Domain Generalization for Semantic Segmentation: A Survey

Manuel Schwonberg[1,2] , Hanno Gottschalk[1]
[1]TU Berlin, Mathematical Modeling of Industrial Life Cycles Group
[2]CARIAD SE
schwonberg@campus.tu-berlin.de, gottschalk@math.tu-berlin.de

Abstract

The generalization of deep neural networks to unknown domains is a major challenge despite their tremendous progress in recent years. For this reason, the dynamic area of domain generalization (DG) has emerged. In contrast to unsupervised domain adaptation, there is no access to or knowledge about the target domains, and DG methods aim to generalize across multiple different unseen target domains. Domain generalization is particularly relevant for the task semantic segmentation which is used in several areas such as biomedicine or automated driving. This survey provides a comprehensive overview of the rapidly evolving topic of domain generalized semantic segmentation. We cluster and review existing approaches and identify the paradigm shift towards foundation-model-based domain generalization. Finally, we provide an extensive performance comparison of all approaches, which highlights the significant influence of foundation models on domain generalization. This survey seeks to advance domain generalization research and inspire scientists to explore new research directions.

1. Introduction

Over the last decade, deep neural networks (DNNs) have paved the way for major progress in the field of computer vision [8, 27, 28, 43]. That includes dense perception tasks such as semantic segmentation and object detection, which are relevant for several different applications such as automated driving. However, domain shifts, i.e. when the training and inference domain differs, challenge DNNs and cause significant performance decreases when applied in the target domain. In particular, in the field of automated driving, domain shifts are a huge challenge. Several domain shifts, like the synthetic-to-real and real-to-real shifts, are major obstacles to the perception of automated vehicles in the real world. Therefore, several research fields have emerged to mitigate the performance decrease caused by domain shifts. Unsupervised domain adaptation (UDA) is very popular, and UDA for semantic segmentation has gained a lot of attention in research, with a large diversity of approaches [13, 14, 50, 54, 60]. UDA assumes that unlabeled target data is given and aims for the adaptation to this particular domain. However, unlabeled target data is not always available, and adaptation to a particular, single domain can be hard because of the large number of different domains. In contrast, domain generalization (DG) approaches overcome these flaws by only accessing the source domain and aiming to generalize to multiple unseen target domains. In recent years, a highly relevant and dynamic research field of DG approaches has emerged with a wide variety of approaches. This necessitates a survey on this topic to first get an overview about the variety of existing approaches and second identify important research trends and future research. Although there are general surveys on domain generalization [56, 74], these works do not provide an in-depth review of DG for semantic segmentation. Only the survey by Rafi *et al*. [44] specifically addresses domain generalization for segmentation, but in our survey we cover a larger number of approaches, include foundation-model-based DG approaches, and provide a more fine-grained taxonomy. Our survey focuses on synthetic-to-real domain generalization because of the large number of approaches and the uniform benchmarking. Overall, the contributions of this survey are as follows:

- we review and cluster over 40 synthetic-to-real domain generalization approaches for DG making it the most extensive survey of DG for semantic segmentation so far
- we identify and analyze the paradigm shift from classic DG approaches towards foundation-model-based DG approaches
- we provide an extensive performance comparison of all reviewed approaches which reveals differences between source datasets and underlines the influence of foundation models

Our survey is structured as follows. First, we will introduce and explain our taxonomy which highlights the paradigm shift towards foundation models. Subsequently, we will review and analyze each of the categories of the taxonomy in

detail and conclude with an extensive performance comparison of the approaches.

2. Taxonomy

Recently, a paradigm shift in DG has been observed, driven by foundation models. In contrast to "classic" DG approaches, which primarily aim at extracting generalizing features from the source domain, the new foundation-model-based methods focus on preserving and exploiting the broad and already generalizing knowledge of pre-trained models. Due to their large-scale vision or vision-language pre-training on huge web-crawled datasets, foundation models already encode highly generalized knowledge that can be adapted to the segmentation task. The labeled source dataset is then used to adapt the pre-trained model, but not primarily to learn generalizing representations. For this reason, DG approaches can first be divided into classic- and foundation-model-based approaches, as shown in the taxonomy in Figure 1.

Among the classic approaches, a wide variety of methods have been developed. They can be further categorized into input, feature, and output space approaches, whereas for the recently emerged foundation model approaches, only feature space methods exist so far. Several DG approaches may propose multiple contributions, for example, in both input and feature space. In this case, they appear multiple times in the taxonomy in Figure 1. Here, each subcategory, like, e.g. feature regularization, provides a comprehensive overview of all works. In addition to taxonomy, a comprehensive performance comparison of domain generalization approaches is provided in Table 1 which will be discussed in Section 5.

3. Classic Domain Generalization Approaches

3.1. Input Space

Image Augmentation Image augmentations are a simple and straightforward way to improve generalization by extending/diversifying the input distribution of the network. Therefore, in the context of DG, image augmentations are often referred to as domain randomization. Augmentations are essential for domain generalization, as they are relevant not only for input space methods, but also for, e.g. consistency losses, feature augmentations, or contrastive learning methods to obtain multiple versions of the same image.

DG approaches perform domain randomization by applying either image transformations or a style transfer, but without modifying the content. Similar to Schwonberg *et al*. [49] with simple standard image augmentations, both RICA [52] and MoDify [22] rely on image transformations. RICA [52] converts RGB images to the CIELAB color space and randomizes the mean and standard deviation along with the range of the color values, arguing that augmentations can be applied more effectively in the CIELAB color space. MoDiFy [22] relies on a simpler randomized RGB channel swapping whose intensity is controlled by a loss memory bank. This is done in combination with a so-called difficulty-aware training strategy to adapt the difficulty of the training samples to the training progress of the network. Following the basic idea of RICA [52], FSDR [17] and PASTA [5] randomize the input of the source domain in the frequency space. PASTA [5] proposes a simple augmentation by first applying a Fast Fourier Transform to the RGB image and then augmenting the amplitude spectrum. The higher the frequency, the stronger the modification. An inverse Fourier transform provides the augmented RGB image, and PASTA can be used complementary to other DG methods. FSDR [17] also works in the frequency domain, but accesses ImageNet samples to transfer their style to the frequency domain via histogram matching, and employs a bi-directional learning scheme to identify domain-invariant frequencies which are not randomized. As the only DG work, SIL [69] by Zhang *et al*. proposes a shape augmentation to enforce shape-invariant learning. The authors use a so-called random elastic deformation algorithm to distort both the inner parts of objects and their boundaries in small local regions of the image. In addition, they use edge and contour transformations, like e.g. Canny edge detection, to extract shape information and further enforce shape-invariant learning.

