branch O_2 in Fig. 2. The intuition behind this extra perturbation in the decoder of the base network is that it adversarially helps the segmentation head and makes it more robust against domain-specific HF noise. Additionally in this case, three instance normalization layers [29] have been adopted in a similar fashion as [7].

Style Perturbation: From our conjecture that low-resolution features correspond to LF features, we aim to increase the variations in the low-frequencies by perturbing feature channel statistics in the spatial domain. To facilitate an increase in the diversity of style which is known to correspond to LF components in the amplitude spectrum, feature channel statistics in the spatial domain are modified using normalized perturbation [11]. NP+ helps the model perceive potentially diverse domains and not overfit to the source domain, and is given by,

$$y = \sigma_s^* \frac{x - \mu_c}{\sigma_c} + \mu_s^*, \quad \sigma_s^* = \alpha\sigma_c, \quad \mu_s^* = \beta\mu_c \tag{4}$$

where $\{\mu_c, \sigma_c\} \in \mathbb{R}^{B\times C}$ are mean and variance of input channels, and $\{\alpha, \beta\} \in \mathbb{R}^{B\times C}$ are drawn from normal distribution. For a batch B with feature channel statistics μ_c, the statistic variance $\Delta \in \mathbb{R}^{1\times C}$ is given by,

$$\Delta = \frac{1}{B}\sum_{b=1}^{B}(\mu_c^b - \tilde{\mu}_c)^2, \quad \tilde{\mu}_c = \frac{1}{B}\sum_{b=1}^{B}(\mu_c^b) \tag{5}$$

where, μ_c^b is the feature channel mean of b^{th} sample in the batch. Setting the normalized variance $\delta = \Delta/\max(\Delta)$ and using Eq. 4, the output feature map y is given as,

$$y = \alpha x + \delta(\beta - \alpha)\mu_c \tag{6}$$

where max represents the maximum operation. Since the features being perturbed using NP+ have a smaller spatial resolution than that of the HRFP block, these perturbations can be viewed as low-resolution feature perturbations.

MRFP/MRFP+ consists of both high-resolution feature perturbation (HRFP/HRFP+) and normalized feature perturbation. The perturbations caused due to HRFP/HRFP+ enables the model to learn domain invariant representations in the learned feature space from distinct domains generated from NP+. The proposed method is fundamentally different from previous convolutional randomization techniques [8, 51, 54] wherein, the task is to induce various style domains in the image space. In contrast, MRFP/MRFP+ aims to induce domain agnostic model behaviour by not only generating diverse style-information in the feature space but also by not letting the model overfit on HF source domain-specific features. The proposed module is a simple, computationally efficient, transferable technique, and thus can be attached to any deep backbone network while adding no additional learnable parameters, nor extra objective functions to optimize in the training process of the base network. During inference, the MRFP module is removed, and only the baseline segmentation network is used.

4. Experiments

4.1. Experimental Setup

The assessment of the proposed MRFP technique involves the use of two synthetic datasets, GTAV [32] individually and collectively with Synthia [33] to address the challenge of Sim-2-Real domain generalization. The models that undergo training are subsequently tested on unseen domains that were not part of the training process. These new domains encompass BDD100k (B), Cityscapes (C), Mapillary (M), Foggy Cityscapes (F), and either GTAV (G) or Synthia (S), depending on the specific training configuration. In scenarios involving single and multi-domain generalization, the setups are as follows: G→ {B, C, M, S}, G→ {F, Rainy Cityscapes with intensity level of 25mm, 50mm, 75mm, 100mm} and (G + S)→ {B, C, M}. In order to ensure fair comparisons, we re-implement IBN-Net [29], RobustNet (ISW) [7], and SAN-SAW [31], and evaluate them on F using their open source codes*. As explained in Section 3.2, MRFP is a transferable plug-and-play module that can be incorporated into any existing model backbone. Extensive experimentation is carried out using two distinct backbones, namely Resnet-50 and MobileNetv2, showcasing the effectiveness and wide applicability of the proposed module. We use Mean Intersection over Union (mIoU) as our quantitative metric.

