Strong but simple: A Baseline for Domain Generalized Dense Perception by CLIP-based Transfer Learning

Christoph Hümmer[1,3]*, Manuel Schwonberg[1,3]*, Liangwei Zhou[2]*, Hu Cao[2], Alois Knoll[2], and Hanno Gottschalk[1]
* equal contribution, alphabetically ordered

[1] Technical University Berlin
gottschalk@math.tu-berlin.de
[2] Technical University Munich
{hu.cao,liangwei.zhou,k}@tum.de
[3] CARIAD SE
{christoph.huemmer,manuel.schwonberg}@cariad.technology

Abstract. Domain generalization (DG) remains a significant challenge for perception based on deep neural networks (DNNs), where domain shifts occur due to synthetic data, lighting, weather, or location changes. Vision-language models (VLMs) marked a large step for the generalization capabilities and have been already applied to various tasks. Very recently, first approaches utilized VLMs for domain generalized segmentation and object detection and obtained strong generalization. However, all these approaches rely on complex modules, feature augmentation frameworks or additional models. Surprisingly and in contrast to that, we found that simple fine-tuning of vision-language pre-trained models yields competitive or even stronger generalization results while being extremely simple to apply. Moreover, we found that vision-language pre-training consistently provides better generalization than the previous standard of vision-only pre-training. This challenges the standard of using ImageNet-based transfer learning for domain generalization. Fully fine-tuning a vision-language pre-trained model is capable of reaching the domain generalization SOTA when training on the synthetic GTA5 dataset. Moreover, we confirm this observation for object detection on a novel synthetic-to-real benchmark. We further obtain superior generalization capabilities by reaching **77.9%** mIoU on the popular Cityscapes→ ACDC benchmark. We also found improved in-domain generalization, leading to an improved SOTA of **86.4%** mIoU on the Cityscapes test set marking the first place on the leaderboard.

Keywords: Domain Generalization · Semantic Segmentation · Object Detection

1 Introduction

Despite the recent advancements in computer vision due to deep learning [50,74], domain shifts are still a major challenge for computer vision tasks. They can

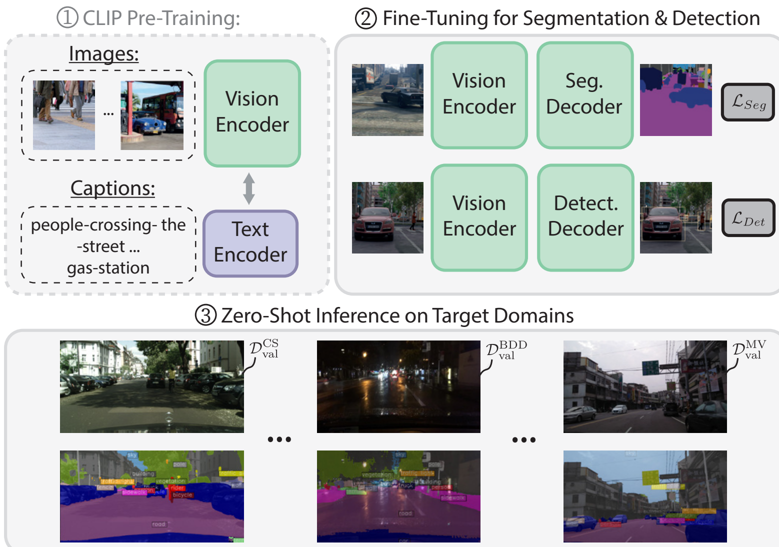

Fig. 1: CLIP-Based Fine-tuning for Domain Generalized Dense Perception: Previous works in domain generalization research mostly focus on applying methods (e.g. augmentations, consistency losses) in downstream training. In contrast, we simply transfer CLIP-based pre-trained weights (1), conduct fine-tuning on two different tasks (2) and evaluate the performance across various unseen real-world domains (3). Hereby, we reach a SOTA performance in domain generalized semantic segmentation and object detection.

