A Study on Unsupervised Domain Adaptation for Semantic Segmentation in the Era of Vision-Language Models

Manuel Schwonberg[*1,3]
schwonberg@campus.tu-berlin.de
Claus Werner[*2,3]
claus.werner@cariad.technology
Hanno Gottschalk[1]
gottschalk@math.tu-berlin.de
Carsten Meyer[2,4]
carsten.meyer@ostfalia.de

[*] equal contribution

[1] Mathematical Modeling of Industrial Life Cycles, TU Berlin
[2] Department of Computer Science, Ostfalia University of Applied Sciences
[3] CARIAD SE
[4] Department of Computer Science, Kiel University

Abstract

Despite the recent progress in deep learning based computer vision, domain shifts are still one of the major challenges. Semantic segmentation for autonomous driving faces a wide range of domain shifts, e.g. caused by changing weather conditions, new geolocations and the frequent use of synthetic data in model training. Unsupervised domain adaptation (UDA) methods have emerged which adapt a model to a new target domain by only using unlabeled data of that domain. The variety of UDA methods is large but all of them use ImageNet pre-trained models. Recently, vision-language models have demonstrated strong generalization capabilities which may facilitate domain adaptation. We show that simply replacing the encoder of existing UDA methods like DACS by a vision-language pre-trained encoder can result in significant performance improvements of up to 10.0% mIoU on the GTA5→Cityscapes domain shift. For the generalization performance to unseen domains, the newly employed vision-language pre-trained encoder provides a gain of up to 13.7% mIoU across three unseen datasets. However, we find that not all UDA methods can be easily paired with the new encoder and that the UDA performance does not always likewise transfer into generalization performance. Finally, we perform our experiments on an adverse weather condition domain shift to further verify our findings on a pure real-to-real domain shift.

1 Introduction

Computer vision has experienced several breakthroughs in the past decade enabled by deep neural networks (DNNs) [11, 18, 30, 44]. Domain shifts, i.e. when the training and inference distribution differ, are still a major challenge for DNNs and can cause severe performance drops, e.g. when models are trained on synthetic data and inference is done on real data [21,

54, 60]. These can significantly hamper the application especially to safety-critical areas like autonomous driving and medical image analysis. Consequently, the mitigation of domain shifts is a major objective and several research fields emerged.

The most popular field is unsupervised domain adaptation (UDA) where only unlabeled samples from the target domain are available and the objective is to adapt the model towards this specific target domain. A broad variety of UDA methods have been developed in the past years utilizing methods like adversarial adaptation [60, 61], contrastive learning [28, 67, 72], self-training [69, 75] and knowledge distillation frameworks [21, 55, 71]. Next to unsupervised domain adaptation so-called domain generalization (DG) methods gained increasing research attention [4, 7, 13, 25, 29, 32, 41, 53, 58, 70, 73]. Here, no target data is available and the objective is to obtain a model which generalizes well across multiple unseen target domains. Very recently, the utilization of vision-language models (VLMs) enabled a major performance increase as well as methodological progress in the field of domain generalization with works like Rein [66], CLOUDS [2], and VLTSeg [26]. These approaches demonstrate that large-scale vision-language pre-training like CLIP [44] can be leveraged to significantly improve domain generalized semantic segmentation as well as detection [63] performance. Surprisingly, the field of UDA research falls back behind DG research because so far all existing UDA methods use ImageNet pre-trained models leaving the strong potential of vision-language models unexplored except a very recent study from Englert *et al*. [12]. For this reason, we equip selected UDA methods with a state-of-the-art vision-language pre-trained encoder and show their strong potential for unsupervised domain adaptation. We also evaluate the domain generalization capabilities of the UDA methods on unseen datasets and show that a simple UDA method with a vision-language pre-trained backbone provides state-of-the-art generalization capabilities. Next to the established synthetic-to-real and real-to-real benchmarks which contain a mixture of different domain shifts (e.g. GTA5→Cityscapes contains synthetic and geolocation shift) we evaluate both the adaptation and generalization performance for a single pure adverse weather condition shift on the ACDC [50] dataset. This is motivated by the results from Sakaridis *et al*. [50] that UDA methods perform very different on different domain shifts and in some cases worsen the performance. Overall, our study makes the following contributions and novel findings:

- Extensive evaluation of UDA methods, in particular DACS [59], equipped with a state-of-the-art vision-language pre-trained backbone demonstrating that this can boost the UDA target performance by 10.0% mIoU and across three unseen datasets by 13.7% mIoU
- Analysis of UDA methods revealing that not all methods are similarly compatible with a VLM-based encoder; this indicates the need for new UDA methods tailored towards vision-language models
- Extensive evaluation on unseen datasets showing that the UDA and generalization performance are not necessarily correlated and that recent DG methods can provide better generalization than UDA methods

2 Related Work

Unsupervised Domain Adaptation (UDA) Methods There exists a broad variety of UDA methods which can be coarsely clustered into input, feature, output space and hybrid adap-

tation methods [54]. In the input space several approaches apply a GAN-based style transfer between domains [3, 6, 20, 31, 65] or augmentation and image mixing techniques [1, 39, 59]. In the feature space adversarial and contrastive learning UDA methods are often used [19, 34, 37, 38, 60], whereas in the output space self-training, contrastive learning and knowledge distillation are common techniques [60, 67, 72, 75]. Hybrid approaches combine multiple of the mentioned techniques and have become increasingly popular for UDA in recent years [24, 65]. However, hybrid approaches are mostly complex and for this reason we also include simpler approaches like DACS [59] in our study.

