Exploiting Domain Properties in Language-Driven Domain Generalization for Semantic Segmentation

Seogkyu Jeon[1] Kibeom Hong[2*] Hyeran Byun[1*]
[1]Yonsei University [2]Sookmyung Women's University
jone9312@yonsei.ac.kr

Abstract

Recent domain generalized semantic segmentation (DGSS) studies have achieved notable improvements by distilling semantic knowledge from Vision-Language Models (VLMs). However, they overlook the semantic misalignment between visual and textual contexts, which arises due to the rigidity of a fixed context prompt learned on a single source domain. To this end, we present a novel domain generalization framework for semantic segmentation, namely Domain-aware Prompt-driven Masked Transformer (DPMFormer). Firstly, we introduce domain-aware prompt learning to facilitate semantic alignment between visual and textual cues. To capture various domain-specific properties with a single source dataset, we propose domain-aware contrastive learning along with the texture perturbation that diversifies the observable domains. Lastly, to establish a framework resilient against diverse environmental changes, we have proposed the domain-robust consistency learning which guides the model to minimize discrepancies of prediction from original and the augmented images. Through experiments and analyses, we demonstrate the superiority of the proposed framework, which establishes a new state-of-the-art on various DGSS benchmarks.

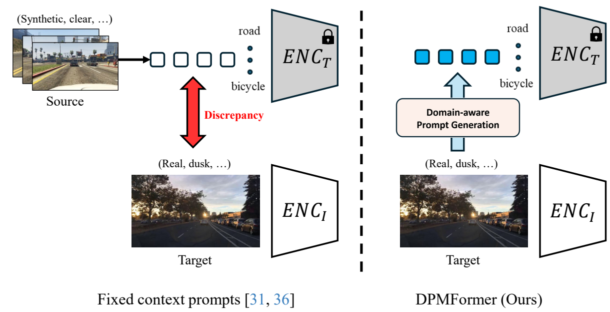


Figure 1. Motivation of DPMFormer. Using a fixed context prompts [31, 36] tend to retain source domain properties, causing contextual misalignment with the target domain. On the other hand, DPMFormer translates domain properties of the input image into context prompts, enhancing semantic alignments.

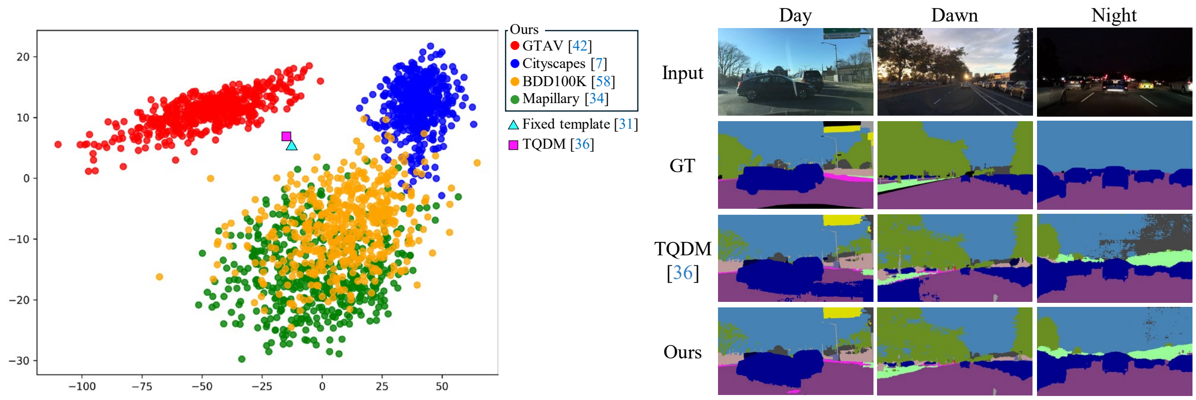
1. Introduction

Over the decades, semantic segmentation has made remarkable progress, now being able to precisely classify each pixel in an image into categories. However, one of the shadows that lies in these advancements is that models often exhibit inconsistent and degraded performance when deployed in various real-world environments. To this end, the task of Domain Generalized Semantic Segmentation (DGSS) has arisen to overcome the discrepancy between the training and test domains, *i.e.*, domain shift. DGSS has progressed through diverse approaches, such as feature whitening [6, 37–39] and domain randomization [4, 14, 19, 24, 27, 28, 56, 59, 61, 62]. Furthermore, some studies [3, 10] exploited multi-scale object queries with a transformer-based architecture, *i.e.*, Mask2Former [5].

Despite these advancements, learning domain-robust representations solely from the single source domain remains a significant hurdle to performance improvement. Recently, several studies [13, 21, 36, 55] employed Vision-Language Models (VLMs) [41, 45] owing to their semantic knowledge learned from diverse large-scale text-image datasets. For this, pioneering VLM-based DGSS works [13, 21, 55] have adopted their pre-trained visual encoder for initialization and fine-tuning [21, 55]. In addition, TQDM [36] has introduced a framework that constructs object queries from the textual descriptions, leveraging semantic concepts from linguistic expressions.

Yet, they still have limitations in fully utilizing textual knowledge to improve domain generalizability. Concretely, the context prompts that decorates textual descriptions of each category are either a predefined template [31] (e.g., '*a photo of*') or a single learnable text embedding [36]. However, we contend that fixed context prompts have limited capability for target domain images due to the discrepancy

[*]Corresponding author



(a) Visualization of textual prompts

(b) Results on BDD100K [58]

Figure 2. PCA visualization of textual prompts (left) and qualitative results on various environments (*e.g.*, Day, Dawn, Night) in BDD100K [58] (right). The models are trained on GTAV [42] with the CLIP-pretrained backbone (ViT-B) [41]. A fixed single-context prompt lacks flexibility in adapting to various domain shifts due to its rigidity. In contrast, our framework utilizes domain-specific properties from input images as context prompts, enhancing semantic alignment between text and images. As a result, as shown in (b), our approach exhibits improved robustness across diverse environments.

between visual and textual contexts as depicted in Fig. 1. Firstly, a handcrafted template inherently encodes the characteristics of a specific domain which restricts its generalizability to other unseen domains. Secondly, prompt optimization is prone to source domain overfitting, especially in the single-source setting. Consequently, both approaches may enlarge the gap between textual and visual contexts, leading to suboptimal performances on target domains.