4.2. Datasets Description

Synthetic Datasets: GTAV [32] is a synthetic image dataset generated using the GTA-V game engine, comprising of 24966 images with pixel-wise semantic labels, with a resolution of 1914x1052. Similarly, Synthia [33] is also a synthetically generated dataset which includes 9400 images with a resolution of 1280x760. Meanwhile, Synthia has 6580 training images and 2820 validation images. Both datasets have 19 common object categories.

Real-world Datasets: Five real-world datasets are used, namely Cityscapes (C), Foggy Cityscapes (F), BDD-100k (B), Mapillary (M) and Rainy Cityscapes (R), maintaining a common label space of 19 classes. These datasets are exclusively employed for testing purposes, using their respective validation sets. Cityscapes [9] is a large-scale dataset, with the resolution as 2048x1024, which contains 500 validation samples. In F [35], synthetic fog is added to the Cityscapes

*Details about re-implemented methods are given in Supplementary material.

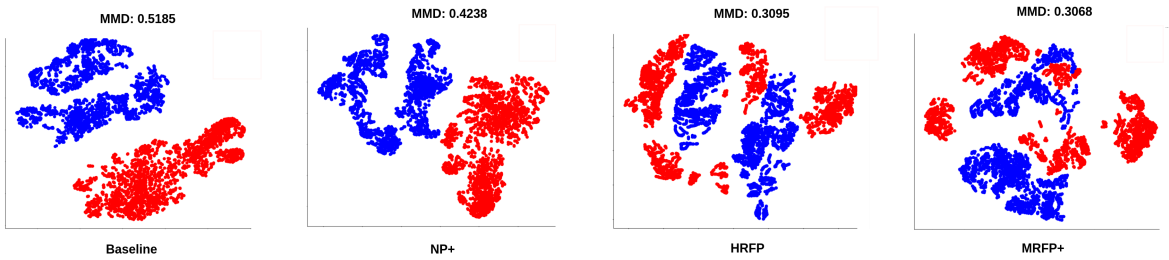

Figure 3. t-SNE visualization for the feature channel statistics of different components of MRFP+, MRFP, NP+ and baseline on GTAV (source domain - red color) and Mapillary (target domain - blue color). The corresponding MMD scores are also reported.

Models(GTAV)	B	C	M	S	Avg
Baseline	31.44	34.66	32.93	25.84	31.21
IBN-Net [29]	32.30	33.85	37.75	27.90	32.95
ISW [7]	35.20	36.58	40.33	28.30	35.10
SAN-SAW [31]	37.34	39.75	41.86	26.70	36.41
WildNet* [24]	34.65	39.13	39.05	28.41	35.31
WEDGE [22]	37.00	38.36	44.82	N/A	N/A
DIRL [50]	39.15	41.04	41.60	N/A	N/A
ProRandConv [8]	37.03	42.36	41.63	25.52	36.63
MRFP(Ours)	38.80	40.25	41.96	27.37	37.09
MRFP+(Ours)	**39.55**	**42.40**	**44.93**	**30.22**	**39.27**

Table 1. Performance comparison of domain generalization methods in terms of mIoU using ResNet-50 backbone. Models are trained on G, and validated on B, C, M, and S. * indicates that the external dataset (i.e., ImageNet) used in WildNet is replaced with the source dataset for fair comparison.

Models(GTAV)	F	25mm	50mm	75mm	100mm	Avg
IBN-Net [29]	33.18	20.07	14.85	11.16	8.80	17.61
ISW [7]	36.30	23.7	17.93	13.53	10.39	20.37
WildNet*† [24]	38.75	27.04	19.17	13.94	9.45	21.67
MRFP(Ours)	37.90	28.29	21.86	17.05	12.79	23.57
MRFP+(Ours)	**40.67**	**30.42**	**23.74**	**19.25**	**15.28**	**25.87**

Table 2. Performance comparison of domain generalization methods using ResNet-50 backbone, in terms of mIoU. Models are trained on GTAV and tested on adverse weather conditions like fog (foggy Cityscapes) and different levels of rain intensity in mm (Rainy Cityscapes). †denotes re-implementation of the method. The best result is highlighted, and the second best result is underlined.

images to simulate reduced visibility conditions, and contains the 1500 images in validation set based on different foggy settings. BDD-100k [53] and Mapillary [28] both contain diverse street view images of resolution 1280x720 and 1920x1080 respectively. The images in validation set of B are 1000 and 2000 for M.