UDA Architectures The vast majority of UDA approaches employed ImageNet pre-trained VGG-16 [56] and ResNet-101 [18] networks as their backbones [20, 33, 51, 60], so that these architectures became the de-facto standard. With the emerging vision transformers and the foundational work DAFormer [21] the mix vision transformer (MiT-B5) proposed by [68] became a popular backbone for UDA methods [22, 67]. However, all of these backbones are ImageNet pre-trained and do not harness the strong generalization power of vision-language pre-training. Only the recent study from Englert *et al*. [12] investigates foundation models for UDA methods but our study differs in three important aspects. First, we also include previous, simpler UDA methods like DACS [59] and demonstrate that these methods also significantly benefit from vision-language pre-training. Second, Englert *et al*. [12] focus on the DINOv2 model [42] where the target domain Cityscapes [9] is used to retrieve similar samples from the web and therefore results may not transfer to other settings. We instead focus on the vision-language pre-trained model EVA02-CLIP [57]. Third, we follow the benchmark protocol from DG approaches [4, 25, 26, 41, 66] enabling a comparison to those studies. Our study shares similarities with the work from Piva *et al*. [43] w.r.t. their evaluation methodology; however, vision-language models are missing in their study and UDA and DG have made significant progress since then.

UDA Benchmarks Synthetic-to-real and real-to-real shifts are mostly used for benchmarking in UDA. The most common synthetic datasets are GTA5 [46] and SYNTHIA [47] and recently Urbansyn [17]. As target domain for the synthetic-to-real domain shift the Cityscapes dataset [9] is widely employed which is also often used as real source domain for the real-to-real domain shift. Real target domains are often the ACDC [50], FoggyCityscapes [48] and DarkZurich [49] datasets. All these domain shifts represent a mixture of at least two domain shifts or rely on artificially generated shifts like in the FoggyCityscapes dataset.

Vision-Language Models (VLMs) CLIP by Radford *et al*. [44] was the foundational work in the field of vision-language models which was trained with image-text pairs and a contrastive loss for the alignment of vision and text embeddings. VLMs like CLIP benefit from large-scale multi-modal datasets like Laion-5B [52] or Commonpool [15] and their size is an inherent advantage since the image-text pairs are easier to collect than e.g. single-class labels for ImageNet [44]. VLMs are commonly used for pre-training and then employed in a transfer learning setting for a downstream task, e.g. semantic segmentation [10, 16, 36, 45, 74] but so far not for UDA in semantic segmentation except the study from Englert *et al*. [12].

3 Method

In this section, we describe the details of our study, focusing on the evaluated UDA methods, the model architectures and the domain shifts of the ACDC dataset.

3.1 UDA Methods

In contrast to Englert *et al.* [12] we decide to not only focus on the most recent and usually more complex UDA methods, but deliberately also include previous methods. That shows how previous UDA methods and principles like adversarial adaptation benefit from the new vision-language pre-trained backbone and how they perform on a pure domain shift. We select previous, highly influential UDA methods: AdaptSegNet [60], ADVENT [64] and DACS [59]. In addition, we include the recent state-of-the-art methods SePiCo [67], DAFormer [21], and MIC [24]. AdaptSegNet [60] is one of the earliest UDA works and utilizes adversarial domain adaptation by employing a domain discriminator in both the feature and output space. Similar to AdaptSegNet, ADVENT [64] applies adversarial learning and self-training on the entropy maps of the output space. DACS [59] is an easy-to-apply method which is incorporated by several subsequent UDA methods, combining input space cross domain image mixing and adaptive self-training. SePiCo [67] as one of the current state-of-the-art methods proposes multiple contrastive losses along with a teacher-student framework to align the source and target domains. DAFormer [21] was the first work using a vision transformer backbone for UDA and applied self-training, rare class sampling and an ImageNet feature distance loss to preserve ImageNet knowledge. We include both the initial DAFormer method and its follow-up approach MIC [24] with HRDA [22].

3.2 UDA Architectures

Encoder & Initialization For the encoder choice, we follow recent domain generalization approaches [26, 66] and employ the EVA02-CLIP-L-14 vision encoder [57] which has shown strong generalization capabilities for segmentation. EVA02-CLIP [57] relies on a sequence of CLIP and masked image modeling pre-training. Note that we only use the vision encoder of the pre-training and refer to it as EVA02-L. Hümmer *et al.* [26] demonstrated the strong generalization capabilities of the EVA02-CLIP encoder which makes it a natural candidate for our study. Both Wei *et al.* [66] and Englert *et al.* [12] focused on DINOv2 [42] pre-trained weights as their initialization. We are not using DINOv2 pre-training in our study since the Cityscapes dataset, which is one of our main target domains, is used to sample the pre-training dataset of DINOv2, reducing the significance of evaluations. Moreover, we include the established UDA architectures DeepLabv2 with a ResNet-101 backbone [5] and the DAFormer architecture with a MiT-B5 backbone [21] as it is common practice in UDA and DG benchmarking [21, 23, 41, 57, 72].