In this perspective, we point out that the semantic form of the textual representation for an object category should be changed with respect to the input visual context in order to strengthen their semantic correspondence. For instance, in the real-world driving scene at night of Fig. 2 (b), target categories (*e.g.*, *car* and *sky*) possess distinct textures from those of synthetic daytime images. Hence, it would be more appropriate to decorate *'a car'* with the textual prompt *'at night in the real-world'* which is reflecting the domain characteristics. As a result, the modified text feature will include details such as darkened car exteriors and light reflections, enhancing the semantic alignment. Nevertheless, it is challenging to design domain-adaptive prompts in the DGSS setting where the training dataset covers only a single domain. Moreover, this setting hinders the visual pipeline from learning robust image features against domain shifts.

To address these challenges, we introduce a novel framework, namely Domain-aware Prompt-driven Masked Transformer (DPMFormer), focusing on two key aspects: (1) Leveraging domain-specific properties of the input image (domain-awareness) and (2) Generating accurate outputs on images with dissimilar domain characteristics (domain-robustness). To cultivate the domain-awareness, we propose a novel **domain-aware prompt learning**, which translates the domain-specific properties of an input image into context prompts via an auxiliary network. In addition, to obtain diverse domain properties in the single-source setting, we apply texture perturbations to synthesize novel domain images. Furthermore, exploiting both source and novel domain images, we propose a **domain-aware contrastive loss** to ensure that the derived prompts effectively capture domain-specific properties of the input image. This loss encourages the anchor context prompts to be distinguishable from those of different domain characteristics, while being closer to those from the anchor domain. With the proposed domain-aware prompt, the model accurately identifies target classes while being aware of the input domain.

Moreover, we strive to provide better domain robustness guidances to the visual encoder and decoders. We carefully organize the texture perturbations with structure-preserving image transformations, ensuring that the original visual context remains intact. Instead of simply reusing the original ground truths for novel domain images, we introduce **domain-robust consistency loss** to guarantee reliable predictions under severe domain shifts. The loss consists of class consistency and mask consistency losses, which penalize discrepancies between class and mask predictions of given image pairs, respectively. Furthermore, domain-robust consistency losses are applied at every layer of the transformer decoder, preventing discrepancies in earlier layers from propagating to later parts.

Through evaluations on various DGSS benchmarks, we demonstrate the superiority of the proposed framework. Notably, our framework achieves state-of-the-art semantic segmentation performance across domain generalization scenarios. Additionally, ablation studies and detailed analyses validate the effectiveness of each component.

2. Related Works

2.1. Domain Generalized Semantic Segmentation

Domain Generalized Semantic Segmentation (DGSS) aims to learn domain-invariant representations that generalize robustly to various unseen target domains. The task assumes a single-source setting, where only one dataset is available during training. Unlike domain adaptation [8, 49, 66] and test-time domain adaptation [53, 54, 57], access to target domains is strictly prohibited, making DGSS more challenging. Previous studies approached DGSS in two ways primarily: feature whitening and normalization [6, 37–39], and domain randomization approaches [4, 14, 19, 24, 27, 28, 37, 38, 56, 59, 61, 62].

Feature whitening and normalization approaches [6, 37–39] mainly focus on removing features which are variant to domain shifts. Exploiting the characteristics of instance whitening [30] and instance normalization [50] which can effectively remove texture and style from the input image, these approach leverages those operations in between the backbone module to minimize the effect from domain and texture changes. Representatively, RobustNet [6] introduced instance selective whitening module that finds gram matrix components sensitive to photometric changes and minimize their changes. However, due to the natural difficulty in disentangling domain-specific and domain-invariant features, these approaches show limited performance gains.

Domain randomization approaches [4, 14, 19, 24, 27, 28, 56, 59, 61, 62] augment novel domains from the source domain by modifying either the images or their features through various methods, *e.g.* affine transformations [14, 20, 28, 56, 61, 62], image translation [59, 65], frequency decomposition [4, 19],and photometric transformations [24]. Synthesized samples increase the domain diversity of the training dataset, reducing the domain gap between the learned representation and the test data. Moreover, most of the generative DGSS methods adopt context-preserving transformations to compute the output discrepancy between the original sample and the augmented one. Representatively, SHADE [61] proposed a style consistency loss to encourage model to learn invariant pixel-level semantic information by minimizing the Jenson-Shannon Divergence (JSD) between the output predictions of original and augmented images.

Meanwhile, several studies [3, 10] have built DGSS upon Mask2Former [5], that leverages attention mechanism [51] renowned for robustness against domain shifts [16, 17, 46]. Mask2former [5] involves a transformer decoder that exploits object query features to group pixels of same objects or categories. Based on this, HGFormer [10] first proposed a hierarchical framework that groups pixels to form part-level masks for complementing whole-level pixel grouping procedure. Similarly, CMFormer [3] utilizes down-sampled features additionally via feature fusion which are more domain-invariant than the original features.

2.2. Language-driven Domain Generalized Semantic Segmentation

Recent studies [13, 21, 22, 36, 55] has investigated to exploit Vision Language Models (VLMs) [9, 23, 33, 40, 41, 45, 47] for DGSS, owing to its powerful generalizability learned from large-scale datasets of image-text pairs. VLTseg [21] employs the image encoder of VLMs for initializing backbone parameters, and fine-tune the network with the segmentation objective function. FAMix [13] proposes to yield class-specific novel styles by concatenating random style description and class names, then performs style randomization by locally mixing the source and the novel styles. DAP [22] exploits text encoder to distill the semantic textual knowledge of target categories to the visual backbone. TQDM [36] points out deficient use of language information in aforementioned works, and introduces a textual-query driven framework for DGSS. They utilize the textual description of each categories for initializing input query features for transformer decoder, and also for computing the text-to-pixel attention to enhance semantic clarity of pixel features. Despite their advancements, we argue that two points are overlooked : (1) the domain discrepancy between the learned textual prompt and the domain-specific property of the input image (2) lack of domain robustness guidance. To mitigate these limitations, we carefully refine the textual queries with the captured domain-specific property of the input image, and augment novel domain images to improve domain-awareness of the model as well as prediction consistency learning.