4.3. Implementation Details

The baseline segmentation model employed is DeepLabv3+ [4], implemented with Resnet50 [15] and MobileNetv2 [36] backbones, both utilizing a dilation rate of 16. All backbones used for training are pretrained on ImageNet [10]. All experiments use SGD [23] optimizer with a momentum of 0.9 and 1e-4 as the weight decay. Initial learning rate is set to 0.01 and is decreased according to polynomial rate scheduler with a power of 0.9. All Resnet50 and MobileNetv2 backbone models are trained for 40k iterations with a batch size of 16 for both single and multi-domain generalization settings. Our augmentation, dataset splits and validation settings are consistent with [7]. For the randomly initialized HRFP module, the batch-norm layer parameters are sampled from a Gaussian distribution with a standard deviation of 0.5. During the training process, NP+, HRFP and HRFP+ modules are subjected to independent randomization, each with a probability of 0.5. Inference of the trained models are conducted on the baseline segmentation network DeepLabv3+ only. To ensure equitable evaluations, all the results reported in this subsection are averaged over three separate runs for fair comparisons. Further implementation details are provided in the supplementary material.

Models (GTAV)	B	C	S	M	F	Avg
Baseline	26.76	26.95	22.72	27.34	27.26	26.20
IBN-Net [29]	27.66	30.14	24.98	27.07	30.03	27.98
ISW [7]	30.05	30.86	24.43	30.67	30.70	29.34
MRFP (Ours)	**33.03**	**32.92**	**25.62**	**30.95**	**32.97**	**31.10**

Table 3. Comparison of mIoU (%). All models are trained on GTAV with DeepLabv3+ with MobileNetv2 as backbone.

4.4. Experimental Results

4.4.1 Quantitative Evaluation

The domain generalization performance of the proposed MRFP technique is compared with existing methods: IBN-Net [29], ISW [7], SAN-SAW [31], ProRandConv [8] and WildNet [24]. Table 1 and Table 2 show the generalization performance of state-of-the-art (SOTA) models and MRFP with ResNet-50 backbone for single domain generalization, in a Sim-2-Real setting. As shown in Table 1, both the MRFP and MRFP+ achieve superior performance compared to the SOTA methods on B, C, M and S. It has an improvement of 7.56% on an average of all target domain datasets compared to baseline DeepLabv3+ model, and an improvement of 2.64% compared to ProRandConv. Table 2 shows the generalization performance when trained on G and tested on adverse weather condition datasets like F and R. Table 3 displays the models trained on GTAV, using MobileNetv2 [36] as the backbone network. MRFP outperforms all the other methods, showing the plug-and-play nature of the proposed method.

Models (G+S)	B	C	M	Avg
Baseline	30.11	38.63	35.90	34.88
ISW [7]	35.99	37.24	38.97	37.40
MRFP (Ours)	40.35	44.54	**45.78**	42.55
MRFP+ (Ours)	**41.13**	**46.18**	45.28	**44.24**

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

Table 4 shows the Sim-2-Real, multi-domain generalization performance of various domain generalization methods trained on G and S (G+S) datasets combined. The proposed model outperforms the baseline model by 9.36 % and ISW by 6.84 %. MRFP+ demonstrates enhanced generalization performance in the Sim-2-Real scenario by enhancing the base network's ability to capture robust and domain-invariant features. In contrast, other approaches primarily focus on normalization, whitening techniques, or introducing random styles. These randomization techniques introduce perturbations with the intention of manipulating input image styles, aiming to cover portions of the target domain. Our method not only covers this aspect, but also induces robustness by perturbing highly domain-specific features in the shallow layers of the encoder. Importantly, this approach doesn't introduce any increase in computational complexity, as only the baseline DeepLabv3+ segmentation model is used during inference.

