Prompting Diffusion Representations for Cross-Domain Semantic Segmentation

Rui Gong[1,3*]
gongr@vision.ee.ethz.ch
Martin Danelljan[1]
martin.danelljan@vision.ee.ethz.ch
Han Sun[2]
han.sun@epfl.ch
Julio Delgado Mangas[4]
anonymous@meta.com
Nikolay Marin[3]
anonymous@amazon.com
Luc Van Gool[1]
vangool@vision.ee.ethz.ch

[1] ETH Zurich
[2] EPFL
[3] Amazon
[4] Meta

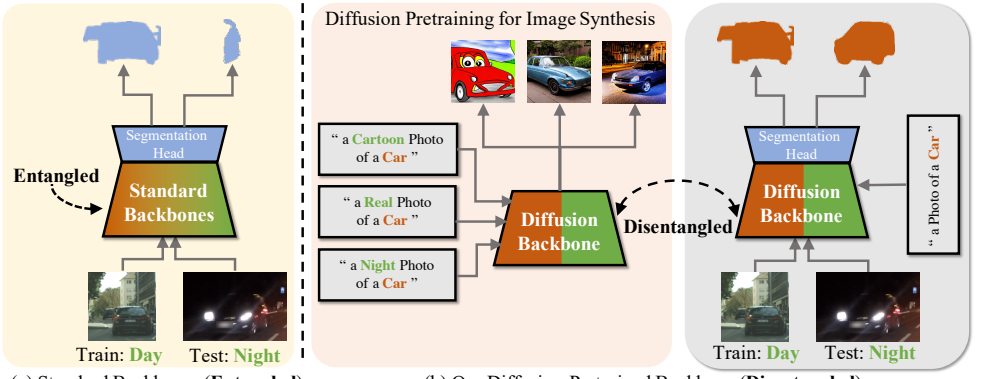
Abstract

While originally designed for image generation, diffusion models have recently shown to provide excellent pretrained feature representations for semantic segmentation. Intrigued by this result, we set out to explore how well diffusion-pretrained representations generalize to new domains, a crucial ability for any representation. We find that diffusion-pretraining achieves extraordinary domain generalization results for semantic segmentation, outperforming both supervised and self-supervised backbone networks. Motivated by this, we investigate how to utilize the model's unique ability of taking an input prompt, in order to further enhance its cross-domain performance. We introduce a scene prompt and a prompt randomization strategy to help further disentangle the domain-invariant information when training the segmentation head. Moreover, we propose a simple but highly effective approach for test-time domain adaptation, based on learning a scene prompt on the target domain in an unsupervised manner. Extensive experiments on four synthetic-to-real and clear-to-adverse weather benchmarks demonstrate the effectiveness of our approaches. Without resorting to any complex techniques, such as image translation, augmentation, or rare-class sampling, we set a new state-of-the-art on all benchmarks.

1 Introduction

Deep neural networks for semantic segmentation have achieved remarkable performance when trained and tested on the data from the same distribution [1, 26, 51, 55]. However, their ability to generalize to new and diverse data remains limited [45, 47, 49, 65]. Deep semantic segmentation models are sensitive to domain shifts, which occurs when the distribution of the

*This work was done before Rui joined Amazon.



(a) Standard Backbones (**Entangled**) (b) Our Diffusion-Pretrained Backbone (**Disentangled**)

Figure 1: **Standard *vs.* Diffusion Backbones.** (a) In standard backbones, the **domain-invariant** (*e.g.* semantics) and **domain-variant** (*e.g.* styles, lighting) information are partially entangled in the latent visual representations. Consequently, semantic segmentation models trained on these backbones often suffer from significant performance degradation when confronted with domain shifts. (b) Conversely, diffusion models naturally learn a more disentangled representation through text-to-image synthesis pretraining, as it is capable of generating the same semantic content under diverse styles. When used as a backbone network for semantic segmentation, it therefore offers a feature representation robust to large domain shifts, leading to superior generalization capabilities across diverse domains.

testing (target) data differs from that of the training (source) data. This often leads to drastic performance degradation. To enhance the generalization ability of deep models to unseen scenarios, domain generalization (DG) methods employ specialized training strategies that improve the model's robustness. Alternatively, test-time domain adaptation (TTDA) aims to fast adapt a model trained on source domain by only utilizing unlabelled target domain data.

Diffusion models have recently achieved extraordinary results for image generation and synthesis tasks [11, 12, 34, 37]. At the heart of the diffusion model lies the idea of training a denoising autoencoder to learn the reverse of a Markovian diffusion process. Trained on large-scale paired image-text datasets like LAION5B [42], diffusion models, such as Stable-Diffusion [37], have demonstrated remarkable performance on image synthesis controlled by natural language. The ability of large-scale text-to-image diffusion models to produce visually stunning images with intricate details, varied content, and coherent structures, while retaining the ability to modify and compose semantics, is a remarkable breakthrough. It implies that the diffusion models implicitly learn both high-level and low-level visual representations from the vast collections of image-text pairs dataset. Recently, frozen diffusion models have therefore been shown to provide excellent feature representations for semantic segmentation [56, 60], providing an alternative to standard supervised [6] or self-supervised pretraining [10].