3. Methods

3.1. Preliminaries

We design our DPMFormer based on mask classification architecture, *i.e.*, Mask2Former [5]. Mask2Former consists of an image encoder ENC_I, pixel decoder DEC_{pix}, and transformer decoder DEC_{tr}. The input RGB image $x \in \mathbb{R}^{3\times H\times W}$ is first fed to the image encoder for feature extraction, and then converted to pixel-wise features $z \in \mathbb{R}^{D\times H\times W}$ through the pixel decoder. The transformer decoder iteratively refines N object queries $q \in \mathbb{R}^{N\times L}$ with multi-scale image features from the pixel decoder. Thereafter, refined object queries are projected into mask embeddings $m \in \mathbb{R}^{N\times D}$ to generate pixel-level mask prediction via dot product as $\hat{y}^{\text{mask}} = m \cdot z, \hat{y}^{\text{mask}} \in \mathbb{R}^{N\times H\times W}$. The category label of each object query q is predicted with a linear classifier, *i.e.*, $c_q \in \mathbb{R}^{K}, c \in \mathbb{R}^{N\times K}$. The final prediction result is derived by the matrix multiplication between the pixel-level mask prediction and class prediction

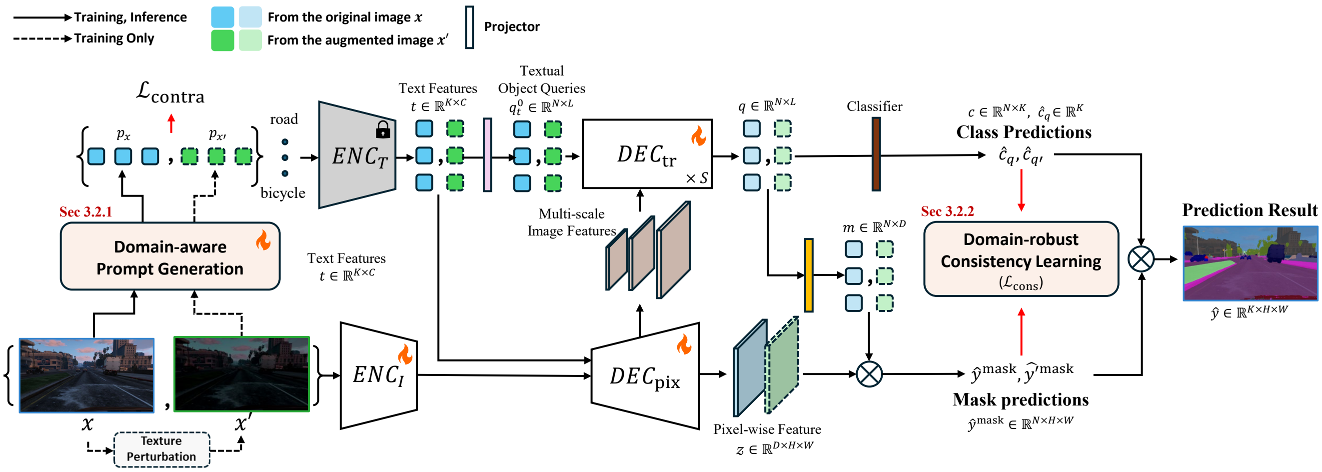


Figure 3. Illustration of DPMFormer. We use Mask2Former [5] based architecture which consists of a backbone image encoder (ENC_I), a pixel decoder (DEC_{pix}), a transformer decoder (DEC_{tr}), and a text encoder (ENC_T). During training, we synthesize images with a novel domain style via texture perturbation. Both images are incorporated to compose a batch and exploited for learning domain-awareness (Sec. 3.2.1) and domain-robustness (Sec. 3.2.2).

of queries as $\hat{y} = c^\top \cdot \hat{y}^{\text{mask}}, \hat{y} \in \mathbb{R}^{K \times H \times W}$.

Furthermore, inspired by [36], we employ a text encoder ENC_T for initializing object queries from textual descriptions. Pre-trained vision language mdoels (VLMs), *e.g.*, CLIP [41], is exploited to initialize both ENC_I and ENC_T parameters, and the parameters of ENC_T remain frozen during training to preserve learned linguistic knowledge. The text encoder ENC_T produces a text feature $t_k \in \mathbb{R}^C$ corresponding to each class label class_k with a learnable context prompt p, *i.e.*, $t_k = ENC_T([p, \{\text{class}_k\}])$. Then, obtained text embeddings $t = \{t_k\}_{k=1}^K \in \mathbb{R}^{K \times C}$ are passed to a multi-layer perceptron (MLP) to derive initial textual object queries q_t^0. Also, the text-to-pixel attention mechanism is established in a pixel decoder layer to enhance pixel semantic clarity enhancement.

3.2. DPMFormer

As illustrated in Fig. 3, Domain-aware Prompt-driven Masked Transformer (DPMFormer), aims to cultivate two core aspects of domain generalization for the model: domain-awareness and domain-robustness. For domain-awareness, we propose the domain-aware context prompt learning which translates domain-specific property of the input image into the text embedding for semantic alignment between textual and visual features. And for domain-robustness, we encourage the model to generate accurate outputs against textural changes via consistency learning.

Texture perturbation. In order to effectively guide the model with diverse domain characteristics, we stylize the training dataset to generate an auxiliary domain dataset. Following RobustNet [6], we adopt photometric transformations—comprising strong color jittering, gaussian blur and noise injection—for their simplicity as well as content-preserving property. The generated image x' is combined with its original image x to form a batch for training.