Decoder We employ an ASPP-based decoder with different dilation rates from the DAFormer architecture [21]. The ASPP-decoder receives multi-level features from different levels of the encoder and performs up-sampling to obtain a common size of the feature maps in case of a hierarchical encoder like a ResNet-101 or a MiT-B5. When using the EVA02-L encoder this upsamling has no effect since the encoder is non-hierarchical.

3.3 Domain Shift Datasets

Most of the real-to-real domain shifts for benchmarking are a mixture of at least two different domain shifts like Cityscapes→ACDC and Cityscapes→DarkZurich. Those benchmarks contain a geolocation shift, a weather/condition shift and also have been recorded with different cameras. In contrast, we aim to evaluate UDA methods in scenarios which exclusively cover only a single, well defined domain shift, e.g. only a geolocation or only an adverse

weather condition shift. The available datasets for such an evaluation are limited. To the best of our knowledge only the ACDC [50] and the DarkZurich [49] datasets offer a well-defined single domain shift with 1:1 scene correspondences. DarkZurich is not included in our experiments because the pure day-to-nighttime shift is already contained in ACDC. Diverse datasets like BDD100K [69] or IDD [62] do not contain the required metadata which enable a pure domain shift evaluation. Since ACDC offers direct scene correspondences between normal daytime weather and night, snow, fog and rain conditions, we chose the clean $\rightarrow$ adverse weather condition domain shift and the ACDC dataset for evaluation. We refer to the normal weather daytime images as $\mathcal{D}^{\text{ACDC}}_{\text{normal}}$ and the adverse weather domains as $\mathcal{D}^{\text{ACDC}}_{\text{snow}}$, $\mathcal{D}^{\text{ACDC}}_{\text{fog}}$ etc. while the official train, validation and test set with all subdomains are denoted as $\mathcal{D}^{\text{ACDC}}_{\text{train}}$, $\mathcal{D}^{\text{ACDC}}_{\text{val}}$ and $\mathcal{D}^{\text{ACDC}}_{\text{test}}$.

Next to this pure shift of the ACDC dataset we follow common practice in the UDA and DG field [14, 21, 41, 60, 66] and employ the GTA5 [46] and the SYNTHIA [47](SYN) dataset as the synthetic source domains with 24966 and 9400 images, respectively. As the real-world domain we utilize Cityscapes [9] (CS), Mapillary Vistas [40] and BDD100K [69] with 2975/500, 18000/2000, 7000/1000 train/validation images respectively. We denote the respective datasets with a subscript as e.g. $\mathcal{D}^{\text{CS}}_{\text{train}}$ or $\mathcal{D}^{\text{CS}}_{\text{val}}$. All values of this study are reported on the respective validation datasets. For the DG evaluation only the validation sets of the corresponding domains are used. The ACDC dataset [50] which is both used as source and target domain contains 1000 images in each sub-domain from which 400 are training, 100 validation and 500 are test images. For half of them reference images under normal weather conditions are available and were used for our new clean ACDC domain shift evaluation.

3.4 Experimental Settings

Implementation Details All experiments are based on the open source framework MMSegmentation [8] and were conducted on a single A100 GPU with 80GB memory. The crop resolution for all experiments was fixed to 512×512 except for MIC [24], where a 1024×1024 resolution was used. The number of training iterations was set to 40k as common practice and a batch size of four was used. For ADVENT and AdaptSegNet with a ResNet-101 backbone the SGD optimizer with a learning rate of $2.5e-03$ was used. In all other cases, the AdamW [35] optimizer was selected in line with previous approaches [21, 26, 66]. For the MiT-B5 backbone a learning rate of $6e-05$ and for the EVA02-L encoder of $1e-05$ was used. Hyperparameters specific to the respective UDA methods were set as given by the authors without change.

Metric As the evaluation metric we use the mean intersection over union (mIoU) averaged across 19 classes which are shared among all synthetic and real datasets. Only for SYNTHIA the mIoU across 13 classes is reported as common standard in UDA [55, 64, 67, 72].

4 Results

In this section, we show the results and start with the evaluation of the UDA performance with vision-language pre-training followed by the domain generalization evaluation. All results obtained with the vision-language pre-trained encoder EVA-02-CLIP will be highlighted with this green color.

4.1 UDA with vision-language pre-training

We equipped four of the selected UDA methods with the vision-language pre-trained EVA02-L encoder and compared it to the current two standard architectures ResNet-101 and MiT-B5, both initialized with ImageNet pre-trained weights. We could not include the combination of MIC and EVA02-CLIP in our study, but future work should investigate this combination.

GTA5→Cityscapes											
Architecture		UDA Method (in % mIoU)									
Encoder	Decoder	AdaptSegNet [60]		ADVENT [64]		DACS [59]		DAFormer [21]		Src. Only	Oracle
		Abs.	Rel.	Abs.	Rel.	Abs.	Rel.	Abs.	Rel.		
ResNet-101	DeepLabV2	44.0	60.4	43.4	59.6	55.3	76.0	57.9	79.5	36.0	72.8
MiT-B5	DAFormer	47.5	60.8	47.1	60.3	60.3	77.2	67.6	86.6	47.6	78.1
EVA02-L	DAFormer	59.5	73.1	59.2	72.7	70.3	86.4	68.0	83.5	60.5	81.4

Table 1: **UDA performance on the GTA5 → Cityscapes** domain shift using different model architectures. "Source only" denotes training on $\mathcal{D}^{\text{GTA5}}$, without any UDA methods, and "Oracle" training on $\mathcal{D}^{\text{CS}}_{\text{train}}$. "Abs." refers to the absolute mIoU on the $\mathcal{D}^{\text{CS}}_{\text{val}}$ dataset whereas "Rel." denotes the performance in % relative to the oracle performance.