In light of the success of diffusion models for supervised segmentation, we are led to contemplate: *How well do diffusion-pretrained semantic segmentation models generalize to unseen domains?* In this paper, we first investigate this question by comparing the generalization performance of diffusion-pretraining with other popular backbones and pretraining. We find that the vanilla diffusion models show exceptional generalization ability, surpassing that of all other backbones. We attribute this to the natural disentanglement of concepts that occur in diffusion models. As illustrated in Fig. 1(b), diffusion models can generate images of the same content, such as a car, under a variety of different styles and environments, e.g. *real*,

cartoon, and *night-time*. Due to this disentangled representation, the segmentation head learns more domain-invariant relations between the underlying features and the scene semantics, such as 'car'. When given an image from a different domain, the diffusion-based segmentation model (Fig. 1(b), right) is therefore able to more robustly identify and segmenting the object, compared to utilizing a standard backbone with a more entangled representation (Fig. 1(a)). These observations motivate us to explore the use of diffusion representations for DG/TTDA.

One key feature that distinguishes diffusion models from other backbones for semantic segmentation is their unique ability to manipulate the backbone using *prompt conditioning*. This unique feature grants us direct control over domains, enabling us to generalize and adapt to new domains directly with parameter-efficient prompts. In this work, we aim at designing ***simple yet effective*** methods for boosting DG and TTDA performance, without resorting to intricate techniques such as image translation, augmentation, or rare class sampling [13, 14, 57, 62]. To this end, we explore how to utilize the prompt in order to achieve even better generalization, or to adapt to new domains.

Domain Generalization: To improve the domain generalization ability of diffusion pretraining semantic segmentation models, we introduce category prompts and scene prompts as conditioning inputs to distinguish domain-invariant features from domain-variant ones. In addition, we propose a prompt randomization strategy to further improve the extraction and disentanglement of domain-invariant representations. This strategy ensures prediction consistency on the same image under different scene prompts, thereby enhancing the robustness of the model to domain shifts.

Test-Time Domain Adaptation: In order to facilitate adaptation of diffusion pretraining semantic segmentation models to the target domain during test time, we propose utilizing the scene prompt as the modulation parameter, which can be optimized via a loss function based on pseudo-labels during inference. The prompt tuning opens a new avenue for TTDA, which is parameter-efficient and mitigates the risk of overfitting.

To summarize, our contributions are four-fold:

- We conduct the first analysis of the generalization performance of diffusion pretrained models for semantic segmentation, demonstrating its superior performance.
- We introduce prompt-based methods, namely *scene prompt* and *prompt randomization*, to further improve the model's domain generalization capability.
- We propose a prompt tuning method to perform test-time domain adaptation of the model.
- Extensive experiments on four benchmarks demonstrate the effectiveness of our approach.

2 Related Work

Domain Generalization. Previous approaches for domain generalization can be categorized into two main strategies: 1) image augmentation and 2) feature normalization and whitening. The first strategy involves randomly stylizing or augmenting images from the source domain, a technique known as domain randomization [8, 44, 46, 59], to learn domain-invariant representations. The second strategy focuses on normalizing and whitening the features [2, 20, 29, 31, 48] to ensure robustness across different domains. In contrast to these previous methods, our approach differs by not relying on stylized or translated images or perturbed features. Instead, we solely regulate the behavior of the model backbone through the use of prompts. This new approach allows us to achieve domain generalization without need for extensive image transformations or feature manipulations.

Test-Time Domain Adaptation. Previous TTDA methods, also known as source-free domain adaptation [27, 50], often focus on tuning the parameters of batch normalization layers, which

are parameter-efficient. However, this approach has limitations in terms of adaptability and compatibility with network architectures other than convolutional neural networks, such as transformers [52]. Alternatively, some methods optimize the entire model or its main components, such as the feature backbone [23, 58]. However, such approaches tend to be parameter-heavy, making them prone to catastrophic overfitting to the noisy unsupervised learning objective, especially when the quantity of target domain data is limited. In contrast, our prompt-based method not only offers greater parameter efficiency compared to tuning batch normalization layers, but it also effectively modulates the behavior of the model.

3 Method

Problem Statement. *Test-time domain adaptation (TTDA):* The objective of TTDA is to adapt a model f_{θ^s}, with pre-trained with parameters θ^s, on a labeled source domain dataset $\{\mathbf{x}^s, \mathbf{y}^s\}$ in order to improve the performance on the unlabeled target domain $\{\mathbf{x}^t\}$. The adaptation $\theta^s \rightarrow \theta^t$ is performed post-training, without access to the source domain data. *Domain generalization (DG):* DG aims to generalize the model f_{θ^s}, trained on labeled source domain data $\{\mathbf{x}^s, \mathbf{y}^s\}$, to the unseen target domain $\{\mathbf{x}^t\}$, but without updating the model.
Diffusion-Pretraining for Segmentation. The basic idea [60] is to 1) utilize the pretrained diffusion model as the backbone network, 2) extract the visual internal representations $\{\mathbf{f}_i(\varepsilon_\theta, \mathbf{x}^s)\}$, and cross-attention maps $\{\mathbf{a}_i(\mathbf{f}_i, \mathcal{C})\}$ between the conditioning input $\mathcal{C}$ and the internal visual representations, and 3) feed the extracted $\{\mathbf{f}_i(\varepsilon_\theta, \mathbf{x}^s)\}$ and $\{\mathbf{a}_i(\mathbf{f}_i, \mathcal{C})\}$ into a learned semantic projection head $\mathcal{D}$, to obtain the predicted semantic segmentation map $\hat{\mathbf{y}}^s$,