3.2.1. Domain-aware context prompt learning

Textual query generation and prompt learning [64] plays a crucial role in enabling the model to leverage the linguistic semantics of each class previously learned by the VLM. Although the learned context prompt is beneficial for performance improvement, it is optimized solely within the source domain without direct consideration of domain shifts. This single context prompt may yield strong results on several domain shift scenarios where target domain has similar domain characteristics with the source. However, in cases of severe domain shift, the performance gain may be limited due to the contextual mismatch between the target domain image and the learned prompt. For example, if the prompt learning is conducted on a dataset containing only sunny day images, the learned prompt is could be misaligned when encountering rainy night scenes. As such, the semantic misalignment should be addressed in order to fully utilize rich semantic knowledge of pre-trained VLM. To this end, we propose to generate domain-aware context prompt that extracts domain-specific properties from the input image as a context prompt.

To obtain domain-aware context prompt from the input image x, we design an auxiliary network $h_\theta(\cdot)$ that takes visual feature as an input and generates a domain-specific prompt embedding $\pi_x = h_\theta(\hat{F}(x))$, where $\hat{F}(x)$ denotes a visual feature extracted from a frozen visual backbone of CLIP [41]. We use the class token as a visual feature $\hat{F}(x)$ for its global representation [11]. Nextly, the obtained domain-specific prompt embedding π_x is integrated with the context prompt embedding p through addition, *i.e.*, $p_x = p + \pi_x$, then concatenated with text embeddings of classes to generate domain-aware textual features as $t_{x,k} = ENC_T([p_x, \{\text{class}_k\}])$. To encourage derived textual features to include domain-specific information of the input image, we introduce a novel domain-aware contrastive learning framework with the original and augmented images. The loss function is depicted as follows:

$$\mathcal{L}_{contra} = -\frac{1}{2B}\sum_{i=1}^{2B}\log\frac{\sum_{j\in\mathcal{P}_i}\exp\text{sim}(\pi_i,\pi_j)/\tau}{\sum_{j\in\mathcal{P}_i\cup\mathcal{N}_i}\exp\text{sim}(\pi_i,\pi_j)/\tau}, \tag{1}$$

where $\text{sim}(\cdot)$ means similarity metric and τ is a temperature parameter. B denotes the batch size of original images, $\mathcal{P}_i$ and $\mathcal{N}_i$ are positive sets and negative sets of the i-th image, respectively. The positive set $\mathcal{P}_i$ is composed of the indices of samples having same domain characteristics with the anchor i whereas the others belong to negative set $\mathcal{N}_i$. For example, when an anchor i is of original source domain images, other original source domain images belongs to $\mathcal{P}_i$ while all augmented images are included in $\mathcal{N}_i$. In case of an augmented image as an anchor, $\mathcal{P}_i = \{i\}$ and $\mathcal{N}_i = \{1,\ldots,2B\}\setminus i$. With the proposed loss, we encourage h_θ to capture domain-specific property of the image and reflect it to the output text feature t_k. Also, the final domain-aware text feature is guided by the task loss $\mathcal{L}_{seg}$ to be aligned with the original image and to minimize segmentation errors. We note that the proposed loss considers the domain information unlike CoCoOp [63] which treats $\mathcal{P}_i$ as the text feature corresponding to a specific image feature while the remaining images forming the negative set. We provide comparative analysis with CoCoOp in Sec.4.4.2.

3.2.2. Domain-robust consistency learning

To further enhance the domain robustness, we encourage model to generate persistent predictions in the domain shift scenario. To enhance prediction consistency, we induce the model to minimize the prediction discrepancy in terms of the mask and class label as follows:

$$\mathcal{L}_{\text{cons}} = \sum_{s=1}^{S}\lambda_{\text{mc}}\cdot\mathcal{L}_{\text{mc}}(\hat{y}_{\text{s}}^{\text{mask}}, \hat{y'}_{\text{s}}^{\text{mask}}) + \lambda_{\text{cc}}\cdot\mathcal{L}_{\text{cc}}(\hat{c}_{q_i,\text{s}}, \hat{c}_{q'_i,\text{s}}). \tag{2}$$

S denotes the number of transformer blocks in the transformer decoder, and $\mathcal{L}_{\text{mc}}$ and $\mathcal{L}_{\text{cc}}$ represent the mask and class consistency losses, respectively. To compute the loss at the s-th transformer decoder block, we obtain $\{\hat{y}_{\text{s}}^{\text{mask}}, \hat{y'}_{\text{s}}^{\text{mask}}\}$ and $\{\hat{c}_{q_i,\text{s}}, \hat{c}_{q'_i,\text{s}}\}$ which are pairs of mask predictions and class predictions of i-th query q_i from the original and augmented image pair $\{x, x'\}$, respectively. We employ binary cross entropy and Jensen–Shannon divergence as a discrepancy measure of the mask consistency loss ($\mathcal{L}_{mc}$) and class consistency loss ($\mathcal{L}_{cc}$), respectively. With the help of domain-aware context prompts and above losses, our model learns to predict not only accurately but also consistently in various domain shift scenarios during training.

3.3. Overall Loss Functions

The overall loss of our framework is a weighted sum of the task loss $\mathcal{L}_{\text{seg}}$, VLM regularization loss $\mathcal{L}_{\text{reg}}$, domain-aware contrastive loss $\mathcal{L}_{\text{contra}}$, and the consistency loss $\mathcal{L}_{\text{cons}}$.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{contra}}\mathcal{L}_{\text{contra}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}, \tag{3}$$

where $\{\lambda_{\text{reg}}, \lambda_{\text{contra}}, \lambda_{\text{cons}}\}$ are constant weighting factors. We note that $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{cons}}$ are calculated with all queries and its predictions from every block of the transformer decoder. We provide details of the baseline losses ($\mathcal{L}_{\text{seg}}$, $\mathcal{L}_{\text{reg}}$) in the supplementary.