Models (GTAV)	B	C	M	S	Avg
Baseline	31.44	34.66	32.93	25.83	31.71
HRFP+	39.28	41.39	42.70	29.70	38.26
HRFP	34.18	38.15	43.33	26.71	35.59
NP+	34.50	40.33	38.85	28.65	35.58
SCFP	38.57	40.65	42.48	28.24	37.48
MRFP(Ours)	38.80	40.25	41.96	27.37	37.09
MRFP+(Ours)	**39.55**	**42.40**	**44.93**	**30.22**	**39.27**

Table 5. Ablation analysis of each setting of the MRFP block, mIoU (%) is reported using ResNet-50 as backbone in the scenario: G $\rightarrow$ {B, C, M, S}

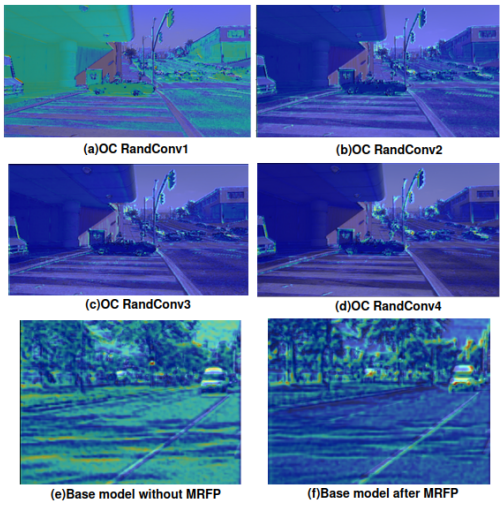

Figure 4. Grad-CAM outputs of subsequent HRFP layers (a-d), shows that constriction of receptive field, forces the module to focus on fine-grained information. (e) showcases model focus on domain-specific features whereas (f) with MRFP the base model focuses on domain in-variant meaningful features.

4.4.2 Qualitative Evaluation

To analyze the MRFP module, Grad-CAM visualizations are shown in Fig. 4. Due to the HRFP block lacking learnable weights, we adopt an ultra-low learning rate in model training for one iteration to create Grad-CAM visualizations for validating our hypothesis. The focus of the HRFP module moves from coarse to fine features from (a) to (d). It is also seen in (f) that with MRFP in the training procedure, the base model tends to focus more on domain in-variant features such as the vehicle and trees (in this case), as opposed to very-fine textures on the tarmac.

Fig. 5 shows the predictions of contemporary generalization models trained on GTAV, and tested on Mapillary. The model extends the source class content to the target do-

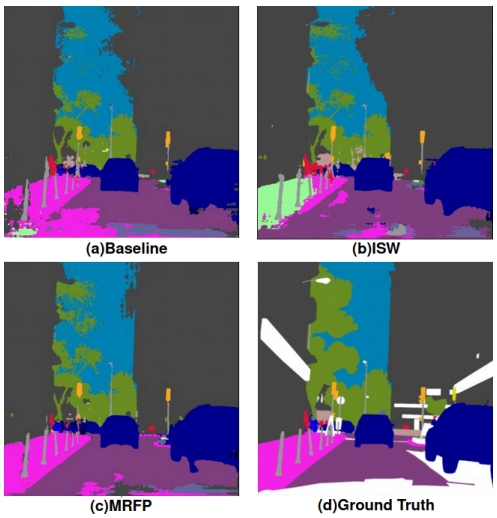

Figure 5. Segmentation outputs of contemporary generalization methods with ground truth.

mains, e.g, it accurately predicts the pavement, vehicles and roads compared to the baseline and ISW. Further experimental results and segmentation outputs are reported in the supplementary material.

4.4.3 Ablation Study

NP+ and HRFP/HRFP+ differ in how they enhance generalization. NP+ targets LF spectrum while HRFP targets the HF spectrum as seen in Fig. 6, which denotes a Fourier analysis of NP+ and HRFP perturbations. A relative increase in the presence of frequencies is studied across 3 bands. The presence of low-frequencies are predominantly increased after NP+. In stark contrast, HRFP predominantly increases the presence of HF components in the feature space. This supports our previously stated conjecture and approach employed to improve out of distribution performance. Fig. 3. depicts the t-SNE plots for the final encoder stage of the ResNet50 backbone along with the corresponding Maximum Mean Discrepancy (MMD) scores. A lower MMD score and a higher degree of overlap between the two distributions indicate better generalizability. This is observed in the components used in the proposed module.