cause a significant performance decrease during inference time when the distribution differs from the training distribution [37,92]. For this reason, two research fields have emerged to tackle domain shifts. Unsupervised domain adaptation (UDA) works with the assumption that unlabeled data from the target domain is available to adapt the network towards this particular domain. A large variety of methods have been developed in this field for both semantic segmentation [36–38, 40, 80, 91, 92, 110, 111] and objection detection [10, 59]. Nevertheless, unlabeled target data or information about the target domain(s) might not always be available during training time. In addition, data collection can be difficult due to the sheer size of the operational domain. For this reason, the field of domain generalization (DG) developed to improve the generalization from only a single (or multiple) source domain(s) to unseen target domains [16, 53, 64]. A variety of methods exist for semantic segmentation and several of them utilize domain randomization techniques in the input space [41,53,67,79] or constrained feature representation learning [66]. For object detection a few DG methods exist yet also utilizing consistency losses and contrastive learning [5, 93, 99]. Recently, vision-language models like CLIP [70] with their remarkable zero-shot capabilities were harnessed to improve domain generalization for street scene segmentation [3, 22, 93, 98]. However, all of these works require additional modules [98], additional models like SAM [50] and large language models [3] or feature augmentations [22,93]. Fahes *et al*. [22] show that fine-tuning with CLIP initialized weights does not work and gives a significantly worse performance compared to ImageNet. With our work, we challenge these results and present a simple yet effective solution solely based on fine-tuning to utilize the generalized vision-language representations. We achieve a competitive or higher performance for both segmentation and object detection compared to the mentioned more complex approaches. For this, we leverage CLIP [70] and EVA-CLIP [86] pre-trained ViT-encoders with a Mask2Former [15] decoder along with a simple fine-tuning protocol. Without any additional modules or losses, our technique demonstrates strong generalization capabilities on synthetic-to-real and real-to-real benchmarks. We refer to these techniques as **VLTSeg** and **VLTDet**, which is an abbreviation for **V**ision-**L**anguage representation **T**ransfer for domain generalized **Seg**mentation or **Det**ection, respectively. Unlike previous DG approaches for dense perception, fine-tuning does not rely on or require input-level augmentations [79,85,93,112,114], consistency losses [5,53,67,107] or complex whitening losses in the feature space [16, 65, 66] but offers a strong and easily applicable method. The contribution of our work is, therefore, multi-folded:

- In contrast to recent works, we show that our simple CLIP-based transfer-learning shows similar or stronger generalization capabilities without any additional modules or methods.
- We compare the generalization of state-of-the-art pre-trainings for dense perception and validate the effectiveness of vision-language pre-training.
- In synthetic-to-real DG we reach a SOTA competitive performance in semantic segmentation of 63.2 % in DG mean. We evaluate synthetic-to-real performance on a new multi-object detection benchmark based on UrbanSyn [28] with an absolute improvement ranging from 5.7 - 8.4% in mAP_{50}. For real-to-real DG, we reach an improved SOTA on Cityscapes→ACDC of 77.9% and for the in-domain Cityscapes test dataset with 86.4% mIoU.

2 Related Work

Domain Generalization Domain generalization (DG) approaches do not require any target data at all and only labeled source data. They can be divided into three major categories: distribution randomization, constrained feature learning and vision-language utilization which are related to this work. To constrain the learning to domain invariant features, instance normalization and whitening techniques are commonly used [16, 64, 65, 67, 103] with different improvements like, e.g. semantic-aware whitening [66].
In domain randomization, the visual appearance of the input is modified to learn domain-invariant features. Several of these approaches use additional external real-world data like ImageNet for style randomization, often in combination with consistency losses [48, 53, 67, 107]. FSDR [41] and PASTA [9] differ by applying style randomization and augmentation in the frequency space. WEDGE [47]

samples real images from the internet and utilizes them and self-training.
In [79], the authors show that augmentations can provide a state-of-the-art DG performance without external data. In contrast, both GBFA [85] and SHADE [112] employ feature-level augmentations. Recently, approaches tailored towards vision transformer architecture emerged in the field of DG, proposing enhanced attention mechanisms [4, 20, 84]. Parallel to our work, other works developed approaches based on vision-language models. PromptFormer, DGinStyle and DIDEX [29,45,62] all utilize generative latent diffusion models for a style transfer-based approach. Very recently, FAmix, CLOUDS and Rein [3, 22, 98] introduced methods with additional modules, feature-based augmentation and multiple different networks in addition to the vision-language trained models. In contrast to that, our simple but strong baseline does not need any additional modules, loss functions or networks but just relies on fine-tuning and achieves a competitive often even superior performance. Also, we demonstrate that our technique achieves SOTA performances for both segmentation and object detection which was only achieved by PASTA [9] before. Similar to our approach, Kerssies *et al*. [46] fine-tuned foundation models for semantic segmentation to investigate the influence of choices like the patch size on the model performance but didn't focus on domain generalization.