4. Experiments

4.1. Implementation Details

Datasets. We validate DPMFormer on *synthetic-to-real* and *real-to-real* scenarios in the single-source setting.

Synthetic datasets. GTAV [42] is a representative synthetic dataset which consists of 24,966 images with a resolution of 1914×1052. The training split contains 12,403 images, while the validation and test set are of 6,382 and 6,181 images, respectively. SYNTHIA [43] dataset provides 6,580 images for training and 2,820 images for validation respectively, with the image resolution at 1280×760.

Real-world datasets. Cityscapes [7] is a dataset collected from the real environment. The resolution of each image is 2048×1024, and the population of training and validation split is 2,975 and 500, respectively. BDD100K [58] includes 7,000 training images and 1,000 validation images with the resolution of 1280×720. Mapillary [34] is composed of images with diverse resolution, where the training and validation size is 18,000 and 2,000, respectively.

Network architecture. Our approach leverages the vision transformer-based models as backbones, initialized with either the CLIP [41] (ViT-B) or EVA02-CLIP [45] (EVA02-L) model. The CLIP backbone is configured with a patch size of 16, while the EVA02-CLIP backbone uses a patch size of 14. For the pixel and transformer decoder, aforementioned, we employ a mask classification architecture [5], consisting of $N = 9$ layers with masked attention mechanisms. Additionally, we design the auxiliary network h_θ for domain-aware context prompt generation as a shallow multi-layer structure (BatchNorm-Linear-ReLU-Linear).

Training and evaluation. To optimize DPMFormer, we employ an AdamW [32] where the learning rate is set as 1×10^{-5} and 1×10^{-4} for synthetic and real training datasets, respectively. Following previous transformer-based studies [18, 36], we apply linear warm-up [15] for

Models (GTAV)	Backbone	Cityscapes	BDD	Mapillary	Avg.
SAN-SAW [39]	ResNet-101	45.33	41.18	40.77	42.43
WildNet [28]	ResNet-101	45.79	41.73	47.08	44.87
SHADE [61]	ResNet-101	46.66	43.66	45.50	45.27
TLDR [27]	ResNet-101	47.58	44.88	48.80	47.09
FAMix* [13]	ResNet-101	49.47	46.40	51.97	49.28
SHADE [61]	MiT-B5	53.27	48.19	54.99	52.15
IBAFormer [44]	MiT-B5	56.34	49.76	58.26	54.79
VLTSeg* [21]	ViT-B	47.50	45.70	54.30	49.17
TQDM* [36]	ViT-B	<u>57.50</u>	<u>47.66</u>	<u>59.76</u>	<u>54.97</u>
DPMFormer* (ours)	ViT-B	**59.00**	**51.80**	**63.62**	**58.14**
VLTSeg** [21]	EVA02-L	65.60	58.40	66.50	63.50
Rein** [55]	EVA02-L	65.30	**60.50**	64.90	63.60
Rein†	ViT-L	66.40	60.40	66.10	64.30
TQDM** [36]	EVA02-L	<u>68.88</u>	59.18	<u>70.10</u>	<u>66.05</u>
DPMFormer** (ours)	EVA02-L	**70.08**	<u>60.48</u>	**70.66**	**67.07**

Table 1. Comparison with the state-of-the-art DGSS methods on *synthetic-to-real* scenario with GTAV [42] as a source. We note that *, **, and † denotes the models initialized with pretrained CLIP [41], EVA02-CLIP [45], and DINOv2 [35], respectively. The best and the second best performances are highlighted with **bold** and <u>underline</u>, respectively.

Models (Synthia)	Backbone	Cityscapes	BDD	Mapillary	Avg.
SAN-SAW [39]	ResNet-101	40.87	35.98	37.26	38.04
TLDR [27]	ResNet-101	42.60	35.46	37.46	38.51
IBAFormer [44]	MiT-B5	50.92	44.66	50.58	48.72
VLTSeg** [21]	EVA02-L	56.80	50.50	54.50	53.93
TQDM** [36]	EVA02-L	<u>57.99</u>	<u>52.43</u>	<u>54.87</u>	<u>55.10</u>
DPMFormer** (ours)	EVA02-L	**58.92**	**54.39**	**60.08**	**57.80**

Table 2. Comparison with the state-of-the-art DGSS methods where SYNTHIA [43] is set as a source dataset.

initial 1,500 iterations and rare class sampling [18]. The optimizer settings are identical for both CLIP and EVA02-CLIP backbone models. We set the total training iterations and the batch size as 20,000 and 8, respectively. Weighting factors $\{\lambda_{\text{reg}}, \lambda_{\text{contra}}, \lambda_{\text{cons}}\}$ are set as 1, 1, 10. We crop the input image to have a resolution of 512×512. The texture perturbation is only applied during training. We use mean Intersection over Union (mIoU) [12] to quantitatively evaluate the results following the convention.

4.2. Quantitative Results

Synthetic-to-real. As shown in Tab. 1, we achieve state-of-the-art in every target domain with both backbones. Notably, with the ViT-B backbone initialized with pretrained CLIP [41], DPMFormer consistently surpasses previous state-of-the-art [36] by 3.17% on average mIoU among target domains. In detail, on Cityscapes which contains mostly daytime real images, the domain robustness empowered by consistency learning assists DPMFormer to cope with the domain gap caused by visual realism, improving the state-of-the-art performance by 1.5%. Meanwhile on BDD100K and Mapillary which have higher environmental variation in terms of weather, time and location, the domain-aware prompt generation enables the model to adaptively leverage