Results for the three model architectures on the common synthetic-to-real domain shift GTA5→ Cityscapes are presented in Table 1. We observe that equipping the simple DACS [59] method with the EVA02-L backbone causes a performance gain on Cityscapes of 10.0% mIoU compared to the standard MiT-B5 encoder and also significantly raises performance relative to the oracle performance (supervised training on $\mathcal{D}^{\text{CS}}_{\text{train}}$) from 77.2% to 86.4%. We reason that similar to previous works [2, 26, 66] the vision-language pre-training provides a stronger backbone which better adapts to the target domain. However, the DAFormer [21] approach does not benefit from the EVA02-L backbone and the performance remains at a level similar to that for the MiT-B5 backbone. This may be caused by the ImageNet feature distance (FD) loss which is designed to preserve the ImageNet pre-trained knowledge by minimizing the feature distance between the ImageNet and the synthetic object classes. We analyze the feature distance loss of the DAFormer [21] training as plotted in Figure 1. The FD-loss is 5-10× higher for the EVA02-L backbone than for the MiT-B5 backbone and shows a different behavior at the beginning. This is reasonable since the loss is based on the ImageNet classes which do not align with the vision-language pre-trained EVA02-L backbone. The FD-loss is part of DAFormer [21], HRDA [22] and MIC [24] and provides a performance increase of 3.5% mIoU according to the original paper [21] but makes those methods hard to transfer to backbones which are not initialized with ImageNet pre-trained weights. The UDA methods AdaptSegNet [60] and ADVENT [64] do not yield further gains to the source only performance with the EVA02-L backbone. These results indicate that the gain of a new backbone depends on the UDA method.

In Table 2 we compare the performance of DACS [59] and DAFormer [21] with the newly employed EVA02-L backbone to the published state-of-the-art performances of other approaches using a ResNet-101 and MiT-B5 backbone. We observe that the performance with the DACS method performs similar to recent works like SePiCo [67] for both GTA5 and SYNTHIA as the source domain. However, compared to MIC [24] the performance is less which may be also caused by the lower resolution. Notably, the performance of DAFormer [21] reduces for SYNTHIA→Cityscapes with the EVA02-L backbone but increases significantly for Cityscapes→ACDC. We further evaluated on a pure adverse weather domain

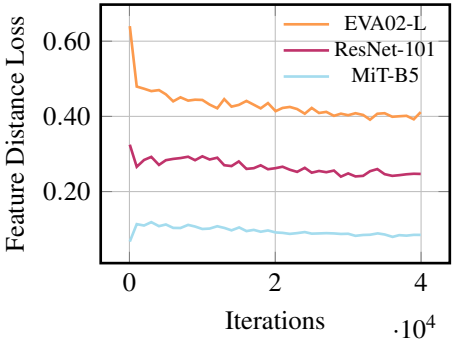

Figure 1: **Feature Distance Loss** over the 40k training iterations of the DAFormer [21] adaptation with three different backbones.

Encoder	Method	GTA5 → CS	SYN → CS	CS → ACDC
ResNet-101	AdaptSegNet [60]	42.4	46.7	33.4*
	ADVENT [64]	45.5	48.0	32.7*
	DACS [59]	52.1	54.8	-
	DAFormer [21]	56.0	-	-
	SePiCo [63]	61.0	66.5	-
	HRDA [24]	63.0	69.2	57.6*
	MIC [24]	64.2	70.7	60.4*
MiT-B5	DAFormer [21]	68.3	67.4	55.4*
	SePiCo [63]	70.3	71.4	59.1*
	HRDA [24]	73.8	72.4	68.0*
	MIC [24]	**75.8**	**74.0**	70.4*
EVA02-L	DACS [59]	70.3	72.3	**72.0**
	DAFormer [21]	68.0	64.7	68.0

Table 2: **Comparison with state-of-the-art UDA methods.** * marks the performance on the ACDC test set $\mathcal{D}_{\text{test}}^{\text{ACDC}}$. All values taken from respective papers except EVA02-L.

shift, using the ACDC [50] reference images under normal weather conditions $\mathcal{D}_{\text{normal}}^{\text{ACDC}}$ and their counterparts taken under adverse weather conditions e.g. $\mathcal{D}_{\text{fog}}^{\text{ACDC}}$ etc. As shown in Table 3 we both adapted the model solely to the respective sub-domains and also to $\mathcal{D}_{\text{train}}^{\text{ACDC}}$ which contains all sub-domains. First, we can see that the ranking of the different UDA methods differs to the ranking on the GTA5→Cityscapes benchmark. Both AdaptSegNet [60] and ADVENT [64] perform better with the EVA02-L encoder on the ACDC shift than DAFormer. This may be related to the FD-loss of DAFormer. In contrast, the state-of-the-art UDA method MIC [24] with the MiT-B5 backbone also shows the best performance on the ACDC shifts, mostly with a significant margin of up to 8.6% mIoU on the rain dataset and 5.8% difference for the mean. For certain cases we observe that UDA methods which perform better on the synthetic-to-real benchmark may perform worse on this benchmark. The DACS method with the EVA02-L backbone performs 5.6% mIoU worse on the synthetic-to-real benchmark than MIC while being better by 5.0% in the mean across the adverse weather conditions. However, the gain over the source only performance with the EVA02-L encoder