Similarly to UDA, in the context of style transfer and also for feature augmentation methods (see Section 3.2), adaptive instance normalization (AdaIN) [19] is highly relevant and widely used. AdaIN [19] allows style transfer from a style reference image to a content image by transferring the feature mean and variance of the style image while preserving the content. However, unlike UDA, a target domain is not available as the style input. Several solutions have been proposed to mitigate this problem. DPCL [63] by Yang *et al*. employs standard image augmentations to create different stylized images and employs AdaIN to transfer the style of the augmented images. The authors then train a network with a $L1$ reconstruction loss that reconstructs the original images from the augmented images, which should provide images similar to the source style during inference. Zhong *et al*. [73] with AdvStyle mitigate the style references problem by formulating the mean and variance of AdaIN as learnable parameters that are learned in an adversarial manner. In a two-step process, the adversarial style-parameters are first learned and then used to augment the original source dataset. Other works use external data to randomize the style of the source domain [26, 34, 66], often with samples from ImageNet and various style transfer techniques. DRPC [66] employs an image-to-image translation model, TLDR [26] uses PhotoWCT [32] for style transfer from ImageNet, and CSDS [34] integrates AdaIN into its class-discriminative style transfer approach based on Ima-

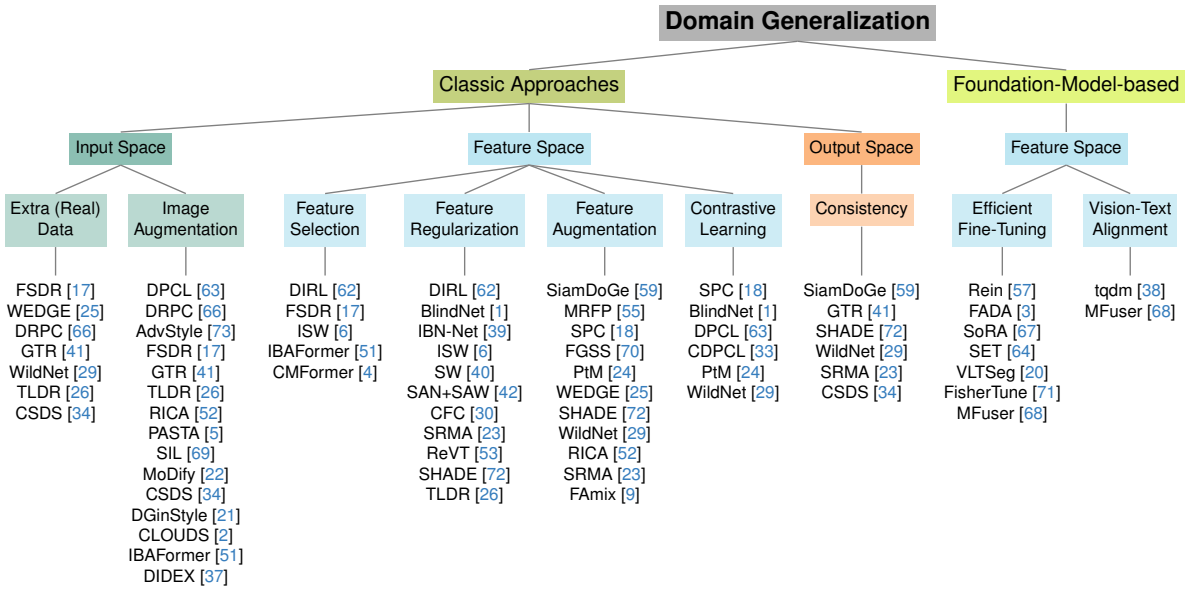

Figure 1. **Taxonomy of Domain Generalization Approaches for Semantic Segmentation**

geNet. Peng *et al*. also use AdaIN for their GTR [41] approach, but as the only DG approach, painted images are used to randomize the style of the source domain. Since the paintings contain a variety of styles, an additional module is needed to filter out too low- and too high-complexity paintings, which were found to be detrimental for style transfer.

In contrast to unsupervised domain adaptation, GANs are rarely used for style transfer in domain generalization approaches because the target domain is not available. However, the recently emerged text-to-image diffusion models [46] are capable of generating or translating photorealistic images without any target domain samples. DGinStyle [21] and CLOUDS [2] exploit this capability to augment the original synthetic source dataset with photorealistic samples generated by a text-to-image diffusion model. CLOUDS [2] generates prompts to describe traffic scenes with an LLM and uses them to generate images with stable diffusion [46]. Since no pixel-level annotations are available for the diffusion-generated images, the authors train on these images as additional training data with pseudo-labels in a teacher-student setting. DGinStyle [21] proposes a more complex framework, since it fine-tunes the diffusion models on the source domain and develops a so-called multi-resolution latent fusion module to better generate smaller objects. However, the fine-tuning step raises the problem of forgetting pre-trained knowledge, and DGinStyle proposes a three-step process including DreamBooth [48] to preserve the original diffusion knowledge. Equipped with these two enhancements, photorealistic images conditioned on text and semantic maps are generated, and a new source domain dataset is created. It consists of half original source domain samples and half diffusion-generated samples and can be combined with any DG method. DIDEX [37] also exploits diffusion models and creates a photorealistic pseudo-target domain using image-to-image translation. Using that, it obtains a generalization by using UDA methods to adapt the model towards this pseudo-target domain.

Extra (Real) Data Existing methods mainly use different types of extra data: ImageNet data [17, 26, 29, 34, 66], paintings [41] and real web-crawled data such as Wedge [25]. All methods that access additional data are listed separately in the taxonomy in Figure 1. The domain generalization paradigm assumes that there is only one (or more) labeled source domain(s), but no target domain data, which in the synthetic-to-real shift means no real-world data. However, as described above, many methods access real-world data, such as ImageNet for style transfer. The active use of real-world data gives these methods a significant advantage, because real-world styles and patterns can be incorporated into the training. While this strategy of incorporating additional real-world data is pragmatically reasonable, it compromises scientific comparability and also circumvents the real problem of domain generalization by using proxy real-world domains.