Models (Cityscapes)	Backbone	BDD	Mapillary	Avg.
SAN-SAW [39]	ResNet-101	54.73	61.27	42.43
WildNet [28]	ResNet-101	47.01	41.73	44.87
SHADE [61]	ResNet-101	50.95	43.66	45.27
TQDM* [36]	ViT-B	50.54	65.74	58.14
DPMFormer* (ours)	ViT-B	**54.81**	**67.72**	**61.27**
HGFormer [10]	EVA02-L	61.50	72.10	66.80
VLTSeg** [21]	EVA02-L	64.40	<u>76.40</u>	70.40
Rein** [55]	EVA02-L	64.10	69.50	66.80
Rein†	ViT-L	**65.00**	72.30	68.65
TQDM [36]**	EVA02-L	<u>64.72</u>	76.15	<u>70.44</u>
DPMFormer** (ours)	EVA02-L	64.2	**76.67**	**70.44**

Table 3. Comparison with the state-of-the-arts trained with Cityscapes [7] on the *real-to-real* scenario.

the textual knowledge for segmentation, impressively escalating the performance by 4.14%, and 3.17%, respectively. Moreover, even with the larger backbone [45], we mark the highest average performance of 67.07%. In Cityscapes and Mapillary dataset, we outperform TQDM [36] by 1.2% and 0.56% respectively, while nearly reaching the score of Rein [55] in BDD. The overall results demonstrate the superiority of DPMFormer, emphasizing the efficacy of domain-awareness as well as domain-robustness.

In Tab. 2, we also present synthetic-to-real results where SYNTHIA [43] is set as a source domain. Again, DPMFormer outperforms all competitors by a large margin, achieving the state-of-the-art performance in every target domain. Remarkably, our framework surpasses the previous state-of-the-art [36] by an average of 2.7% mIoU across target domains. In particular, the performance on Mapillary [34] improves significantly by 5.21%. These results demonstrate the effectiveness of DPMFormer in addressing the domain gap in terms of texture and perspective.

Real-to-real. Furthermore, as shown in Tab.3, DPMFormer records the highest average mIoU with both backbones. With the CLIP-pretrained backbone, we significantly boost the state-of-the-art performance by 4.27% and 1.98% on BDD and Mapillary, respectively. These results indicate that enhanced domain robustness supports the model in predicting consistently under environmental changes, while the domain-aware prompts facilitate the semantic alignment on the unseen domain. As a result, DPMFormer successfully achieves an average performance gain of 3.13% over the previous state-of-the-art. Equipped with the EVA-CLIP [45] pretrained backbone, we attain the highest mIoU of 76.67% on Mapillary, while performing slightly lower on BDD. Overall, the average performance is 70.44%, which is comparable to TQDM [36].

4.3. Qualitative Results

In Fig. 4, we qualitatively compare our method with FAMix [13] and TQDM [36] on the *synthetic-to-real* sce-

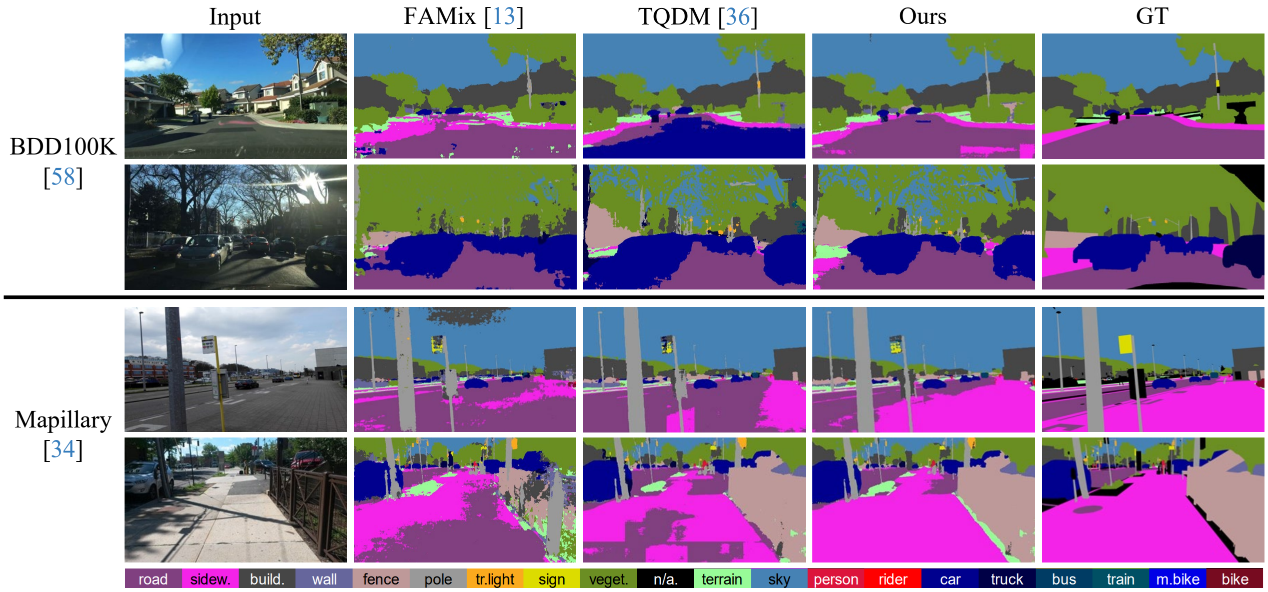


Figure 4. Qualitative comparison on synthetic-to-real scenario with the CLIP-pretrained backbone (ViT-B). The training source domain is set as GTAV [42] while the target domains are BDD100K [58] and Mapillary [34]. The overall result shows that DPMFormer accomplishes precise segmentation with the images of strong illumination contrast as well as confusing textures.

nario with the pretrained CLIP backbone (ViT-B). As depicted in the results on BDD100K, both competitors struggle to discriminate accurately under environments having confusing textures or large variations of illumination. In particular, the road in the first-row image has a texture similar to the exterior of a car, leading TQDM to a mismatch between the visual and textual features. In addition, FAMix shows sensitivity to the shades, mislabeling the road as a sidewalk. On the contrary, our method effectively distinguishes the road from other categories by reflecting the domain-specific properties of real-world roads in clear weather. In the case of the second row image with the high illumination contrast, both FAMix and TQDM fail to notice sidewalks in the image due to their low brightness. On the other hand, DPMFormer carries out precise predictions under severe photometric changes owing to the domain robustness acquired from handling diverse texture changes.