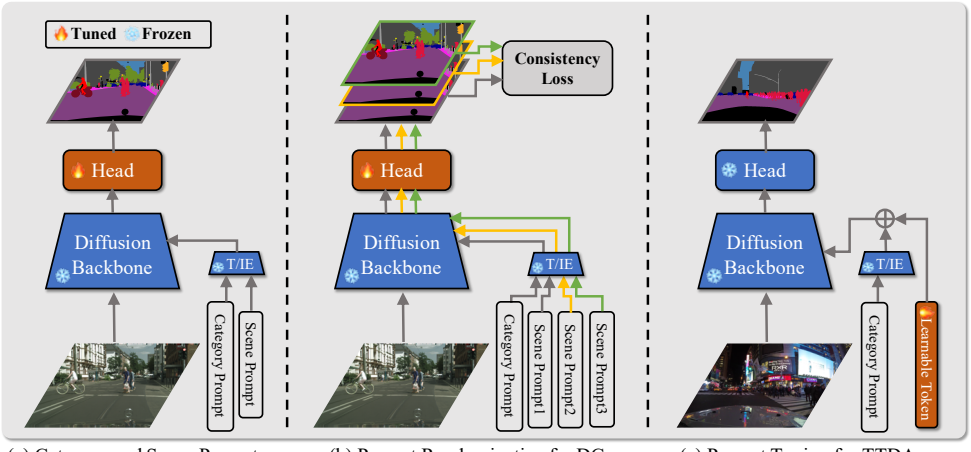
$$\hat{\mathbf{y}}^s = \mathcal{D}(\mathbf{f}_i(\varepsilon_\theta, \mathbf{x}^s), \mathbf{a}_i(\mathbf{f}_i, \mathcal{C})) \tag{1}$$

Then, the semantic projection head $\mathcal{D}$ is trained with the standard cross entropy loss, $\mathcal{L}_s = CE(\hat{\mathbf{y}}^s, \mathbf{y}^s)$. During the training, the diffusion model ε_θ is frozen and only the semantic projection head $\mathcal{D}$ is optimized, *i.e.* $\min_{\mathcal{D}} \mathcal{L}_s$.

3.1 Prompting Diffusion Representations for Domain Generalization

3.1.1 Generalization Abilities of Diffusion-Pretraining

Given the remarkable success of diffusion-pretraining for semantic segmentation, a natural inquiry arises: *To what extent do diffusion-pretrained segmentation models maintain their effectiveness under severe domain shift?* Motivated by this question, we first set out to investigate the diffusion-pretrained segmentation model's generalization performance in face of significant domain shift. In Table 1, we compare popular pretraining strategies: (1) supervised pretraining for image classification on ImageNet-22k [6]; (2) self-supervised pretraining for pixel reconstruction on ImageNet-1k [6]; (3) CLIP pretraining, consisting of contrastive visual-language pairing [33]; and (4) diffusion-driven pretraining for text-to-image synthesis on LAION-5B [42]. The oracle is the same network trained in a fully supervised manner on the target Cityscapes dataset. Note that the reported oracle mIoU values may appear lower than those in the literature on supervised learning [7, 10, 24, 25, 55], as we follow the established convention of prior DG and TTDA works that downsample the Cityscapes images by a factor of two, to ensure fair comparison. In all cases, we assess the model's performance on the Cityscapes validation set by reporting the mean intersection-over-union (mIoU). To evaluate the generalization ability of each model, we present the relative mIoU compared to the oracle, following [14]. Interestingly, higher oracle performance does not necessarily equate to better generalization on unseen target domains. This indicates that certain backbone models struggle with overfitting and do not effectively capture domain-invariant knowledge.



(a) Category and Scene Prompt (b) Prompt Randomization for DG (c) Prompt Tuning for TTDA

Figure 2: **Method Overview.** (a) We employ a frozen diffusion-pretrained backbone and a trained segmentation head. Our backbone is conditioned on a *category prompt* and the introduced the *scene prompt*, which are first input to the text/image prompt encoder "T/IE". The prompts aids the extraction of domain-invariant representations (Sec. 3.1.2). (b) We improve domain generalization by sampling random scene prompts and enforcing consistent predictions during training, facilitating learning of domain-invariant relations (Sec. 3.1.3). (c) For TTDA, we further propose a *prompt tuning* strategy, which adapts the representation by learning the scene prompt token (Sec. 3.2.2).

Table 1: **Comparison to other pretraining methods**, under GTA → Cityscapes.

Architecture	Swin-B[IN]	MiT-B5[IN]	ConvNeXt-B[IN]	MAE-ViT-L/16[IN]	CLIP-ViT-B[IN]	Stable-Diffusion[SD]
Pretraining Type	*supervised*	*supervised*	*supervised*	*self-supervised*	*visual-language*	*text-to-image*
Pretraining Dataset	*ImageNet-22k*	*ImageNet-22k*	*ImageNet-22k*	*ImageNet-1k*	*CLIP*	*LAION-5B*
Generalization	38.9	45.6	46.0	42.7	39.0	**49.2**
Oracle	79.2	76.4	79.9	76.8	70.0	74.7
Relative	49.1%	59.7%	57.6%	55.6%	55.7%	**65.9%**

However, we observe that the model using diffusion pretraining achieves superior performance in both absolute (49.2 mIoU) and relative (65.9% of the oracle) generalization. This demonstrates an exceptional generalization ability compared to other pretraining. This remarkable generalization performance reached by the vanilla diffusion pretrained model, encourages us to further investigate their potential benefits in domain adaptation and generalization.