3.2. Feature Space

Feature Selection While all DG approaches aim at learning domain-invariant features, some approaches explicitly aim to distinguish between domain-dependent and domain-independent features in order to treat them differently. These approaches can be categorized under the term feature selection, since they treat both types of features differently.

We can further distinguish between direct feature selection and indirect feature selection methods. DIRL [62] utilizes augmentations to identify domain-invariant features. Since the augmentations introduce a style shift, the authors compute the $L2$ distance between features from the original and augmented images, and a higher distance indicates a higher domain sensitivity. This sensitivity vector guides the learning of attention weights that force the model to focus on domain-invariant features. In a very similar way with augmented images, Choi *et al*. [6] identify style-sensitive features for their instance selective whitening approach (ISW), but unlike DIRL [62] they find these features by k-means clustering on the covariance matrix. FSDR [17] by Huang *et al*. aims to identify domain-invariant components in the frequency space. The prediction entropy serves as a criterion for identifying domain-variant frequency bands. High entropy indicates high domain dependence, and these frequencies are then randomized by histogram matching of real-world frequencies from ImageNet samples.

With the emergence of vision transformer architectures for domain generalization, a line of work has proposed advanced attention mechanisms to better facilitate generalized representation learning [4, 51]. Since these improvements implicitly allow the networks to focus on domain-invariant features, they can be seen as indirect feature selection methods. IBAFormer [51] proposes two intra-batch attention mechanisms that allow cross-image attention within a batch and should help the network to extract domain-invariant features. The mean intra-batch attention (MIBA) computes the mean of the features and allows the network attention between single images and the mean features. In contrast, the element-wise intra-batch attention (EIBA) computes the attention between all individual images of the batch separately. Finally, IBAFormer adds intra-batch attention fusion modules directly before the decoder. The authors of CMFormer [4] observed that high-resolution features are more focusing on content, while the low-resolution features are more sensitive to style. Therefore, they propose a mask-attention mechanism that processes high- and low-resolution features separately and fuses both features through a concatenation followed by a linear layer.

Feature Regularization This category groups DG approaches that employ feature regularization objectives to constrain representation learning to domain-invariant features. Among them, several works employ instance normalization (IN) or instance whitening (IW) techniques, as these techniques have been found to be beneficial for domain generalized learning. While IN only normalizes each channel, instance whitening forces the feature covariance across channels to be decorrelated. Several works have shown that style-sensitive information are encoded in the feature covariance [6, 10, 31] and thus decorrelation can help to learn style-invariant features.

IBN-Net [39] is one of the foundational works for instance normalization and batch normalization (BN). Pan *et al*. find that IN removes style-related information in the earlier layers of the network, thereby improving style-invariant learning, but has a detrimental effect in the deeper layers, where BN layers are beneficial to preserve content. IBN-Net is used as a complementary method in later works [6, 62]. Li *et al*. [30] argue that instance normalization can be detrimental to the spatial feature distribution. Therefore, they introduce a contour learning module to regularize the network and enforce the learning of clear object boundaries. IBN-Net [39] shows that designing a network with the exact positioning of normalization layers is a challenge. For this reason, Pan *et al*. [40] introduce switchable whitening (SW), which learns importance weights for either only batch and instance whitening or for five regularization techniques including batch, instance and layer normalization. In this way, the network can select the best whitening or normalization technique depending on the task and the dataset. Two other problems of whitening are the lack of awareness of already domain-invariant features and semantic classes [6, 42]. Choi *et al*. [6] propose instance selective whitening (ISW) based on their observation that naive whitening hinders domain-invariant discriminative features. ISW computes the variance between two covariance matrices of a non-augmented and an augmented image to identify areas sensitive to style. k-means then clusters high-variance and domain-variant features so that a filter can be applied to the naive whitening loss and only domain-variant features are decorrelated. Similarly, DIRL [62] uses ISW, but replaces the covariance-based sensitivity feature identification with its own sensitivity vector described above and forces the 30% most domain-variant features to be decorrelated. Peng *et al*. [42] develop semantic-aware normalization and whitening (SAN + SAW). Based on the classifier predictions, SAW rearranges the feature channels into groups and each group contains channels of different classes. Group instance whitening is then applied to these groups so that channels of the same class are not decorrelated, which is detrimental to the semantic features. Unlike any other whitening work, BlindNet [1] introduces a covariance regularization that enforces the covariance between the features of the original and an augmented image to be the identity matrix. In addition, Ahn *et al*. also introduce a direct covariance consistency by computing the covariance matrices for both images and enforcing similarity with a $L2$-loss.

ImageNet knowledge can be used either actively by directly accessing ImageNet samples as described in Section 3.1, or passively by using the encoded knowledge in pre-trained ImageNet models as a regularization. SHADE [72] and SRMA [23] both utilize this pre-trained knowledge of the real-world for feature space regularization. While SHADE, similar to DAFormer [14], simply uses a $L2$-loss between

the frozen ImageNet features and the features of its trained model, SRMA extends this idea significantly. First, the regularization is applied to multiple layers, and second, the regularization loss has three parts: similar to SHADE [72], an element-wise $L2$-loss, another $L2$-loss between the semantic centers of the thing classes, and a global regularization loss for the feature maps after global average pooling. TLDR [26] also implements ImageNet-based regularization, but the authors argue that direct regularization may not be optimal. For this reason, they compute the Gram matrix of both the synthetic and ImageNet features and force the Gram matrices to be similar by a $L2$-loss.

In addition to BlindNet [1], DRPC [66] is the only work that applies a consistency loss in the feature space, which can be understood as a form of regularization. Yue *et al*. average the feature maps over multiple images of different styles and apply a $L1$ consistency loss that forces the features of each pyramidal feature map to be similar to the mean across all images. The approach ReVT [53] differs from all other feature regularization methods in that it transfers the concept of model soups [58] to the area of domain generalization. ReVT averages the encoder weights of three encoders trained independently but with the same augmentations, and uses the new averaged encoder weights for inference. Also, averaging the decoder was found to be detrimental, so one of the three decoders is selected.