As observed with the samples from the Mapillary (the third and fourth rows), FAMix and TQDM lack discriminability between the road and the sidewalk. To be specific with the first image (third row), these classes appear to have a similar color and texture, resulting in misclassifications. Meanwhile, the domain-specific properties from the image transfer the semantic knowledge of the appearance of real-world sidewalk to the model, DPMFormer accurately distinguish between the road and the sidewalk. With the second image, TQDM [36] shows vulnerability against small textural changes on the sidewalk (*e.g.*, shades of the poles). Contrarily, DPMformer effectively performs segmentation owing to its domain-robustness as well as the obtained domain-aware textual queries.

4.4. Analysis

4.4.1. Ablation studies

To demonstrate the effectiveness of the components in DPMFormer, *i.e.*, texture perturbation, domain-robust consistency learning ($\mathcal{L}_{cons}$), and domain-aware context prompt learning ($\mathcal{L}_{contra}$), we conduct ablation studies with the CLIP pretrained backbone on *synthetic-to-real* scenario. As presented in Tab. 4, every component contributes to the performance gain. Specifically, the texture perturbation enlarges the observable domain during training, boosts the average performance by 0.65%. In addition with the $\mathcal{L}_{\text{cons}}$, the model successfully equips the robustness against domain-shift and enhances the prediction accuracy on unseen target domains, increasing the average mIoU by 0.77%. Furthermore, $\mathcal{L}_{\text{contra}}$ allows the model to acquire domain-specific property from the input image in the form of textual prompt embedding. Consequently, the model exploits both visual and textual cues properly aligned for the target domain, considerably elevate the average mIoU by 1.99%. Combining all components together, DPMFormer effectively learns both domain-awareness and domain-robustness to generate accurate results consistently on unseen target domains.

4.4.2. Comparison with prompt learning methods

In Tab. 5, we compare our domain-aware context prompt learning ($\mathcal{L}_{\text{contra}}$) with previous prompt learning methods [25, 26, 63] to validate the efficacy for domain gen-

Models (GTAV)	Cityscapes	BDD	Mapillary	Avg.
Baseline	57.5	47.66	59.76	54.97
+ Perturbation	57.04	48.19	60.91	55.38
+ $\mathcal{L}_{\text{cons}}$	58.22	49.39	60.84	56.15
+ $\mathcal{L}_{\text{contra}}$	**59.0**	**51.8**	**63.62**	**58.14**

Table 4. Ablation study of proposed components. The models are trained on GTAV with CLIP pretrained backbone (ViT-B). The best performance for each column is highlighted in bold.

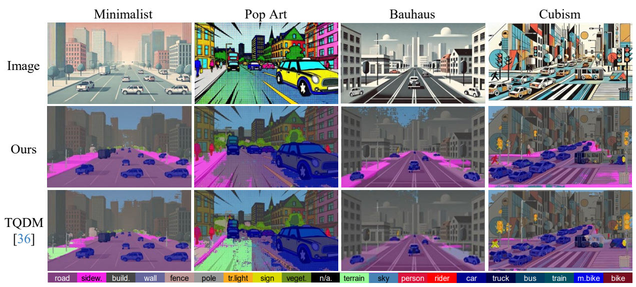


Figure 5. Qualitative results on diverse styles, *i.e.*, *Minimalist*, *Pop Art*, *Bauhaus*, and *Cubism*. The models are trained on GTAV with CLIP pretrained backbone (ViT-B).

eralization. Given the textual features $t = ENC_T(p_x)$ and visual features $v = \hat{F}(x)$, the contrastive loss of CoCoOp [63] encourages the similarity between the anchor text feature and its corresponding visual feature to be higher than others. As observed in the second row, CoCoOp marginally improves the performance of prompt generator h_θ, indicating that the generated feature become instance-specific which is less generalizable to other domains. Moreover, additionally assigning the visual feature from augmented sample as positives (CoCoOp$^+$) resulted in negligible performance gain, since the derived prompt contains domain-invariant information rather than domain-specific properties. Furthermore, although both MaPLe [25] and PromptSRC [26] employ multi-modal prompts, their performance gains are modest due to the limited generalizability of fixed prompts. On the other hand, our domain-aware contrastive loss design enable the model to proficiently capture domain-specific property from the image which are more helpful for the semantic alignment between the visual features and the textual knowledge. From the fifth to the eighth row, we verify $\mathcal{L}_{\text{contra}}$ with different similarity calculation targets for sim$(\cdot)$, *i.e.*, text features t, text and visual features (t, v), and (3) output context embeddings π. The results confirm that computing the domain-aware contrastive loss empowers the domain generalizability especially when computed with the context embeddings π which can provide direct domain guidance to h_θ.

$\mathcal{L}_{\text{contra}}$ (GTAV)	Cityscapes	BDD	Mapillary	Avg.
–	57.65	49.63	61.10	56.13
CoCoOp [63]	57.84	49.91	61.33	56.36
CoCoOp$^+$	57.60	50.04	60.96	56.20
MaPLe [25]	57.87	50.12	61.04	56.34
PromptSRC [26]	58.10	49.73	62.51	56.78
Ours (sim(t, t))	58.17	50.18	61.37	56.57
Ours (sim(t, v))	**59.49**	50.34	62.18	57.34
Ours (sim(π, π))	59.00	**51.80**	**63.62**	**58.14**

Table 5. Comparison with prompt learning methods [25, 26, 63] in synthetic-to-real scenario. – denotes DPMFormer without $\mathcal{L}_{\text{contra}}$. CoCoOp$^+$ indicates the modified loss that additionally include the feature from augmented sample of anchor as positives. t, v, and π refers to the text feature, visual feature, and context embedding.