The purpose of this work is to develop *simple yet effective* method for domain adaptation and generalization problems, without any complex tricks, such as image translation, data augmentation and class sampling. Building upon the characteristics of diffusion models, we note that these models are distinguished by their capacity to be finely controlled by the conditioning input $\mathcal{C}$, derived from image or text prompts. Different from previous methods, that change the backbone behaviors by modulating specific networks layers, stylizing images or introducing additional networks, prompts tuning opens a new avenue of manipulating the backbones representation in a effective and efficient way. In the next sections, we propose novel prompt tuning methods for domain generalization and test-time domain adaptation, based on diffusion-pretrained segmentation models. An overview is depicted in Fig. 2.

3.1.2 Category and Scene Prompt

To improve the generalization ability of diffusion-pretrained segmentation models, we first introduce the *category prompt* $\mathcal{C}_c$ and the *scene prompt* $\mathcal{C}_s$ as the conditioning inputs $\mathcal{C} =$

$[\mathcal{C}_c;\mathcal{C}_s]$. These are used to disentangle the domain-invariant features, such as object classes and scene layout, and the domain-variant features, such as object color, scene style and lighting.

Category Prompt. The category prompt is typically defined as a template of "`a photo of a [Class]`", where "`[Class]`" are *category names* (*e.g.* road, sidewalk and sky for the street scene image). *I.e.,* the category prompt only provides the class names as the guidance, to extract the domain-invariant knowledge. For instance, using the "car" class as an example, diffusion models can synthesize car images with varying attributes by providing different prompts. However, despite the diverse attribute inputs, the fundamental identity of the object as a car remains unchanged as long the prompts include "`a photo of a car`". This highlights the ability of category prompts to effectively capture knowledge related to the object's core identity, i.e. "what is a car?", from other attributes, such as color/body type. The core identity of object is domain-invariant and exactly what the domain generalization needs. For $C-$class segmentation, category prompts are C tokens, each of which is $N-$dim vector.

Scene Prompt. The category prompt can capture the main features of an object that stay the same across different scenarios, *i.e.* domain-invariant knowledge. To better extract domain invariant representations, we further condition the network on an introduced scene prompt. Our hypothesis is that the diffusion network can better extract domain-invariant representations if it is aware of the image domain. Consider e.g. a night photo of a street scene. It might be difficult to recognize objects, such as cars, pedestrians, and buildings in such conditions. However, by making the diffusion representation explicitly aware of the conditions through a style prompt "A dark night photo", we believe that it can partly revoke the domain-specific effect as it will explicitly consider a night-time view of a car, pedestrian, or building. Thus, to further facilitate the extraction of domain-invariant knowledge across different domains, we introduce the scene prompt, $\mathcal{C}_s$. One example of scene prompt is a template "`a [scene] photo`", *e.g.* "a GTA5 photo" or "a night photo".

By combining the category prompt $\mathcal{C}_c$ and the scene prompt $\mathcal{C}_s$ as the conditional inputs, the predicted semantic segmentation map in Eq. (1) is rewritten as,

$$\hat{\mathbf{y}}^s = \mathcal{D}(\mathbf{f}_i(\varepsilon_\theta, \mathbf{x}^s), \mathbf{a}_i(\mathbf{f}_i, [\mathcal{C}_c;\mathcal{C}_s])) \tag{2}$$

Note that the scene prompt $\mathcal{C}_s$ can not only be defined as the aforementioned text template, but also be designated as a $N-$dim learnable prompt, or an image prompt obtained by feeding an example image into pretrained language-image encoder (*e.g.* CLIP). With category and scene prompts employed, there are $M = C+1$ tokens of $N-$dim vector as conditioning inputs $\mathcal{C}$.

3.1.3 Prompt Randomization for Generalization

By incorporating the category and scene prompts as conditional inputs, the diffusion-pretraining segmentation model is able to extract domain-invariant knowledge, leading to enhanced generalization ability. To further capture domain-invariant knowledge and boost generalization capabilities, we propose a prompt randomization strategy. Our idea is to enforce consistency between the semantic predictions under different scene prompts $\{\mathcal{C}_s^k\}_{k=1}^K$. The intuition is that a model capable of generalizing well would make similar predictions for images containing the same content, irrespective of their domain-variant attributes, such as weather or style.

By feeding various scene prompts $\{\mathcal{C}_s^k\}_{k=1}^K$ into the diffusion backbone in Sec. 3.1.2, the corresponding semantic segmentation maps are obtained as $\{\hat{\mathbf{y}}_k^s\}_{k=1}^K$, where $\hat{\mathbf{y}}_k^s = \mathcal{D}(\mathbf{f}_i(\varepsilon_\theta, \mathbf{x}^s), \mathbf{a}_i(\mathbf{f}_i, [\mathcal{C};\mathcal{E}_k]))$. Then, the consistency loss, $\mathcal{L}_c$, between different scene prompts are,

$$\mathcal{L}_c = \sum_{p,q\in\{1,\dots,K\},q\neq p} KL(\hat{\mathbf{y}}_p^s || \hat{\mathbf{y}}_q^s) = -\sum_{p,q\in\{1,\dots,K\},q\neq p} \hat{\mathbf{y}}_p^s \log \frac{\hat{\mathbf{y}}_p^s}{\hat{\mathbf{y}}_q^s}, \tag{3}$$

where $KL(\cdot||\cdot)$ represents the Kullback–Leibler (KL) divergence [4], which aligns the semantic prediction under different scene prompts. The complete learning objective for prompt randomization is the combination of semantic segmentation loss $\mathcal{L}_s$ and consistency loss $\mathcal{L}_c$,

$$\mathcal{L}_{total} = \sum_{k=1}^{K} CE(\hat{\mathbf{y}}_k^s, \mathbf{y}^s) + \lambda \mathcal{L}_c. \tag{4}$$

Here, λ is the hyper-parameter used to balance the semantic segmentation loss and the consistency loss, which is set to 0.1 in this work.