Feature Augmentation Similar to the input space augmentations, the feature maps of the network can also be augmented, mixed, or randomized. These techniques are closely related to the input space style transfer methods, since methods such as AdaIN [19] perform style transfer or enhancement in the feature space. In addition, the mean and covariance of the features that AdaIN exploits are highly relevant, as many approaches manipulate them for feature randomization or augmentation.

Similar to the UDA approach DLOW [11], SiamDoGe [59] proposes to augment the features by interpolating the feature statistics within the source domain. For this, a non-augmented and an augmented source image are processed by a Siamese-like network. Based on the resulting two feature maps, SiamDoGe randomly linearly interpolates between the original and augmented feature maps to compute new intra-domain features. SHADE [72] follows a similar idea with its style hallucination module. Starting with the base source styles, SHADE applies fast point sampling to sample styles,i.e., mean and variance. Then, similar to AdaIN, new feature statistics are computed with randomly sampled weights.

Style or memory banks are another common technique for feature augmentation [18, 24, 70], and in all cases a current feature map is augmented with stored knowledge from a memory bank. FGSS [70] implements a simple memory bank mechanism in which class-wise prototypes are stored and updated by exponential moving average. The augmented feature map is computed by a random combination of current features and the semantic class prototype. In contrast, both SPC [18] and Pin the memory [24] propose a weighted feature memory bank augmentation scheme. SPC [18] collects a memory bank of styles during training via EMA. For new samples during inference, the Wasserstein distance is calculated for all styles in the memory bank, and the augmented feature statistics are obtained by a weighted recombination of styles in the memory bank. The smaller the distance, the higher the weights for that style. During training, this serves as a feature augmentation with other styles, and during inference as a projection of new unknown styles into the known learned style space. In contrast, PtM [24] employs a memory matrix that is trained to contain domain-invariant features from a meta-learning framework. The update process is performed by a small update network, and a cosine similarity matrix is computed during read access. The fused feature map is obtained as a concatenation of the original features and the weighted memory features. FAmix [9] stands out among memory bank approaches because it is the first work to use CLIP [43]. Fahes *et al*. propose a two-step process in which the first step is the collection of a style bank. Unlike other works, FAmix uses prompting and the CLIP text encoder to increase the feature diversity of its style bank. The second is the training step, which uses AdaIN to linearly augment the original feature statistics with randomly sampled style bank statistics.

Both MRFP [55] and RICA [52] employ a separate network for feature randomization. MRFP [55] proposes a simple setup where a separate lightweight inverted encoder-decoder network of random convolutions randomizes the features coming from a certain layer and adds them to the output of the same layer. RICA [52] shares a similar idea, but implements a separately trained GAN directly after the first convolutional layer to augment the earlier features. Especially for deeper layers, such feature augmentation is detrimental for the generalization. The CycleGAN [75] architecture and principle is adopted with two generators, two discriminators, and two network streams with original and augmented input images. Hereby, the GAN effectively learns to generate features that are indistinguishable from the feature maps of the augmented network stream.

WEDGE [25] and WildNet [29] both propose a style transfer in the feature space. Similarly, SRMA [23] also uses AdaIN to normalize the features. However, instead of accessing real features, it recalculates the synthetic feature statistics by a randomly weighted linear recombination of all the individual class statistics. While both SRMA [23] and WildNet [29] normalize the original feature statistics following AdaIN [19], WEDGE introduces a more complex style transfer into the feature space without using AdaIN. The core element of its so-called style injection is a pro-

jection matrix. First, the cosine similarity between the synthetic and real features is computed. The projection matrix then maps the synthetic features into a similar embedding space as the real features, weighted by the cosine similarity. Similarly to RICA [52], the authors find that the style injection works best in the earlier layers and worse when applied only to the deeper layers.

Contrastive Learning Contrastive learning is a popular technique in the field of self-supervised learning and has been extensively studied for UDA [50]. It is also widely used for domain generalization, although the construction of positive and negative pairs is more difficult due to the lack of samples from the target domain. As a substitute, DG approaches often use augmentations or style transfer to obtain style-modified image pairs for contrastive learning. Another common feature of contrastive DG approaches is the utilization of several contrastive loss functions.

One strategy for contrastive learning can be two different loss functions, where one loss enforces intra-cluster compactness and the other loss enforces inter-cluster dispersion of class prototypes. SPC [18] and PtM [24] follow this strategy and both rely only on samples from the source domain without the need for augmentations. SPC [18] proposes to perform segmentation with clustering, which requires a well-structured feature space. For this, the authors implement a clustering loss that minimizes the distance between pixel embeddings and class prototypes, and another loss adopted from the InfoNCE loss to disperse the class prototypes. The idea behind PtM [24] is the same, but with different loss functions due to the meta-learning scheme. During the meta-training phase of PtM, a memory bank is constructed and a customized compactness loss is employed to obtain features that are semantically close to the memory elements. The dispersion loss, on the other hand, ensures that the different class elements of the memory bank are well distributed in the feature space.

Other approaches focus on the commonly used InfoNCE [12] loss. Both WildNet [29] and BlindNet [1] rely on similar contrastive losses and both require style-modified input pairs. WildNet [29] has two InfoNCE-based losses for two differently stylized images that train the network to learn style-invariant features. They define the same pixel of their image pair as the positive pair and all pixels from other locations as the negative pairs, while excluding pixels of the same class. BlindNet [1] relies on the same definition of positive and negative pairs, but the second contrastive loss aims to learn from misclassified pixels by taking the wrong classes as negative samples. CDPCL [33] modifies the InfoNCE loss with an uncertainty weighting matrix. This is computed from the cosine similarity between the original and augmented class prototypes and reflects how likely the class prototypes will differ between domains.

Unlike other works, DPCL [63] introduces three contrastive loss functions: pixel-pixel, class-to-pixel, and instance-to-class. Only the pixel-class loss is based on the InfoNCE loss. The pixel-pixel loss is based on transition probability matrices for both the prediction and the ground truth and is optimized by the Jensen-Shannon divergence. The ground truth transition probabilities force the features to be similar for the same class and dissimilar for other classes.