Enc.	UDA Method	GTA5→CS	Target Domain $\mathcal{D}^{\text{T}}$ $\mathcal{D}_{\text{fog}}^{\text{ACDC}}$	$\mathcal{D}_{\text{rain}}^{\text{ACDC}}$	$\mathcal{D}_{\text{snow}}^{\text{ACDC}}$	$\mathcal{D}_{\text{night}}^{\text{ACDC}}$	mean	$\mathcal{D}^{\text{T}}$ $\mathcal{D}_{\text{train}}^{\text{ACDC}}$
ResNet-101	Source Only	36.0	65.2	51.9	52.8	32.4	50.6	50.0
	AdaptSegNet [60]	44.0	64.4	55.6	54.5	35.8	52.6	51.6
	ADVENT [64]	43.4	64.1	53.0	54.1	34.1	51.3	52.2
	DACS [59]	55.3	68.0	**57.8**	**59.4**	37.8	**55.7**	55.6
	DAFormer [21]	**57.9**	**68.2**	54.4	58.5	**38.5**	54.9	**56.6**
MiT-B5	Source Only	47.6	70.9	63.6	61.4	36.4	58.1	59.2
	DAFormer [21]	67.6	73.5	59.2	64.5	47.1	61.1	64.2
	SePiCo [63]	67.3	76.7	62.7	63.7	47.6	62.7	66.2
	MIC [24]	**75.9**	**79.6**	**71.3**	**69.1**	**53.5**	**68.4**	**72.0**
EVA02-L	Source Only	60.5	81.4	75.6	74.0	**59.0**	72.5	74.1
	AdaptSegNet [60]	59.5	81.2	**75.8**	74.3	54.7	71.5	72.2
	ADVENT [64]	59.2	80.9	75.5	69.9	54.0	70.1	73.1
	DACS [59]	**70.3**	**82.6**	75.6	**77.7**	57.6	**73.4**	**75.6**
	DAFormer [21]	68.0	78.3	70.4	74.2	53.9	69.2	70.3

Table 3: **UDA performance from $\mathcal{D}_{\text{normal}}^{\text{ACDC}}$ to $\mathcal{D}_{\text{fog}}^{\text{ACDC}}$, $\mathcal{D}_{\text{rain}}^{\text{ACDC}}$, $\mathcal{D}_{\text{snow}}^{\text{ACDC}}$ and $\mathcal{D}_{\text{night}}^{\text{ACDC}}$.** The values for GTA5 → Cityscapes are given for comparison.

on the pure ACDC shift is limited and only DACS [59] provides a minor improvement. That shows that different UDA methods perform differently for different shifts and the 5.0% gap to the performance of MIC [24] is mainly attributed to the pre-training of the encoder. In line

with the results from [50], we observe that certain methods even show worse performances than a source-only trained model for certain domains. DAFormer [21] with MiT-B5 adapted to the rain images reaches a 4.4% mIoU lower performance. Also ADVENT [64] and AdaptSegNet [60] with a ResNet-101 encoder slightly reduce the performance on the fog domain. With the EVA02-L backbone all UDA methods except DACS [59] reach a lower performance in average and the DAFormer [21] performance drops by 3.3% mIoU compared to source only. While this may be related to the FD-loss also AdaptSegNet [60] and ADVENT [64] undergo a clear performance drop of up to 2.4% mIoU in average. The adaptation to $\mathcal{D}_{\text{train}}^{\text{ACDC}}$ mostly leads to higher performance compared to the mean of adapting to single sub-domains especially for the MiT-B5 backbone, e.g. a gain of 3.6% mIoU for MIC [24], and EVA02-L. For the ResNet-101 backbone the performance is mostly similar or even smaller. This might be caused by the different abilities of the vision transformer backbone who can utilize the larger amount of target data more effectively and benefit from the larger diversity of the target domain.