4.4.3. Qualitative results on diverse styles

To further verify the domain generalization capability of DPMFormer, we compare our model with TQDM [36] on images with diverse artistic styles generated by ChatGPT[1]. As shown in Fig. 5, DPMFormer correctly predicts objects and their surroundings even in severe domain shift scenarios, *e.g.*, *Cubism*. With modern art styles (*i.e.*, *Minimalist*, *Pop Art*, and *Bauhaus*), TQDM perplexes among road, sidewalk, and terrain due to their textural changes. Notably, DPMFormer produces more reliable results owing to the enhanced robustness against texture variations that are learned from the texture perturbation and consistency learning. In case of the scene in cubism style, TQDM mispredicts the person on the left side because of its visual similarity with the person in the traffic sign. On the other hand, our method reflects the cubism style to the textual object queries, making an accurate classification with the same instance.

5. Conclusion

In this paper, we presented DPMFormer, a novel domain generalization framework for semantic segmentation. To address the limited generalizability of fixed context prompt learned from a single source dataset, we introduced a novel domain-aware prompt learning which can reflect domain-specific properties of the input image into textual prompts. The proposed component enhanced the semantic alignment between visual and textual cues, assisting the model to fully leverage the abundant semantic knowledge of VLMs. Moreover, to empower domain-robustness, we simulated various domain shifts via texture perturbations, and provided consistency guidance to the model by minimizing prediction discrepancies between original and augmented images. Through extensive experiments and analyses, we demonstrated the effectiveness of DPMFormer, achieving state-of-the-art performance on various benchmarks.

[1] https://chat.openai.com

Acknowledgments. This project was supported by the National Research Foundation of Korea grant funded by the Korea government (MSIT) (No. 2022R1A2B5B02001467; RS-2024-00346364). This research was supported by Sookmyung Women's University Research Grants (1-2403-2034).

References

[1] Shuanghao Bai, Yuedi Zhang, Wanqi Zhou, Zhirong Luan, and Badong Chen. Soft prompt generation for domain generalization. In *European Conference on Computer Vision*, pages 434–450. Springer, 2025. 2

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 1, 2

[3] 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. 1, 3

[4] Prithvijit Chattopadhyay, Kartik Sarangmath, Vivek Vijaykumar, and Judy Hoffman. Pasta: Proportional amplitude spectrum training augmentation for syn-to-real domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19288–19300, 2023. 1, 3

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1, 3, 4, 5

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11580–11590, 2021. 1, 3, 4

[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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5, 6, 1, 3, 7, 8

[8] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. *Domain adaptation in computer vision applications*, pages 1–35, 2017. 3

[9] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173, 2021. 3

[10] Jian Ding, Nan Xue, Gui-Song Xia, Bernt Schiele, and Dengxin Dai. Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15413–15423, 2023. 1, 3, 6

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 4

[12] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 6

[13] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. A simple recipe for language-guided domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23428–23437, 2024. 1, 3, 6

[14] 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 (ICLR 2023)*. OpenReview, 2023. 1, 3

[15] P Goyal. Accurate, large minibatch sg d: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 5

[16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 3, 1

[17] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 3

[18] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9924–9935, 2022. 5, 6

[19] 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*, pages 6891–6902, 2021. 1, 3

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 3

[21] Christoph Hümmer, Manuel Schwonberg, Liangwei Zhong, Hu Cao, Alois Knoll, and Hanno Gottschalk. Vltseg: Simple transfer of clip-based vision-language representations for domain generalized semantic segmentation. *arXiv preprint arXiv:2312.02021*, 2023. 1, 3, 6

[22] Xinyue Huo, Lingxi Xie, Hengtong Hu, Wengang Zhou, Houqiang Li, and Qi Tian. Domain-agnostic priors for semantic segmentation under unsupervised domain adaptation and domain generalization. *International Journal of Computer Vision*, 132(9):3954–3976, 2024. 3

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International*

conference on machine learning, pages 4904–4916. PMLR, 2021. 3

[24] 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. 1, 3

[25] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023. 7, 8

[26] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023. 7, 8

[27] Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. Texture learning domain randomization for domain generalized segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 677–687, 2023. 1, 3, 6

[28] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9936–9946, 2022. 1, 3, 6

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, broader and artier domain generalization. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5543–5551, 2017. 1

[30] 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

[31] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2023. 1

[32] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[34] 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. 5, 6, 7, 3, 8

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[36] Byeonghyun Pak, Byeongju Woo, Sunghwan Kim, Daehwan Kim, and Hoseong Kim. Textual query-driven mask transformer for domain generalized segmentation. *European Conference on Computer Vision*, 2024. 1, 3, 4, 5, 6, 7, 8, 2

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

[38] Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1863–1871, 2019. 3

[39] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2594–2605, 2022. 1, 3, 6

[40] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555: 126658, 2023. 3

[41] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 5, 6

[42] 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. 2, 5, 6, 7, 3, 4

[43] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 5, 6

[44] 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*, 2023. 6

[45] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1, 3, 5, 6

[46] C Szegedy. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 3

[47] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

[48] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 1

[49] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 3

[50] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6924–6932, 2017. 3

[51] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[52] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5018–5027, 2017. 1, 2

[53] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *The Nineth International Conference on Learning Representations (ICLR 2021)*. OpenReview, 2021. 3

[54] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 3

[55] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28619–28630, 2024. 1, 3, 6

[56] 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. 1, 3

[57] Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8978–8987, 2021. 3

[58] 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. 2, 5, 7, 3, 4, 6

[59] 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. 1, 3

[60] Xin Zhang, Shixiang Shane Gu, Yutaka Matsuo, and Yusuke Iwasawa. Domain prompt learning for efficiently adapting clip to unseen domains. *Transactions of the Japanese Society for Artificial Intelligence*, 38(6):B–MC2_1, 2023. 2

[61] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *European conference on computer vision*, pages 535–552. Springer, 2022. 1, 3, 6

[62] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. *Advances in neural information processing systems*, 35:338–350, 2022. 1, 3

[63] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 5, 7, 8, 2

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 4

[65] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3

[66] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 3