3.3. Output Space

Consistency Learning Consistency methods rely on a straightforward principle and all have in common that the network is expected to produce the same predictions when the style or texture of the input changes. In this context, augmentations and style transfer are essential for these methods to obtain different samples with the same content, but different styles that can be used for consistency loss. GTR [41] applies one of the simplest forms of a consistency loss, which is implemented as a $L1$ loss between the predictions for its locally and globally randomized input images. SiamDoGe [59] follows the same principle as GTR [41], but applies a consistency loss on both output- and feature-level and weights it by a $L1$ feature distance.

Several other works use the KL-divergence or Jensen-Shannon divergence to enforce prediction consistency between two inputs. Both WildNet [29] and CSDS [34] utilize the KL-divergence. Similar to that, SHADE [72] and SRMA [34] use the Jensen-Shannon divergence between different stylized images.

4. Foundation-model-based Domain Generalization Approaches

This category of DG approaches has only recently emerged, so the number and diversity of works is still limited.

Efficient Fine-tuning Similar to the basic idea of LoRA [16], most of the approaches of this category freeze the pre-trained parameters of the foundation model and propose adapters between the frozen layers to obtain a foundation model adapted for domain generalized segmentation. Next to our own approach, Rein [57] was one of the first approaches to follow this idea. Specifically, Rein aims to overcome the gap between the task and the pre-training by introducing a set of learnable tokens that can learn correspondences between the pre-trained features and the segmentation instances. For this, Rein computes a similarity matrix between the tokens and the features, followed by another feature fusion and an MLP. Building on this idea, both FADA [3] and SET [64] propose adapters in the frequency space. Their basic ideas share conceptual similarities with the classic DG approaches, such as FSDR [17] and PASTA [5]. FADA [3], similar to many classic DG works, argues that domain-invariant information can be better learned in the frequency space and implements the Haar wavelet trans-

form directly on the pre-trained features. The approach then splits the features into a low- and high-frequency stream, both of which have separate learnable tokens and a similarity matrix for correlation between the pre-trained features. Only for the high-frequency components, instance normalization is applied to this matrix to suppress style-dependent information. Similarly, SET [64] obtains the phase and amplitude spectra of the frozen pre-trained layers and introduces learnable tokens for both. Similar to FADA [3], a similarity matrix is calculated and forwarded into a MLP. The amplitude stream, like the high-frequency stream of FADA, includes instances normalization to remove style-variant domain-dependent features. As an extension to these works, the MFuser [68] uses a frozen VLM and a vision-foundation model in parallel and proposes a fusion module as a connection between both networks. In contrast to the very similar, frequency-space-focused works, SoRA [67] proposes an improvement for LoRA [16] tailored to the task of domain generalized segmentation. The authors observed that only the smaller singular components carry domain-specific knowledge, while the larger components encode generalized knowledge. For this reason, SoRA fine-tunes only the minor singular components to adapt the network to the new task and preserve the pre-trained knowledge. In contrast to other approaches, VLTSeg [20] proposes a simple setting for fine-tuning without freezing the weights and training the entire foundation model in the source domain. FisherTune [71] builds on this idea and uses an approximate Fisher information matrix to estimate the domain sensitivity of the parameters. Controlled by this matrix, it only fine-tunes the most domain-sensitive parameters.

Vision-Language Alignment The majority of works in this field do not utilize the text encoder after the vision-language pre-training, and only adopt the vision encoder for the downstream DG training. In contrast to this trend, tqdm [38] and MFuser [68] are the only approaches that used the text encoder during the downstream training. The authors of tqdm use the class-wise text embeddings from a frozen CLIP encoder as textual queries and text clusters for a modified cross-attention mechanism in the network decoder. The cross-attention has a text-to-pixel attention block and computes attention weights based on the text clusters, which are then used to update the pixel features. The text queries refined by the decoder are also used for the final segmentation. In addition, tqdm regularizes the network with a text, a text-vision, and a vision alignment loss. Similar to tqdm, MFuser [68] introduces a vision-text alignment module with an attention and a mamba block [35].

5. Performance Comparison

Similar to UDA, GTA5 [45] and Synthia [47] are used as synthetic source datasets for the vast majority of DG approaches. Cityscapes [7], BDD100k [65] and Mapillary [36] are used as the unseen datasets for DG performance evaluation. This setting allows for a very consistent and comparable benchmarking of DG approaches. For this reason, a list of the performance of all DG approaches is provided for both synthetic source datasets in Table 1.

First, we can clearly observe how the different backbone architectures ResNet-101, MiT-B5 and ViTs and their pre-trainings affect the domain generalization capabilities. While the best performing approach with a ResNet-101 architecture and ImageNet initialization reaches 48.0% mIoU for the DG Mean, this increases to 54.7% mIoU for CLOUDS [2] with a CLIP initialization while the approach clearly benefits from three large foundation models. However, when the SegFormer [61] backbone MiT-B5 is used, the current peak performance reaches 57.8 % mIoU. Finally, the foundation models with their large-scale pre-training raise the DG performance to another level, and the very recent approach SoRA [67] achieves 68.3 % mIoU for the average over three real-world datasets.

The list of DG approaches in Table 1 also confirms the observation from the UDA analysis that generalizing from the Synthia dataset is more difficult than for GTA5. For example, if we look at the absolute peak performance across all architectures and initializations, we see that it is significantly lower at 55.1% mIoU compared to GTA5 at 68.3%. Notably, the performance for GTA5 is reported across 19 classes and for Synthia as the mean of 16 classes. The newer foundation model approaches also clearly struggle when trained on Synthia. While the smaller dataset size (9.400 for Synthia vs. 24.499 for GTA5) is given as a reason [4] for the lower performance, other reasons such as the different viewpoint of Synthia and also the lower rendering quality should be taken into account.

Another trend can be observed in the given performance comparison. It is easy to see that the performance gains of new DG approaches with a ResNet-101 backbone are comparatively small. Between TLDR [26] from early 2023 and the current leading approach CSDS [34], there is a small improvement of only 0.9% mIoU in the DG mean. Also, the other performances show that many approaches achieve a performance in a similar range without a major gain. This trend is similar to UDA, where stagnation in performance improvements was also observed. However, similar to the field of unsupervised domain adaptation, new architectures like the MiT-B5 and foundation models [20, 57, 67] have enabled new breakthroughs in performance and methodology.