4.2 Domain generalization of UDA methods

The domain generalization performance of UDA approaches to entirely unseen domains is rarely evaluated but highly relevant because the adaptation to e.g. a certain real target domain should intuitively also improve the generalization to other unseen target domains. For this reason, we evaluate the domain generalization performance across different backbones following the same protocol as pure DG approaches [14, 26, 41] and compare it with state-of-the-art DG approaches similar to [43]. The results are shown in Table 4. Combining a simple UDA method like DACS [59] with EVA02-L improves the DG performance by 25.2% over the ResNet-101 and by 13.7% mIoU over the MiT-B5 backbone which highlights the strong generalization capabilities of vision-language pre-trained backbones. It also further improves DG performance compared to pure domain generalization methods and outperforms VLTSeg [26] by 2.1% mIoU in average across Mapillary Vistas, BDD100K and ACDC. That confirms the observation from Piva *et al*. [43] that UDA methods can provide a better generalization than DG methods. Intuitively, this can be expected since the adaptation to real images should also increase the performance on other unseen real domains since there are basic patterns which can be learned from the unlabeled real domain. The observation for the MiT-B5 backbone is similar. Recent DG methods perform similarly or outperform the generalization of several UDA methods with this backbone but cannot compete with the recent UDA approach MIC [24]. DIDEX [41] as a recent DG method outperforms both DACS [59] and DAFormer [21] by 4.9% and 3.7% mIoU in the DG mean respectively. However, MIC [24] performs 5.4% mIoU better in the DG mean. In contrast, for the ResNet-101 backbone, CLOUDS [2] outperforms DAFormer with its generalization by a large margin of 7.7% and 10.6% on BDD100K and Mapillary respectively. This is not a contradiction to the results of Piva *et al*. [43] since DG methods made a significant progress recently by e.g. using foundation models like CLOUDS [2] which enabled them to surpass the generalization of UDA methods. Notably, we observe that the UDA target domain performance on Cityscapes does not necessarily translate into a similar generalization performance. While we can see a clear performance gain of DAFormer over DACS with the MiT-B5 backbone on Cityscapes of over 7% mIoU the performance gap is reduced to 1.2% mIoU on the DG mean. For other backbones we make similar observations, like a Cityscapes ResNet-101 performance difference of 11.3% mIoU between AdaptSegNet and DACS but only 3.4% mIoU difference on the DG mean. This may be related to a method-dependent

$\mathcal{D}^S = \mathcal{D}^{GTA5}$		$\mathcal{D}^T = \mathcal{D}^{CS}_{train}$	Domain Generalization			
Enc.	**UDA/DG Method**	$\mathcal{D}^{CS}_{val}$	$\mathcal{D}^{MV}_{val}$	$\mathcal{D}^{BDD}_{val}$	$\mathcal{D}^{ACDC}_{val}$	**DG mean**
ResNet-101	FAmix [12]	49.5	52.0	46.4	**36.1**	**44.8**
	CLOUDS [1]	55.7	**59.0**	**49.3**	-	-
	VLTSeg [28]	51.2	52.2	43.3	-	-
	AdaptSegNet [60]	44.0	40.8	40.0	27.1	36.0
	ADVENT [64]	43.4	40.4	40.2	27.5	36.0
	DACS [59]	55.3	46.3	39.4	32.4	39.4
	DAFormer [21]	**57.9**	48.4	41.6	35.0	41.7
MiT-B5	DGinStyle [23]	58.6	62.5	52.3	46.1	53.6
	HRDA [25]	57.4	61.2	49.1	44.0	51.4
	CLOUDS [1]	58.1	62.3	53.8	-	-
	DIDEX [41]	62.0	63.0	54.3	50.1	55.8
	AdaptSegNet [60]	47.5	48.4	44.2	34.7	42.4
	ADVENT [64]	47.1	47.6	44.8	34.8	42.4
	DACS [59]	60.3	58.0	51.3	43.5	50.9
	DAFormer [21]	67.6	58.6	52.2	45.5	52.1
	SePiCo [67]	67.3	60.0	52.3	47.8	53.4
	MIC [24]	**75.9**	**69.3**	**57.6**	**56.8**	**61.2**
EVA02-L	VLTSeg [28]	65.6	66.5	58.4	62.6	62.5
	Rein [68]	65.3	66.1	60.4	-	-
	AdaptSegNet [60]	59.5	63.1	56.0	54.3	57.8
	ADVENT [64]	59.2	62.8	57.4	54.9	58.4
	DACS [59]	**70.3**	**68.2**	**61.2**	**64.4**	**64.6**
	DAFormer [21]	68.0	64.8	58.1	61.6	61.5

Table 4: **Domain generalization (DG) and UDA performances** on various real datasets of GTA5→Cityscapes UDA models and DG methods. The DG mean is calculated across Mapillary, BDD and ACDC. gray marks domain generalization methods which were trained on GTA5 without any adaptation. For these, values are taken from the respective publications.

overfitting of the UDA methods to the target domain which hampers the generalization of UDA methods.

We also evaluated the DG performance on the pure adverse weather shift of the ACDC dataset. We can observe from Table 5 that similar to the synthetic-to-real shift a higher performance in the target domain not necessarily causes a higher domain generalization performance. DACS [59] with ResNet-101 performs 4% mIoU better on the adverse ACDC weather domains compared to AdaptSegNet [60] but 0.3% worse on the DG mean. SePiCo [67] has a 2% mIoU higher performance on the target domain than DAFormer [21] but performs slightly worse in the DG mean. MIC [24] shows its strong generalization capabilities also in this pure real-to-real benchmark and outperforms DAFormer by a clear margin of 5.8% mIoU. Surprisingly, all four methods AdaptSegNet [60], ADVENT [64], DACS [59] and DAFormer [21] with the EVA02-L backbone outperform MIC not only for the generalization performance by up to 5.1% mIoU but also on the target domain by up to 3.6% mIoU. This may be caused by the smaller target dataset compared to the adaptation to Cityscapes which increases the influence of the pre-trained representations of the EVA02-L backbone. It also shows how different UDA methods can perform on different domain shifts with different encoders since the DG ranking for GTA5→Cityscapes is different.