3.2 Prompting Diffusion Representations for TTDA

3.2.1 Test-Time Domain Adaptation

Test-time domain adaptation (TTDA) presents two main challenges that must be addressed: (1) how can the source-domain initialized model be **modulated** effectively in light of unsupervised learning objective fraught with noise? (2) What **learning objectives** should be adopted to enable optimization if only unlabeled data from the target domain is provided? Our work primarily addresses challenge (1), and employs a pseudo-label based optimization objective for challenge (2) as it is proven simple yet effective in the test-time domain adaptation field.

3.2.2 Prompt Tuning for TTDA

Modulation Parameters. To effectively tackle the aforementioned challenge (1) in TTDA, the main focus is on identifying the relevant parameters that need to be updated in order to control the behavior of the backbone in a desirable manner. Our diffusion pretraining models, described in Sec. 3.1, leverage the category and scene prompts to effectively control the behavior of the backbone. More specifically, the *category prompt*, $\mathcal{C}_c$, captures *domain-invariant* knowledge on the object core identity, shared by the source and target domains. The *scene prompt*, $\mathcal{C}_s$, introduces the domain-specific information to further help disentangeling the representation. Test-time domain adaptation involves a domain shift from the source to the target domain. Therefore, the scene prompts needs to be updated to accommodate this shift in domains. The basic idea of our prompt tuning for TTDA is to learn the scene prompt $\mathcal{C}_s$ to facilitate adaptation from the source to the target domain. That is, the scene prompt serves as the modulation parameter, which can be updated by $\theta^t \leftarrow \theta^s : \mathcal{E} \leftarrow \mathcal{C}_s + \Delta\mathcal{C}_s$.
Learning Objective. Our test-time optimization objective $\mathcal{L}_t$ is to tune the scene prompt $\mathcal{E}$ supervised by the pseudo-label $\tilde{\mathbf{y}}^t = \arg\max \mathcal{D}(\mathbf{f}_i(\varepsilon_\theta, \mathbf{x}^t), \mathbf{a}_i(\mathbf{f}_i, [\mathcal{C}_c; \mathcal{C}_s]))$, formulated as,

$$\mathcal{L}_t = CE(\hat{\mathbf{y}}^t, \tilde{\mathbf{y}}^t), \qquad \mathcal{C}_s \leftarrow \mathcal{C}_s + \partial\mathcal{L}_t / \partial\mathcal{C}_s. \tag{5}$$

The only optimized parameters during test-time is the scene prompt $\mathcal{C}_s$, which is a $N-$dim vector and set as 768-dim in this work. Thus, our prompt-tuning method for TTDA is parameter-efficient, enabling fast adaptation and helping to mitigate the risk of overfitting.

4 Experiments

Setup. We evaluate the effectiveness of our proposed prompt-based method for DG and TTDA under different scenarios, including synthetic-to-real and clear-to-adverse benchmarks. We use the conventional notation A→B to describe the DG and TTDA task, where A and

Table 2: **Comparison to SOTA DG methods.** † denotes results obtained from [43].

(a) Synthetic-to-Real.

Method	Backbone	Extra Data	G→C	S→C
IBN-Net[29]	ResNet-101	✗	37.4	34.2
DRPC[53]	ResNet-101	✓	42.5	37.6
ISW[4]	ResNet-101	✗	37.2	37.2†
FSDR[18]	ResNet-101	✓	44.8	40.8
GTR[31]	ResNet-101	✓	43.7	39.7
SAN-SAW[34]	ResNet-101	✗	45.3	40.9
SHADE[51]	ResNet-101	✗	46.7	-
WEDGE[15]	ResNet-101	✓	45.2	40.9
AugFormer-S[43]	MiT-B5	✗	45.6	40.3
AugFormer[43]	MiT-B5	✗	-	44.2†
Ours (Van.)	*Diffusion*	✗	49.2	47.8
Ours (DG-T)	*Diffusion*	✗	**52.0**	49.1
Ours (DG-I)	*Diffusion*	✗	**52.0**	**49.3**

(b) Clear-to-Adverse (val set).