6. Conclusion & Outlook

Overall, this survey provides a comprehensive overview of the quickly evolving and relevant field of domain generalization for semantic segmentation. We identified a

Table 1. **Performance Comparison** of all domain generalization approaches for semantic segmentation. $\mathcal{D}_{\text{val}}^{\text{CS}}$ denotes the validation set of Cityscapes [7], $\mathcal{D}_{\text{val}}^{\text{BDD}}$ from BDD100k [65] and $\mathcal{D}_{\text{val}}^{\text{MV}}$ from Mapillary [36].

DG Method	Enc.	Init	$\mathcal{D}^{\text{S}}$: **GTA5** [45]				$\mathcal{D}^{\text{S}}$: **SYNTHIA** [47]			
			$\mathcal{D}_{\text{val}}^{\text{CS}}$	$\mathcal{D}_{\text{val}}^{\text{BDD}}$	$\mathcal{D}_{\text{val}}^{\text{MV}}$	**DG mean**	$\mathcal{D}_{\text{val}}^{\text{CS}}$	$\mathcal{D}_{\text{val}}^{\text{BDD}}$	$\mathcal{D}_{\text{val}}^{\text{MV}}$	**DG mean**
DIRL [62]	ResNet-50	ImageNet	41.0	39.2	41.6	40.6	-	-	-	-
SiamDoGe [59]			43.0	37.5	40.6	40.4	-	-	-	-
MRPF [55]			42.4	39.6	44.9	42.3	27.9*	25.7*	26.0*	26.5*
SPC [18]			44.1	40.5	45.5	43.4	-	-	-	-
BlindNet [1]			**45.7**	**41.3**	**47.1**	**44.7**	-	-	-	-
FGSS [70]			44.5	36.5	44.3	41.8	39.2	29.8	33.5	34.2
DPCL [63]			44.9	40.2	46.7	43.9	-	-	-	-
CDPCL [33]			42.3	39.5	42.4	41.4	**41.2**	**35.4**	**35.5**	**37.4**
IBN-Net° [39]	ResNet-101	ImageNet	37.7	34.2	36.8	36.2	-	-	-	-
ISW° [6]			37.3	38.7	38.1	38.0	-	-	-	-
DRPC [66]			42.5	38.7	38.1	39.8	37.6	34.3	34.1	35.4
SW [40]			36.1	36.6	32.6	35.1	31.6	35.5	29.3	32.1
AdvStyle [73]			39.5	36.4	36.1	37.3	37.6	27.5	31.8	32.3
FSDR [17]			44.8	41.2	43.4	43.1	40.8	39.6	37.4	39.3
PtM [24]			44.9	39.7	41.3	42.0	-	-	-	-
SAN+SAW [42]			45.3	41.2	40.8	42.4	40.9	36.0	37.3	38.1
WEDGE [25]			45.2	41.1	48.1	44.8	40.9	38.1	**43.1**	40.7
GTR [41]			43.7	39.6	39.1	40.8	39.7	35.3	36.4	37.1
SHADE [72]			46.7	43.7	45.5	45.3	-	-	-	-
WildNet° [29]			45.8	41.7	47.1	44.9	-	-	-	-
CFC [30]			41.1	37.8	44.4	41.1	-	-	-	-
TLDR [26]			47.6	44.9	48.8	47.1	42.6	35.5	37.5	38.5
RICA [52]			48.0	45.2	46.3	46.5	**45.0**	36.3	41.6	41.0
PASTA [5]			45.3	42.3	48.6	45.4	-	-	-	-
SIL [69]			48.0	40.3	47.4	45.2	40.9	33.4	33.4	35.9
MoDify [22]			**48.8**	44.2	47.5	46.8	43.4	**39.5**	42.3	**41.7**
SRMA [23]			48.4	45.0	49.6	47.7	-	-	-	-
CSDS [34]			48.5	**46.1**	49.4	**48.0**	43.9	38.2	40.6	40.9
DGinStyle [21]			46.9	42.8	**50.2**	46.6	-	-	-	-
FAMix [9]	R101	CLIP	49.5	46.4	52.0	49.3	-	-	-	-
CLOUDS [2]			**55.7**	**49.3**	**59.0**	**54.7**	**49.1**	**40.3**	**50.1**	**46.5**
HRDA [15]	MiT-B5	ImageNet	57.4	49.1	61.2	55.9	-	-	-	-
DGinStyle [21]			58.6	52.3	62.5	57.8	-	-	-	-
DIDEX [37]			**62.0**	**54.3**	**63.0**	**59.7**	**59.8**	**47.4**	**59.5**	**55.6**
ReVT [53]			50.0	48.0	52.8	50.2	46.3	40.3	44.8	43.8
IBAFormer [51]			56.3	49.8	58.3	54.8	50.9	44.7	50.9	48.7
CMFormer [4]	Swin-L	IN1k	55.3	49.9	60.1	55.1	44.6	33.4	43.3	40.4
VLTSeg [20]	EVA02-L	EVA02-CLIP	65.3	58.3	66.0	63.2	56.8	51.9	**55.1**	54.6
tqdm [38]	EVA02-L	EVA02-CLIP	68.9	59.2	70.1	66.1	**58.0**	**52.4**	54.9	**55.1**
MFuser [68]	EVA02-L	EVA02-CLIP	**70.2**	**63.1**	**70.1**	**71.3**	54.2	46.7	53.2	51.4
Rein [57]	ViT-L	DinoV2	66.4	60.4	66.1	64.3	-	-	-	-
SET [64]			68.1	61.6	67.7	65.8	49.7	45.5	49.5	48.2
SoRA [67]			**71.8**	61.3	**71.7**	**68.3**	-	-	-	-
FADA [3]			68.2	61.9	68.1	66.1	**50.0**	**45.8**	**49.9**	**48.6**
FisherTune [71]			68.2	**63.3**	68.0	66.5	-	-	-	-

paradigm change from classic DG approaches towards foundation-model approaches which has enabled significant performance progress. For both classic and foundation-model-based approaches, we define a fine-grained clustering of input, feature, and output space approaches. Domain generalization with foundation models is a promising future research direction because the research in this area is still at an initial stage but shows a strong performance. Beyond this, it can also be valuable to extend the scope of DG research towards end-to-end automated driving approaches and research domain generalization on a system-level.