4.3 Discussion

We did not apply any changes to the UDA methods like e.g. changing hyperparameters, a different resolution or disabling the FD-loss. This may have improved the performance of

$\mathcal{D}^S = \mathcal{D}^{ACDC}_{normal}$		$\mathcal{D}^T = \mathcal{D}^{ACDC}_{train}$	Domain Generalization			
Encoder	**UDA Method**	$\mathcal{D}^{ACDC}_{val}$	$\mathcal{D}^{CS}_{val}$	$\mathcal{D}^{MV}_{val}$	$\mathcal{D}^{BDD}_{val}$	**DG mean**
ResNet-101	AdaptSegNet [60]	51.6	52.5	**52.0**	42.5	49.0
	ADVENT [64]	52.2	53.1	51.6	42.9	49.2
	DACS [59]	55.6	55.0	48.9	42.1	48.7
	DAFormer [21]	**56.6**	**56.3**	51.4	**44.0**	**50.5**
MiT-B5	DAFormer [21]	64.2	65.5	59.1	49.0	57.9
	SePiCo [67]	66.2	64.7	58.6	49.7	57.7
	MIC [24]	**72.0**	**70.0**	**64.7**	**56.4**	**63.7**
EVA02-L	AdaptSegNet [60]	72.2	75.4	68.5	**61.6**	68.5
	ADVENT [64]	73.1	75.3	68.2	61.0	68.2
	DACS [59]	**75.6**	**75.8**	**68.9**	**61.6**	**68.8**
	DAFormer [21]	70.3	72.3	66.5	55.7	64.8

Table 5: **Domain generalization and UDA performances** for adaptation from ACDC clear weather reference images $\mathcal{D}^{ACDC}_{normal}$ to all adverse ACDC conditions. DG Mean is calculated over Mapillary, BDD100K and Cityscapes.

the UDA methods. However, in contrast to Englert *et al.* [12] our aim was to evaluate the UDA methods without any changes to assess how well they transfer to a different domain shift and a new encoder architecture with vision-language initialization. Modifying existing UDA methods may not be trivial because removing or adapting certain components will likely influence the performance and behavior.

5 Conclusion

We equipped existing UDA methods with a state-of-the-art vision-language pre-trained encoder and studied the target performance and the generalization to unseen domains. The results demonstrate the potential of vision-language pre-training for UDA by reaching a competitive target domain performance with a simple UDA method. They also indicate strong generalization capabilities for both established benchmarks and a pure $\rightarrow$ adverse weather condition domain shift based on ACDC. We show that recent state-of-the-art UDA methods rely on a loss function which cannot be directly used for the vision-language pre-trained encoder. Our results indicate that similar to domain generalization new UDA methods are required to fully exploit the potential of vision-language models for UDA. Our domain generalization evaluations showed two novel findings. First, the target domain performance is not necessarily an indicator for their generalization capabilities and second, that recent, pure DG methods are performing in parts similarly or even superior than UDA methods.

References

[1] Nikita Araslanov and Stefan Roth. Self-Supervised Augmentation Consistency for Adapting Semantic Segmentation. In *Proc. of CVPR*, pages 15384–15394, virtual, June 2021.

[2] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lath-

uilière. Collaborating foundation models for domain generalized semantic segmentation. In *Proceedings of the CVPR*, pages 3108–3119, 2024.

[3] Wei-Lun Chang, Hui-Po Wang, Wen-Hsiao Peng, and Wei-Chen Chiu. All about structure: Adapting structural information across domains for boosting semantic segmentation. In *Proc. of CVPR*, pages 1900–1909, 2019.

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

[5] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):834–848, April 2017.

[6] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proc. of CVPR*, pages 1791–1800, 2019.

[7] Jaehoon Choi, Taekyung Kim, and Changick Kim. Self-Ensembling With GAN-Based Data Augmentation for Domain Adaptation in Semantic Segmentation. In *Proc. of ICCV*, pages 6830–6840, Seoul, Korea, October 2019.

[8] MMSegmentation Contributors. Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark, 2020.

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

[10] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proc. of CVPR*, pages 11583–11592, 2022.

[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. In *Proc. of ICLR*, pages 1–21, May 2021.

[12] Brunó B Englert, Fabrizio J Piva, Tommie Kerssies, Daan De Geus, and Gijs Dubbelman. Exploring the benefits of vision foundation models for unsupervised domain adaptation. In *Proc. of CVPR*, pages 1172–1180, 2024.

[13] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. Pøda: Prompt-driven zero-shot domain adaptation, 2023.

[14] 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 CVPR*, pages 23428–23437, 2024.

[15] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022.

[17] Jose L Gómez, Manuel Silva, Antonio Seoane, Agnès Borrás, Mario Noriega, Germán Ros, Jose A Iglesias-Guitian, and Antonio M López. All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes. *arXiv preprint arXiv:2312.12176*, 2023.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of CVPR*, pages 770–778, 2016.

[19] Judy Hoffman, Dequan Wang, Fischer Yu, and Trevor Darrell. FCNs in the Wild: Pixel-Level Adversarial and Constraint-Based Adaptation. *arXiv*, (1612.02649), December 2016.

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

[21] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation. In *Proc. of CVPR*, pages 9924–9935, June 2022.

[22] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-Aware High-Resolution Domain-Adaptive Semantic Segmentation. In *Proc. of ECCV*, pages 372–391, October 2022.

[23] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Domain Adaptive and Generalizable Network Architectures and Training Strategies for Semantic Image Segmentation. *arXiv:2304.13615*, pages 1–15, April 2023.