Impact of different components in MRFP/MRFP+: To investigate the contribution of each component in the overall MRFP technique, we perform an ablation analysis where we validate the need for different components of MRFP as well as different spatial configurations in the proposed HRFP module. Table 5 shows HRFP+ alone outperforms NP+ by 2.68%, beating SOTA methods by 1.63%, and the baseline by 6.55%. Additionally, NP+ only supplements HRFP/HRFP+, aligning with our hypothesis. All experiments are conducted using the same training settings

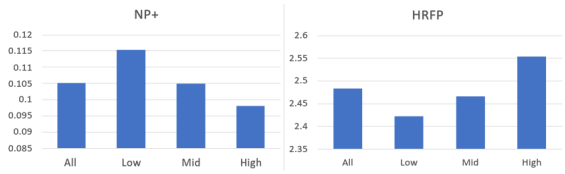

Figure 6. Presence of frequencies in the Fourier domain of NP+ and HRFP features. The most significant rise in the presence of LF features is noted following NP+ perturbation (left), whereas the most notable increase in the presence of high-frequency features occurs with HRFP perturbation (right).

as described in the implementation details. As seen from the results in Table 5, we observe that disabling either component has a detrimental effect on out-of-domain performance in comparison to utilizing both components. To further aid the model to focus on domain in-variant features, combining both style perturbation and HRFP+ showcases the highest increase of 7.56% out-of-domain performance.

Importance of the overcompleteness in HRFP: A different spatial configuration in the fine-grained feature perturbation block is conducted with spatially consistent Feature Perturbation (SCFP) where all layers in the fine-grained feature perturbation block are spatially unchanged. The SCFP configuration causes a drop of 0.78% when compared to HRFP+ module. This could be due to the model's focus not being shifted away from domain-specific fine-grained/HF features. The results show that the combination of HRFP+ and NP+ gives the best average out of domain performance of approximately 7.56%, showcasing the need for perturbing domain-variant high frequency features along with style perturbation. Additional studies on overcompleteness of the HRFP module are provided in the supplementary material.

5. Conclusion

This paper introduces a novel Multi-Resolution Feature Perturbation (MRFP) technique, designed to address both single and multi-domain generalization challenges within Sim-2-Real context for semantic segmentation. The MRFP technique involves perturbing domain-specific fine-grained features using HRFP along with perturbing the feature channel statistics using normalized perturbation. Our approach consistently achieves superior performance across diverse domain generalization scenarios using seven urban scene segmentation datasets. MRFP has an improvement of 7.56 % compared to baseline, when trained on GTAV and tested on all target domain datasets. Importantly, this improvement is achieved without an increase in the number of parameters or computational cost during the inference phase.

References

[1] Euijoon Ahn, Dagan Feng, and Jinman Kim. A spatial guided self-supervised clustering network for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 379–388. Springer, 2021. 1

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010. 1

[3] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 458–467, 2021. 3

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1, 6

[5] Tianle Chen, Mahsa Baktashmotlagh, Zijian Wang, and Mathieu Salzmann. Center-aware Adversarial Augmentation for Single Domain Generalization. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4146–4154, Waikoloa, HI, USA, 2023. IEEE. 2, 3

[6] Zhi Chen, Yadan Luo, Ruihong Qiu, Sen Wang, Zi Huang, Jingjing Li, and Zheng Zhang. Semantics disentangling for generalized zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8712–8720, 2021. 1

[7] 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, 3, 5, 6, 7

[8] Seokeon Choi, Debasmit Das, Sungha Choi, Seunghan Yang, Hyunsin Park, and Sungrack Yun. Progressive random convolutions for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10312–10322, 2023. 1, 2, 3, 5, 6, 7

[9] 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. 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[11] Qi Fan, Mattia Segu, Yu-Wing Tai, Fisher Yu, Chi-Keung Tang, Bernt Schiele, and Dengxin Dai. Towards robust object detection invariant to real-world domain shifts. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 3, 5