Acknowledgment

This work was supported by the German Federal Ministry for Economic Affairs and Climate Action within the project “Safe AI Engineering” under Grant 19A24004U.

References

[1] Woo-Jin 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3616–3626, 2024. 3, 4, 5, 6, 8

[2] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. Collaborating foundation models for domain generalized semantic segmentation. In *Proc. of CVPR*, pages 3108–3119, 2024. 3, 7, 8

[3] Qi Bi, Jingjun Yi, Hao Zheng, Haolan Zhan, Yawen Huang, Wei Ji, Yuexiang Li, and Yefeng Zheng. Learning frequency-adapted vision foundation model for domain generalized semantic segmentation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 3, 6, 7, 8

[4] Qi Bi, Shaodi You, and Theo Gevers. Learning content-enhanced mask transformer for domain generalized urban-scene segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 819–827, 2024. 3, 4, 7, 8

[5] Prithvijit Chattopadhyay*, Kartik Sarangmath*, Vivek Vijaykumar, and Judy Hoffman. Pasta: Proportional amplitude spectrum training augmentation for syn-to-real domain generalization. In *Proc. of ICCV*, 2023. 2, 3, 6, 8

[6] 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 *Proc. of CVPR*, pages 11580–11590, virtual, 2021. 3, 4, 8

[7] 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 *Proc. of CVPR*, pages 3213–3223, Las Vegas, NV, USA, 2016. 7, 8

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[9] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. A simple recipe for language-guided domain generalized segmentation. In *Proc. of CVPR*, pages 23428–23437, 2024. 3, 5, 8

[10] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image Style Transfer Using Convolutional Neural Networks. In *Proc. of CVPR*, pages 2414–2423, Las Vegas, NV, USA, 2016. 4

[11] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. DLOW: Domain Flow for Adaptation and Generalization. In *Proc. of CVPR*, pages 2477–2486, Long Beach, CA, USA, 2019. 5

[12] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 6

[13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Philip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In *Proc. of ICML*, pages 1989–1998, Stockholm, Sweden, 2018. 1

[14] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation. In *Proc. of CVPR*, pages 9924–9935, New Orleans, LA, USA, 2022. 1, 4

[15] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Domain Adaptive and Generalizable Network Architectures and Training Strategies for Semantic Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–15, 2023. 8

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6, 7

[17] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. FSDR: Frequency Space Domain Randomization for Domain Generalization. In *Proc. of CVPR*, pages 6891–6902, 2021. 2, 3, 4, 6, 8

[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*, pages 3061–3071, 2023. 3, 5, 6, 8

[19] Xun Huang and Serge Belongie. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *Proc. of ICCV*, pages 1501–1510, Venice, Italy, 2017. 2, 5

[20] Christoph Hümmer, Manuel Schwonberg, Liangwei Zhou, Hu Cao, Alois Knoll, and Hanno Gottschalk. Strong but simple: A baseline for domain generalized dense perception by clip-based transfer learning. In *Proceedings of the Asian Conference on Computer Vision*, pages 4223–4244, 2024. 3, 7, 8

[21] 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 *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2023. 3, 8

[22] Xueying Jiang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Domain generalization via balancing training difficulty and model capability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18993–19003, 2023. 2, 3, 8

[23] Guanlong Jiao, Chenyangguang Zhang, Haonan Yin, Yu Mo, Biqing Huang, Hui Pan, Yi Luo, and Jingxian Liu. Semantic-rearrangement-based multi-level alignment for domain generalized segmentation. *arXiv preprint arXiv:2404.13701*, 2024. 3, 4, 5, 8

[24] Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. Pin the memory: Learning to general-

ize semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4350–4360, 2022. 3, 5, 6, 8

[25] Namyup Kim, Taeyoung Son, Cuiling Lan, Wenjun Zeng, and Suha Kwak. WEDGE: Web-Image Assisted Domain Generalization for Semantic Segmentation. *arXiv:2109.14196*, pages 1–14, 2021. 3, 5, 8

[26] Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. Texture learning domain randomization for domain generalized segmentation. *Proc. of ICCV*, 2023. 2, 3, 5, 7, 8

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25(6):84–90, 2012. 1

[29] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proc. of CVPR*, pages 9936–9946, 2022. 3, 5, 6, 8

[30] Xinhui Li, Mingjia Li, Xiaopeng Li, and Xiaojie Guo. Learning generalized knowledge from a single domain on urban-scene segmentation. *IEEE Transactions on Multimedia*, 25: 7635–7646, 2022. 3, 4, 8

[31] 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. 4

[32] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A Closed-Form Solution to Photorealistic Image Stylization. In *Proc. of ECCV*, pages 453–468, Munich, Germany, 2018. 2

[33] Muxin Liao, Shishun Tian, Yuhang Zhang, Guoguang Hua, Wenbin Zou, and Xia Li. Calibration-based dual prototypical contrastive learning approach for domain generalization semantic segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2199–2210, 2023. 3, 6, 8

[34] Muxin Liao, Shishun Tian, Binbin Wei, Yuhang Zhang, Wenbin Zou, and Xia Li. Class-balanced sampling and discriminative stylization for domain generalization semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2024. 2, 3, 6, 7, 8

[35] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024. 7

[36] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proc. of ICCV*, pages 4990–4999, 2017. 7, 8

[37] Joshua Niemeijer, Manuel Schwonberg, Jan-Aike Termöhlen, Nico M Schmidt, and Tim Fingscheidt. Generalization by adaptation: Diffusion-based domain extension for domain-generalized semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2830–2840, 2024. 3, 8

[38] Byeonghyun Pak, Byeongju Woo, Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. Textual query-driven mask transformer for domain generalized segmentation. In *European Conference on Computer Vision*, pages 37–54. Springer, 2025. 3, 7, 8

[39] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net. In *Proc. of ECCV*, pages 464–479, Munich, Germany, 2018. 3, 4, 8

[40] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable Whitening for Deep Representation Learning. In *Proc. of ICCV*, pages 1863–1871, 2019. 3, 4, 8

[41] Duo Peng, Yinjie Lei, Lingqiao Liu, Pingping Zhang, and Jun Liu. Global and Local Texture Randomization for Synthetic-to-Real Semantic Segmentation. In *IEEE TIP*, 30: 6594–6608, 2021. 3, 6, 8