Method	Backbone	Extra Data	C→A	C→D
ISA[21]	ResNet-101	✗	47.4	26.1
ISW+ISA[21]	ResNet-50	✗	47.6	23.1
MixS [33]	ResNet-101	✗	37.0	9.4
MixS+ISA[21]	ResNet-101	✗	41.8	20.6
DSU[19]	ResNet-101	✗	38.3	12.3
DSU+ISA[21]	ResNet-101	✗	43.3	24.6
IBN-Net [29]	ResNet-50	✗	44.1	21.7
ISW+MSA[33]	ResNet-50	✗	47.3	22.5
ISW+MSA [33]	ResNet-101	✗	49.0	24.8
SiamDoGe [52]	ResNet-50	✗	52.3	-
Ours (Van.)	*Diffusion*	✗	57.0	31.2
Ours (DG-T)	*Diffusion*	✗	**58.6**	**34.0**
Ours (DG-I)	*Diffusion*	✗	58.4	**34.0**

B are source and target domain, respectively. *Synthetic-to-Real*: There are two settings, GTA [36] → Cityscapes [3] and SYNTHIA [38] → Cityscapes [3]. *Clear-to-Adverse*: There are also two tasks, Cityscapes [3] → ACDC [41] and Cityscapes [3] → Dark Zurich [39]. For ease of reference, we use the following abbreviations throughout the text: G→C, S→C, C→A, and C→D, respectively. *Scene Prompt*: The scene prompt for prompt randomization by default is composed of two components: (1) the text description for the source domain, such as "a GTA5 photo" for the GTA dataset, and (2) the text description for the target domain, such as "a night photo" for the Dark Zurich dataset, called the text prompt version. As an alternative of (2), we also experiment with an image from the target domain, *i.e.* the image prompt version. Specific prompts used for each experiment are put in the supplementary.

4.1 Comparison with state-of-the-art

Domain Generalization. In Sec. 3.1.3, we propose the prompt randomization strategy for DG with diffusion pretraining models. As shown in Table 2, our prompt randomization method is demonstrated to outperform previous SOTA DG methods by a significant margin. Notably, scene prompts for prompt randomization can be obtained flexibly in different types, including text (DG-T) and image (DG-I) prompts. And both types are proven effective, improving the vanilla diffusion models (Van.) performance significantly.

Test-time domain adaptation. In Sec. 3.2.2, we propose the prompt tuning strategy to adapt the diffusion representation during test time. Results in Table 3 demonstrate the superior performance of our prompt tuning method for TTDA, achieving a remarkable improvement of 5.9%, 6.5%, 2.7%, and 4.9% over existing TTDA methods on different benchmarks.

Unsupervised domain adaptation. As show in Table 4, our proposed DG methods (DG-T and DG-I) and TTDA method exhibit exceptional performance, outperforming even the strong unsupervised domain adaptation (UDA) methods that have access to both the source and target domain at the same time for training. Notably, our DG method achieves a performance gain of 5.8% over the strong UDA method, DAFormer, on the Cityscapes→ACDC benchmark, despite not utilizing any data from ACDC. Fig. 3 presents qualitative comparisons among our DG, TTDA, and DAFormer results, illustrating the effectiveness of our method. For instance, when examining the “road" class, we observe that without prompt conditioning, DAFormer segments the "road" into the upper sky region. In contrast, our category prompt conditioning, specifically "a photo of a road," assists the model in learning that the "road" should appear in the lower part rather than the upper sky region, thus preventing this error from occurring.

Table 3: **Comparison to SOTA TTDA methods.** "Param. Eff." represents parameter efficient. † represents the baseline is combined with [51] and trained with higher resolution images.

(a) Synthetic-to-Real.

Method	Backbone	Param. Eff.	G→C	S→C
TransAdapt[8]	ResNet-101	✓	37.8	33.7
Tent[51]	ResNet-101	✓	38.9	35.5
SFDA[43]	ResNet-50	✗	43.2	39.2
SFUDA[33]	ResNet-101	✗	49.4	44.2
URMA[1]	ResNet-101	✗	45.1	39.6
SHOT[24]	ResNet-101	✗	44.1	-
AUGCO[30]	ResNet-50	✓	47.1	39.5
HCL[14]	ResNet-101	✗	48.1	43.5
C-SFDA[43]	ResNet-101	✗	48.3	44.6
CO-SFDA[43]	ResNet-101	✓	46.3	43.0
Ours (Van.)	*Diffusion*	–	49.2	47.8
Ours (TTDA)	*Diffusion*	✓	**52.2**	**49.5**

(b) Clear-to-Adverse.

Method	Set	Backbone	Param. Eff.	C→A	C→D
TTBN[43]	Val-Set	ResNet-101	✓	-	28.0
TENT[51]		ResNet-101	✓	-	26.6
AUGCO[30]		ResNet-101	✓	-	32.4
CO-SFDA[43]		ResNet-101	✓	-	33.2
MSA[65]		ResNet-101	✓	47.9	22.8
MSA[65]		MiT-B0	✓	46.6	20.2
Ours (Van.)		*Diffusion*	–	57.0	31.2
Ours (TTDA)		*Diffusion*	✓	**58.5**	**37.0**
TENT	Test-Set	ResNet-101	✓	49.0	-
HCL		ResNet-101	✓	46.8	-
URMA		ResNet-101	✓	47.2	-
SegFormer [65] †		MiT-B5	✗	59.3	42.8
Ours (TTDA)		*Diffusion*	✓	**62.0**	**47.7**

Table 4: **Comparison between our DG, TTDA method to UDA methods,** under C→A. All methods are evaluated on the test set through the online public evaluation server. † represents the use of extra auxiliary reference images that are geographically aligned and captured under clear-weather/daytime.