Pre-Training and Generalization Several works investigated the impact of supervised ImageNet pre-training on the downstream task [25,32,43,52,104,118] or up-scaling it to the larger 21k-ImageNet pre-training [73].
In recent years, self-supervised image-based pre-training approaches have emerged. Earlier approaches [12, 30, 31] already showed a strong transferability to downstream tasks. More self-supervised pre-training methods were proposed in the following namely MoCov3 [13], DINO [8], DINOv2 [63], iBOT [115], masked image modeling (MIM) [102] and BEiTv2 [68] demonstrating highly generalizing and robust representations which transfer well to downstream tasks. We exclude DinoV2 as it uses Cityscapes to sample its pre-training dataset and the knowledge of target datasets/domains is a significant advantage compared to the largely uncurated datasets SO400M and Laion-2B employed in CLIP and EVA-02-CLIP.
With the development of CLIP [70], supervised vision-language pre-training showed strong zero-shot robustness, making it a promising candidate for pre-training. Consequently, several works building upon CLIP were proposed, such as EVA-CLIP [86] further improving, upscaling, and robustifying the vision-language pre-training.

Vision-Language Models Vision Language Models (VLMs) combine vision and text features by training on large image-text pair datasets. A significant advantage of this training is the availability of image-text caption data, which can be extracted from the web and does not require manual labeling. That leads to massive datasets like the publicly available LAION-5B [78] dataset. The training methods can be divided into three main branches: The first implements contrastive learning by enforcing the similarity of corresponding text caption and image embeddings [44, 70, 81, 86, 116]. The second branch is about generative cost functions, for instance, masked image modeling and masked text modeling pre-training strategies [2, 26, 83, 96, 109], and the third branch introduces input modality adapters [1, 54, 69]. Several works adopted CLIP for tasks like object detection or semantic segmentation to profit from the classification capabilities and features. A straightforward extension is to apply CLIP to region-text pairs [82, 94, 113] to enable classification in object detection. Similarly, semantic segmentation can be conducted based on CLIP by employing segment-text alignment [34, 57, 106, 117].

3 Method

In this section, we will describe the general setting for our simple transfer learning setting and the investigation of pre-training for domain generalized dense perception.

3.1 General Setting

An input image is defined as $\mathbf{x}_n \in \mathbb{G}^{H\times W\times C}$ with $\mathbb{G}$ being the integer color values, H and W height and width respectively and C denoting the number of channels. A deep segmentation network is defined as $\mathbf{M}$ mapping the input images $\mathbf{x}_n$ to output probability maps denoted as $\mathbf{y}_n = (y_{n,i,s}) \in \mathbb{I}^{H\times W\times S}$. They represent the class posterior probabilities $y_{n,i,s} = P(s|i, \mathbf{x}_n)$ for each class $s \in \mathcal{S}$ at pixel index $i \in \mathcal{I} = \{1, 2, ..., H \cdot W\}$ with $\mathbb{I} = [0, 1]$. We furthermore define the encoder and decoder of a network $\mathbf{M}$ as $\mathbf{M}_E$ and $\mathbf{M}_D$, respectively. For the vision and language encoders we introduce the superscripts V and L respectively. We describe the vision encoder as $\mathbf{M}_E^V$ and the language encoder as $\mathbf{M}_E^L$. $\mathcal{T}$ denotes the space of text descriptions. Superscripts "S" and "T" on $\mathbf{x}_n$ and $\mathbf{y}_n$ denote the domain from which the variables stem, with e.g., $\mathcal{D}^{\mathrm{S}}$ being the source domain and $\mathcal{D}^{\mathrm{T}}$ being the target domain. Since there are multiple possible unseen target domains in domain generalization, the target domains $\mathcal{D}^{\mathrm{T}_k}$ are indexed with $k \in \mathcal{K} = \{1, 2, ..., K\}$. In domain generalization, we seek to train a model $\mathbf{M}$ which generalizes well to these unseen target domains $\mathcal{D}^{\mathrm{T}_k}$. We provide a more extensive mathematical motivation of our approach in the supplement.