[42] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-Aware Domain Generalized Segmentation. In *Proc. of CVPR*, pages 2594–2605, 2022. 3, 4, 8

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of ICML*, pages 8748–8763. PMLR, 2021. 1, 5

[44] Taki Hasan Rafi, Ratul Mahjabin, Emon Ghosh, Young-Woong Ko, and Jeong-Gun Lee. Domain generalization for semantic segmentation: a survey. *Artificial Intelligence Review*, 57(9):247, 2024. 1

[45] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for Data: Ground Truth from Computer Games. In *Proc. of ECCV*, pages 102–118, Amsterdam, Netherlands, 2016. Springer. 7, 8

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. of CVPR*, pages 10684–10695, 2022. 3

[47] 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 *Proc. of CVPR*, pages 3234–3243, Las Vegas, NV, USA, 2016. 7, 8

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[49] Manuel Schwonberg, Fadoua El Bouazati, Nico M Schmidt, and Hanno Gottschalk. Augmentation-Based Domain Generalization for Semantic Segmentation. In *Proc. of IV - Workshops*, pages 1–8, 2023. 2

[50] Manuel Schwonberg, Joshua Niemeijer, Jan-Aike Termöhlen, Nico M Schmidt, Hanno Gottschalk, Tim Fingscheidt, et al. Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving. *IEEE Access*, 11:54296–54336, 2023. 1, 6

[51] Qiyu Sun, Huilin Chen, Meng Zheng, Ziyan Wu, Michael Felsberg, and Yang Tang. Ibaformer: Intra-batch attention transformer for domain generalized semantic segmentation. *arXiv preprint arXiv:2309.06282*, pages 1–10, 2023. 3, 4, 8

[52] Qiyu Sun, Pavlo Melnyk, Michael Felsberg, and Yang Tang. Augment Features Beyond Color for Domain Generalized Segmentation. *arXiv:2307.01703*, pages 1–10, 2023. 2, 3, 5, 6, 8

[53] Jan-Aike Termöhlen, Timo Bartels, and Tim Fingscheidt. A Re-Parameterized Vision Transformer (ReVT) for Domain-Generalized Semantic Segmentation. In *Proc. of ICCV - Workshops*, pages 1–10, Paris, France, 2023. 3, 5, 8

[54] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to Adapt Structured Output Space for Semantic Segmentation. In *Proc. of CVPR*, pages 7472–7481, Salt Lake City, UT, USA, 2018. 1

[55] Sumanth Udupa, Prajwal Gurunath, Aniruddh Sikdar, and Suresh Sundaram. Mrfp: Learning generalizable semantic segmentation from sim-2-real with multi-resolution feature perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5904–5914, 2024. 3, 5, 8

[56] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022. 1

[57] Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *Proc. of CVPR*, pages 28619–28630, 2024. 3, 6, 7, 8

[58] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model Soups: Averaging Weights of Multiple Fine-Tuned Models Improves Accuracy Without Increasing Inference Time. In *Proc. of ICML*, pages 23965–23998, Baltimore, MD, USA, 2022. PMLR. 5

[59] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Lili Ju, and Song Wang. Siamdoge: Domain generalizable semantic segmentation using siamese network. In *European Conference on Computer Vision*, pages 603–620. Springer, 2022. 3, 5, 6, 8

[60] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. SePiCo: Semantic-Guided Pixel Contrast for Domain Adaptive Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(7), 2022. 1

[61] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Proc. of NeurIPS*, pages 12077–12090, virtual, 2021. 7

[62] 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 *Proc. of AAAI*, pages 2884–2892, 2022. 3, 4, 8

[63] Liwei Yang, Xiang Gu, and Jian Sun. Generalized semantic segmentation by self-supervised source domain projection and multi-level contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10789–10797, 2023. 2, 3, 6, 8

[64] Jingjun Yi, Qi Bi, Hao Zheng, Haolan Zhan, Wei Ji, Yawen Huang, Yuexiang Li, and Yefeng Zheng. Learning spectral-decomposited tokens for domain generalized semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8159–8168, 2024. 3, 6, 7, 8

[65] 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 *Proc. of CVPR*, pages 1–14, Seattle, WA, USA, 2020. 7, 8

[66] 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 *Proc. of ICCV*, pages 2100–2110, Seoul, Korea, 2019. 2, 3, 5, 8

[67] Seokju Yun, Seunghye Chae, Dongheon Lee, and Youngmin Ro. Sora: Singular value decomposed low-rank adaptation for domain generalizable representation learning. *arXiv preprint arXiv:2412.04077*, 2024. 3, 7, 8

[68] Xin Zhang and Robby T Tan. Mamba as a bridge: Where vision foundation models meet vision language models for domain-generalized semantic segmentation. *arXiv preprint arXiv:2504.03193*, 2025. 3, 7, 8

[69] Yuhang Zhang, Shishun Tian, Muxin Liao, Guoguang Hua, Wenbin Zou, and Chen Xu. Learning shape-invariant representation for generalizable semantic segmentation. *IEEE Transactions on Image Processing*, 32:5031–5045, 2023. 2, 3, 8

[70] Yuhang Zhang, Shishun Tian, Muxin Liao, Zhengyu Zhang, Wenbin Zou, and Chen Xu. Fine-grained self-supervision for generalizable semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1):371–383, 2023. 3, 5, 8

[71] Dong Zhao, Jinlong Li, Shuang Wang, Mengyao Wu, Qi Zang, Nicu Sebe, and Zhun Zhong. Fishertune: Fisher-guided robust tuning of vision foundation models for domain generalized segmentation. *arXiv preprint arXiv:2503.17940*, 2025. 3, 7, 8

[72] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-Hallucinated Dual Consistency Learning for Domain Generalized Semantic Segmentation. In *Proc. of ECCV*, pages 535–552, 2022. 3, 4, 5, 6, 8

[73] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial Style Augmentation for Domain Generalized

Urban-Scene Segmentation. In *Proc. of NeurIPS*, pages 338–350, 2022. 2, 3, 8

[74] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022. 1

[75] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward Multimodal Image-to-Image Translation. In *Proc. of NeurIPS*, pages 465–476, Long Beach, CA, USA, 2017. 5