Setting	Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MC	Bike	mIoU
	Cityscapes → ACDC (**Test Set**)																				
UDA	ADVENT[65]	72.9	14.3	40.5	16.6	21.2	9.3	17.4	21.2	63.8	23.8	18.3	32.6	19.5	69.5	36.2	34.5	46.2	26.9	36.1	32.7
	GCMA[65] †	79.7	48.7	71.5	21.6	29.9	42.5	56.7	57.7	75.8	39.5	87.2	57.4	29.7	80.6	44.9	46.2	62.0	37.2	46.5	53.4
	MGCDA[65] †	73.4	28.7	69.9	19.3	26.3	36.8	53.0	53.3	75.4	32.0	84.6	51.0	26.1	77.6	43.2	45.9	53.9	32.7	41.5	48.7
	DANNet[65] †	84.3	54.2	77.6	38.0	30.0	18.9	41.6	35.2	71.3	39.4	86.6	48.7	29.2	76.2	41.6	43.0	58.6	32.6	43.9	50.0
	DAFormer[65]	58.4	51.3	84.0	42.7	35.1	50.7	30.0	57.0	74.8	52.8	51.3	58.3	32.6	82.7	58.3	54.9	82.4	44.1	50.7	55.4
DG	Ours (DG-T)	89.6	62.5	84.4	48.6	39.9	49.2	48.7	55.6	74.5	48.3	86.1	60.3	39.9	84.9	62.6	63.5	73.6	37.7	52.3	**61.2**
	Ours (DG-I)	91.1	85.9	84.7	83.8	76.4	66.3	66.1	62.6	56.4	56.1	54.6	54.4	51.4	50.6	49.6	47.8	42.2	41.1	38.8	61.0
TTDA	Ours	88.2	86.0	85.6	85.5	74.0	73.8	62.3	61.2	60.0	57.6	56.5	55.9	52.3	52.0	50.8	50.3	42.4	42.2	41.8	**62.0**

4.2 Ablation Study and Prompts Analysis

Different Scene Prompts Comparison. We conduct ablation experiments to evaluate the effectiveness of our proposed scene prompt in improving the domain generalization performance of diffusion pretraining models. *With and Without Scene Prompt*: First, we compare the performance of models with and without scene prompt to verify its effectiveness. As shown in Table 6, the models with any of the different scene prompts (target, learned, and source) all outperformed the baseline model without scene prompt, achieving mIoU scores of 50.9%, 51.4%, and 51.4% respectively, compared to 49.2% for the baseline on the GTA→Cityscapes benchmark. *Different Scene Prompts*: Next, we investigate the impact of different scene prompt choices by comparing the prompts obtained from 1) text description of target domain, referred to as "Target"; 2) text description of source domain, referred to as "Source"; and 3) a learnable parameter, referred to as "Learned". Our results show that the "Source" scene prompt outperforms the "Target" and "Learned" prompts on different benchmarks. This finding confirms our statement in Sec. 3.1.2 that the scene prompt is used to disentangle domain-invariant knowledge and revoke the effect of domain-variant factors in the source domain. Hence, the "Source" prompt, which captures domain-variant factors in the source domain, works best.

Increased Number of Class Prompts. To assess the impact of classes in category prompts, we conduct an experiment in which more classes were utilized in the category prompt. More specifically, in all GTA→Cityscapes experiments in this work, the category prompt consisting of 19 classes is used. To obtain the category prompt with more classes, we utilize 150 classes from the ADE20K dataset [62]. This expanded set of classes not only includes the 19 object

Table 5: **Prompts analysis on,** 1) increased number of category prompts (150 classes); 2) scene prompts (a "water, grass, sand, painting" photo) irrelevant to source/target domain, under G → C.

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MC	Bike	mIoU
*Increased number of **category prompts***																				
19 Classes	70.9	25.7	88.1	54.5	43.6	43.5	46.3	32.7	87.8	52.1	90.9	63.0	24.9	87.7	34.5	42.6	2.9	22.2	20.8	49.2
150 Classes	76.0	29.9	88.2	41.5	41.3	45.0	45.2	29.5	87.8	52.4	90.2	63.7	27.7	86.9	34.6	47.3	8.1	28.6	28.4	50.1
*Prompt randomization with irrelevant **scene prompts***																				
Irrelevant	85.6	36.7	87.8	52.7	44.9	41.7	45.8	31.4	87.5	51.2	89.7	64.1	29.5	87.4	29.7	35.7	13.0	31.7	36.9	51.7
DG-Text	87.7	36.2	87.7	43.3	38.2	38.1	44.4	31.2	87.6	48.3	89.8	63.9	31.7	89.4	61.9	50.6	0.1	24.5	33.9	52.0

classes used in our standard category prompt, but also encompasses a variety of additional classes. Our analysis in Table 5 demonstrates that increasing the number of classes in the category prompt leads to a improvement in the generalization ability of diffusion pretraining semantic segmentation models, 50.1% *vs.* 49.2%. This improvement can be attributed to the fact that providing a greater number of classes as the category prompt enables the diffusion representations to more accurately distinguish between a broader range of class objects, thereby mitigating the issue of mis-classification. This finding suggests a promising direction for future work to further improve the generalization ability by incorporating more auxiliary classes into the category prompt.

Prompt Randomization with Irrelevant Prompts. In our primary experiments, the scene prompt is generated from text/image prompts that are relevant to the target domain. However, to evaluate the flexibility and scalability of the scene prompt, we conduct an experiment in which the scene prompt is generated from a random, unrelated scene description. For instance, in the GTA→Cityscapes experiment, we used scene prompts such as "a sand photo," "a grass photo," "a painting photo," and "a water photo." Results presented in Table 5 demonstrate that the prompt randomization method using irrelevant prompts performs favorably in comparison to the method using target-relevant prompts, achieving 51.7% and 52.0%, respectively.