One of the intriguing properties of the multimodal vision-language models like CLIP [70] is the zero-shot robustness against domain shifts. That indicates that vision-language trained feature extractors exhibit a strong semantic generalization due to the domain-invariant language modality and the massive size of the datasets. This inherent generalization may be easy to transfer to downstream tasks without complex generalization methods as previous approaches. Therefore, we aim to utilize this robustness of VLMs against domain shifts with VLTSeg and VLTDet for domain generalized dense perception.

3.2 Vision-Language based Transfer learning

In contrast to previous works, which employ complex auxiliary objectives, we investigate domain generalization by only fine-tuning strongly generalized encoders with a simple protocol. The straightforward way to employ pre-trained encoder weights for semantic segmentation or other perception tasks such as object detection is the utilization of a task-specific head and then fine-tuning the entire network $\mathbf{M}$ on the downstream task, which is often done in computer vision [8,63,102,115] and domain generalization [48,53,67,107] with additional DG methods. The encoder in our experiments $\mathbf{M}_E^V$ is initialized with either CLIP [70] or EVA-CLIP [24] pre-trained weights, which are trained on image-text pairs. We focus on transformer-based architectures ViT [21] and EVA [86] since they benefit more from large-scale pre-training and have a higher capacity than CNNs, which benefits generalization, as shown previously by [27]. Moreover, we select the recently introduced transformer-based Mask2Former [15] as the decoder for semantic segmentation. This decoder based on mask classification was robust across multiple segmentation tasks and, therefore, fits well with our target objective of domain generalization. We chose the ViTDet [56] as a simple decoder for our object detection experiments on top of the backbone. We are fully fine-tuning the entire network $\mathbf{M}$ and will validate and discuss this strategy with an ablation study of (partial) layer freezing.

3.3 Relative Performance under Domain Shift

In domain adaptation and generalization the common evaluation metric is the mean intersection over union (mIoU) or mean average precision (mAP). However, for the real-to-real domain shift evaluation (see Table 6) we analyze the performance drop from in- to out-of-domain performance. For this reason we adapt the common robustness metric relative performance under corruption (rPC) [60] since corruptions are technically one form of domain shift. Based on their definition we define our relative performance under domain shift (rPD) metric as:

$$\text{rPD} = \frac{\frac{1}{K}\sum_{k=1}^{K} \text{mIoU}^{\text{T}_k}}{\text{mIoU}^{S}} \tag{1}$$

The intuition behind the rPD is similar to the rPC. It measures the, over all target domains K averaged, relative performance drop caused by domain shift between the source domain with its oracle performance mIoU^S and the reduced mIoU^{T_k} on the unseen target domains T_k. We apply the same metric to object detection by exchanging mIoU with mAP_{50}.

4 Experimental Settings

In the following, we introduce firstly the employed datasets and metrics. Afterwards, we introduce the network architectures and implementation details.

4.1 Datasets and Metrics

Datasets: According to the common domain generalization standard setting [29, 53, 66, 85, 97] we employ the two synthetic datasets SYNTHIA (SYN) [75]

and GTA5 [72] as our source domains for our synthetic-to-real experiments. The datasets contain 9400 and 24966 images, respectively. UrbanSyn [28] with 7540 synthetic images is used for object detection. We utilize Cityscapes [18], BDD100k [105], Mapillary [61] and ACDC [77] as our real-world target domains with 500, 1000, 2000 and 406 validation images, respectively. The training sets of these datasets remain unused in the synthetic-to-real experiments. As a common practice in domain generalization [47, 84, 108], we evaluate the experiments on the validation sets $\mathcal{D}_{\text{val}}$ of the real-world target domain datasets and compute the domain generalization (DG) mean over Cityscapes, Mapillary, BDD and ACDC.
Metrics: Next to our novel rPD metric, we follow common DG research practice for the evaluation metrics [42, 51, 88]. The mean intersection over union (mIoU) of $S = 19$ (SYNTHIA trainings only $S = 16$) segmentation classes [18, 72, 77] and the mAP_{50} of $S = 8$ classes is used for the majority of experiments. In addition, the mean of mIoUs and mAP_{50} on the target datasets is computed and described as DG mean.