[12] Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. Adversarially adaptive normalization for single domain generalization. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8208–8217, 2021. 1

[13] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 4

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

[16] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 3

[17] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6891–6902, 2021. 2

[18] 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 (CVPR)*, pages 3061–3071, 2023. 1

[19] 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

[20] Philip TG Jackson, Amir Atapour Abarghouei, Stephen Bonner, Toby P Breckon, and Boguslaw Obara. Style augmentation: data augmentation via style randomization. In *CVPR workshops*, pages 10–11, 2019. 3

[21] Jeya Maria Jose, Vishwanath Sindagi, Ilker Hacihaliloglu, and Vishal M. Patel. KiU-Net: Towards Accurate Segmentation of Biomedical Images using Over-complete Representations, 2020. arXiv:2006.04878 [cs, eess]. 3

[22] Namyup Kim, Taeyoung Son, Jaehyun Pahk, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Wedge: web-image assisted domain generalization for semantic segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9281–9288. IEEE, 2023. 6

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural net-

works. *Advances in neural information processing systems*, 25, 2012. 6

[24] 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. 3, 6, 7

[25] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3

[26] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017. 3

[27] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 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*, pages 2507–2516, 2019. 1

[28] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017. 6

[29] 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. 3, 5, 6, 7

[30] Shivam K Panda, Yongkyu Lee, and M Khalid Jawed. Agronav: Autonomous navigation framework for agricultural robots and vehicles using semantic segmentation and semantic line detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6271–6280, 2023. 1

[31] 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. 3, 5, 6, 7

[32] 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. 1, 5

[33] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, Las Vegas, NV, USA, 2016. IEEE. 1, 5

[34] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 1

[35] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 5

[36] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6, 7

[37] Aniruddh Sikdar, Sumanth Udupa, Prajwal Gurunath, and Suresh Sundaram. Deepmao: Deep multi-scale aware overcomplete network for building segmentation in satellite imagery. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 487–496, 2023. 1, 3

[38] Hugues Thomas, Ben Agro, Mona Gridseth, Jian Zhang, and Timothy D Barfoot. Self-supervised learning of lidar segmentation for autonomous indoor navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14047–14053. IEEE, 2021. 1

[39] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 23–30. IEEE, 2017. 3

[40] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999. 3

[41] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 1096–1103, Helsinki, Finland, 2008. ACM Press. 3

[42] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019. 1

[43] Nils Philipp Walter, David Stutz, and Bernt Schiele. On fragile features and batch normalization in adversarial training. *arXiv preprint arXiv:2204.12393*, 2022. 3

[44] Chaoqun Wan, Xu Shen, Yonggang Zhang, Zhiheng Yin, Xinmei Tian, Feng Gao, Jianqiang Huang, and Xian-Sheng Hua. Meta convolutional neural networks for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4682–4691, 2022. 1

[45] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8681–8691, 2020. 1, 2, 4

[46] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 1

[47] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 834–843, 2021. 1

[48] Zijian Wang, Yadan Luo, Zi Huang, and Mahsa Baktashmotlagh. FFM: Injecting Out-of-Domain Knowledge via Factorized Frequency Modification. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4124–4133, Waikoloa, HI, USA, 2023. IEEE. 2, 3

[49] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. OpenEarthMap: A Benchmark Dataset for Global High-Resolution Land Cover Mapping. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6243–6253, Waikoloa, HI, USA, 2023. IEEE. 1

[50] Qi Xu, Liang Yao, Zhengkai Jiang, Guannan Jiang, Wenqing Chu, Wenhui Han, Wei Zhang, Chengjie Wang, and Ying Tai. Dirl: Domain-invariant representation learning for generalizable semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2884–2892, 2022. 6

[51] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. *arXiv preprint arXiv:2007.13003*, 2020. 2, 3, 5

[52] Luona Yang, Xiaodan Liang, Tairui Wang, and Eric Xing. Real-to-virtual domain unification for end-to-end autonomous driving. In *Proceedings of the European conference on computer vision (ECCV)*, pages 530–545, 2018. 1

[53] 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. 6

[54] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2100–2110, 2019. 5

[55] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. 4

[56] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 1, 2