[24] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: Masked Image Consistency for Context-Enhanced Domain Adaptation. In *Proc. of CVPR*, pages 11721–11732, June 2023.

[25] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *Proc. of CVPR*, pages 6891–6902, 2021.

[26] 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.

[27] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2023.

[28] Zhengkai Jiang, Yuxi Li, Ceyuan Yang, Peng Gao, Yabiao Wang, Ying Tai, and Chengjie Wang. Prototypical contrast adaptation for domain adaptive semantic segmentation. In *European conference on computer vision*, pages 36–54. Springer, 2022.

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

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[31] Kuan-Hui Lee, German Ros, Jie Li, and Adrien Gaidon. SPIGAN: Privileged Adversarial Learning from Simulation. In *Proc. of ICLR*, pages 1–14, New Orleans, LA, USA, April 2019.

[32] Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. WildNet: Learning Domain Generalized Semantic Segmentation from the Wild. In *Proc. of CVPR*, pages 9936–9946, June 2022.

[33] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6758–6767, 2019.

[34] Weizhe Liu, David Ferstl, Samuel Schulter, Lukas Zebedin, Pascal Fua, and Christian Leistner. Domain adaptation for semantic segmentation via patch-wise contrastive learning. *arXiv preprint arXiv:2104.11056*, 2021.

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. of ICLR*, 2018.

[36] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proc. of CVPR*, pages 7086–7096, 2022.

[37] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a Closer Look at Domain Shift: Category-Level Adversaries for Semantics Consistent Domain Adaptation. In *Proc. of CVPR*, pages 2507–2516, Long Beach, CA, USA, June 2019.

[38] Robert A. Marsden, Alexander Bartler, Mario Döbler, and Bin Yang. Contrastive Learning and Self-Training for Unsupervised Domain Adaptation in Semantic Segmentation. In *Proc. of IJCNN*, pages 1–8, Padua, Italy, July 2022.

[39] Luke Melas-Kyriazi and Arjun K Manrai. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proc. of CVPR*, pages 12435–12445, 2021.

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

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

[42] 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.

[43] Fabrizio J Piva, Daan De Geus, and Gijs Dubbelman. Empirical generalization study: Unsupervised domain adaptation vs. domain generalization methods for semantic segmentation in the wild. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 499–508, 2023.

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

[45] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proc. of CVPR*, pages 18082–18091, 2022.

[46] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Proc. of ECCV*, pages 102–118. Springer, 2016.

[47] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proc. of CVPR*, pages 3234–3243, 2016.

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

[49] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[50] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding. In *Proc. of ICCV*, pages 10765–10775, October 2021.

[51] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3752–3761, 2018.

[52] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Proc. of NeurIPS Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

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

[54] Manuel Schwonberg, Joshua Niemeijer, Jan-Aike Termöhlen, Jörg P. Schäfer, Nico M. Schmidt, Hanno Gottschalk, and Tim Fingscheidt. Survey on Unsupervised Domain Adaptation for Semantic Segmentation for Visual Perception in Automated Driving. *IEEE Access*, 11:54296–54336, May 2023.

[55] Fengyi Shen, Akhil Gurram, Ziyuan Liu, He Wang, and Alois Knoll. Diga: Distil to generalize and then adapt for domain adaptive semantic segmentation. In *Proc. of CVPR*, pages 15866–15877, 2023.

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[57] 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.

[58] Jan-Aike Termöhlen, Timo Bartels, and Tim Fingscheidt. A re-parameterized vision transformer (revt) for domain-generalized semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4376–4385, 2023.

[59] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: Domain Adaptation via Cross-Domain Mixed Sampling. In *Proc. of WACV*, pages 1379–1389, January 2021.

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

[61] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. Domain Adaptation for Structured Output via Discriminative Patch Representations. In *Proc. of ICCV*, pages 1456–1465, Seoul, Korea, October 2019.

[62] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1743–1751. IEEE, 2019.

[63] Vidit Vidit, Martin Engilberge, and Mathieu Salzmann. Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3219–3229, 2023.

[64] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Perez. ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation. In *Proc. of CVPR*, pages 2517–2526, June 2019.

[65] Zhonghao Wang, Mo Yu, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-Mei Hwu, Thomas S. Huang, and Honghui Shi. Differential Treatment for Stuff and Things: A Simple Unsupervised Domain Adaptation Method for Semantic Segmentation. In *Proc. of CVPR*, pages 12635–12644, June 2020.

[66] 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 CVPR*, pages 28619–28630, 2024.

[67] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. SePiCo: Semantic-Guided Pixel Contrast for Domain Adaptive Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45 (7), January 2023.

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

[69] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proc. of CVPR*, pages 2636–2645, 2020.

[70] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proc. of ICCV*, pages 2100–2110, 2019.

[71] Kai Zhang, Yifan Sun, Rui Wang, Haichang Li, and Xiaohui Hu. Multiple fusion adaptation: A strong framework for unsupervised semantic segmentation adaptation. *arXiv preprint arXiv:2112.00295*, 2021.

[72] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical Pseudo Label Denoising and Target Structure Learning for Domain Adaptive Semantic Segmentation. In *Proc. of CVPR*, pages 12414–12424, 2021.

[73] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial Style Augmentation for Domain Generalized Urban-Scene Segmentation. In *Proc. of NeurIPS*, pages 338–350, December 2022.

[74] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.

[75] 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.