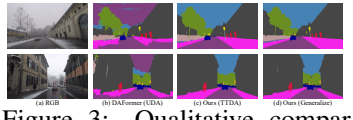


Figure 3: Qualitative comparisons between our DG, TTDA and UDA [14] method. DAFormer is a UDA method, using full target domain data during training.

Table 6: **Ablation study and comparisons between different prompts types,** under synthetic-to-real and clear-to-adverse benchmarks. Results are evaluated on the val-set of target domain.

Setting	w/o $\mathcal{C}_s$	Target ($\mathcal{C}_s$)	Learned ($\mathcal{C}_s$)	Source ($\mathcal{C}_s$)	TTDA	DG-T	DG-I
G→C	49.2	50.9	51.4	51.4	52.2	52.0	52.0
S→C	47.8	48.4	48.2	48.8	49.5	49.1	49.3
C→D	31.2	32.2	30.4	32.8	37.0	34.0	34.0
C→A	57.0	57.0	58.0	58.0	58.5	58.6	58.4

5 Conclusion

In this work, we conduct the first study on the generalization performance of diffusion pretraining semantic segmentation models, showing their superiority over other pretraining. We introduce novel prompt-based methods—the scene prompt and prompt randomization—to enhance domain generalization. Also, we propose prompt tuning for efficient and effective test-time domain adaptation. Extensive experiments validate our simple yet powerful approach.

Limitations. The current approach employs hand-designed prompts. An interesting future direction is to leverage other large language models to automatically generate accurate prompts for our method.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2017.

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

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

[4] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[5] Debasmit Das, Shubhankar Borse, Hyojin Park, Kambiz Azarian, Hong Cai, Risheek Garrepalli, and Fatih Porikli. Transadapt: A transformative framework for online test time adaptive semantic segmentation. In *ICASSP*, 2023.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[7] 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.

[8] Qi Fan, Mattia Segu, Yu-Wing Tai, Fisher Yu, Chi-Keung Tang, Bernt Schiele, and Dengxin Dai. Normalization perturbation: A simple domain generalization method for real-world domain shifts. *arXiv preprint arXiv:2211.04393*, 2022.

[9] Francois Fleuret et al. Uncertainty reduction for model adaptation in semantic segmentation. In *CVPR*, 2021.

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018.

[14] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*, 2022.

[15] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *CVPR*, 2021.

[16] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *NeurIPS*, 2021.

[17] Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-pang Chiu, Supun Samarasekera, and Nazanin Rahnavard. C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. In *CVPR*, 2023.

[18] Namyup Kim, Taeyoung Son, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Wedge: web-image assisted domain generalization for semantic segmentation. *arXiv preprint arXiv:2109.14196*, 2021.

[19] Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. *arXiv preprint arXiv:2202.03958*, 2022.

[20] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *NeurIPS*, 2017.

[21] Yumeng Li, Dan Zhang, Margret Keuper, and Anna Khoreva. Intra-source style augmentation for improved domain generalization. In *WACV*, 2023.

[22] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, 2020.

[23] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *CVPR*, 2021.

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

[25] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[27] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: dynamic unsupervised domain adaptation by normalization. In *CVPR*, 2022.

[28] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

[29] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018.

[30] Duo Peng, Yinjie Lei, Lingqiao Liu, Pingping Zhang, and Jun Liu. Global and local texture randomization for synthetic-to-real semantic segmentation. *TIP*, 2021.

[31] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. In *CVPR*, 2022.

[32] Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Augco: augmentation consistency-guided self-training for source-free domain adaptive semantic segmentation. *arXiv preprint arXiv:2107.10140*, 2021.

[33] 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 *ICML*, 2021.

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[35] Nikhil Reddy, Abhinav Singhal, Abhishek Kumar, Mahsa Baktashmotlagh, and Chetan Arora. Master of all: Simultaneous generalization of urban-scene segmentation to all adverse weather conditions. In *ECCV*, 2022.

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

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[38] 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 *CVPR*, 2016.

[39] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *ICCV*, 2019.

[40] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *TPAMI*, 2020.

[41] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021.

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[43] Manuel Schwonberg, Fadoua El Bouazati, Nico M Schmidt, and Hanno Gottschalk. Augmentation-based domain generalization for semantic segmentation. *arXiv preprint arXiv:2304.12122*, 2023.

[44] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IROS*, 2017.

[45] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021.

[46] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *CVPR Workshops*, 2018.

[47] 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 *CVPR*, 2018.

[48] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *CVPR*, 2017.

[49] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019.

[50] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

[51] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020.

[52] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *CVPR*, 2022.

[53] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *CVPR*, 2021.

[54] Zhenyao Wu, Xinyi Wu, Xiaoping Zhang, Lili Ju, and Song Wang. Siamdoge: Domain generalizable semantic segmentation using siamese network. In *ECCV*, 2022.

[55] 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 *NeurIPS*, 2021.

[56] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. *arXiv preprint arXiv:2303.04803*, 2023.

[57] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020.

[58] Mucong Ye, Jing Zhang, Jinpeng Ouyang, and Ding Yuan. Source data-free unsupervised domain adaptation for semantic segmentation. In *ACM MM*, 2021.

[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 *ICCV*, 2019.

[60] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. *arXiv preprint arXiv:2303.02153*, 2023.

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

[62] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017.

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

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[65] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018.