4.2 Data Augmentations and Baselines

We investigate a set of augmentations and domain generalization methods that were shown to be effective. The augmentations consist of PixMix [35], RandAugment [19], and PASTA [9] and are applied on every image except PixMix, which is applied randomly with a probability of 0.5. For PixMix, we employ the parameters $k = 4$ for mixing rounds and $\beta = 3$ for the mixing coefficients. For RandAugment, we only use color space augmentations which consist of Color-Transform, AutoContrast, Equalize, Sharpness, Posterize, Solarize, Color, Contrast, Brightness. Moreover, we incorporate the recent method TLDR [48], which relies on texture regularization losses, and ReVT [89], which aims to improve generalization by model averaging.

4.3 Network Architectures

Encoder We utilize vision transformer (ViT)-based backbones [21] for all our experiments but with different initialization. Our default versions are the ViT-Large encoder with patch-size 14, and the EVA-02 [86] encoder, denoted as `ViT-L-14` and `EVA-02-L-14` respectively. For our ablation studies also ViT-Base, denoted as `ViT-B-16`, is used. Note that we do not apply any own vision-language pre-training. For the synthetic-to-real experiments with a ResNet-backbone (see supplementary material), we also employed a `ResNet-101` [33] encoder.
Decoder Architecture Our standard decoder for semantic segmentation architecture is the Mask2Former [15] due to its generality. For ablation purposes, we also used the [71] FPN decoder [49], the ASPP-based DAFormer decoder [37] and the SegFormer decoder [101]. For object detection we used the ViTDet [56] decoder without the mask head. All decoders were randomly initialized in our experiments.

4.4 Implementation Details

Our experiments are implemented based on the open-source toolbox MMSegmentation [17] and MMDetection [11]. All experiments are conducted on 2 $\times$ A100 GPUs with 80GB each. For Mask2Former losses we employ $\lambda_{\mathrm{CE}} = 2.0$, $\lambda_{\mathrm{BCE}} = 5.0$ and $\lambda_{\mathrm{DICE}} = 5.0$ following the authors protocol [15]. For a fair comparison with previous works, we used two different settings for the synthetic-to-real and real-to-real experiments. The synthetic-to-real experiments for semantic segmentation were conducted with a crop size of 512 $\times$ 512 and a batch size 16. Only 5k fine-tuning iterations were necessary in this setting to avoid overfitting to the synthetic source domain. The real-to-real experiments used a larger crop

Table 1: Transfer Learning Domain Generalization Performance of supervised training on $\mathcal{D}^{\mathrm{GTA5}}_{\mathrm{train}}$ and $\mathcal{D}^{\mathrm{SYN}}_{\mathrm{train}}$.* denotes that the `ViT-L-16` was used as the encoder. ∘ denotes that `ViT-L-14` was used as the encoder. 🔒 denotes frozen encoder weights.

	Pre-Training				mIoU in %				
	Method	Data	Sup.	Self-Sup.	$\mathcal{D}^{\mathrm{CS}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{BDD}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{MV}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{ACDC}}_{\mathrm{val}}$	**DG mean**
$\mathcal{D}^S$: **GTA5**	SegFormer [101]	ImgNet-1K	✓	✗	46.6	45.6	50.1	36.4	44.7
	Supervised*	ImgNet-21K	✓	✗	49.3	47.0	52.2	43.5	48.0
	MoCov3* [13]	ImgNet-1K	✗	✓	49.7	46.2	52.4	39.1	46.9
	DeiT3* [90]	ImgNet-21K	✓	✓	53.7	52.6	59.3	49.0	53.7
	SAM* [50]	SA-1B	✓	✗	53.2	50.3	58.8	45.5	52.0
	LDM [74]	LAION-5B	✓	✗	49.2	-	-	-	-
	CLIP 🔒° [70]	WIT	✓	✗	53.2	49.8	57.1	45.0	51.2
	CLIP ° [70]	WIT	✓	✗	55.6	52.5	59.9	51.5	54.9
	EVA-02-CLIP 🔒° [86]	Merged-2B	✓	✗	55.2	51.3	57.4	47.3	52.8
	EVA-02-CLIP° [86]	Merged-2B	✓	✗	**65.3**	**58.3**	**66.0**	**62.6**	**63.1**
$\mathcal{D}^S$: **SYNTHIA**	SegFormer [101]	ImgNet-1K	✓	✗	41.1	36.2	42.4	32.6	38.1
	Supervised*	ImgNet-21K	✓	✗	44.3	37.1	43.1	34.8	39.8
	MoCov3* [13]	ImgNet-1K	✗	✓	40.2	35.4	41.5	31.7	37.2
	DeiT3* [90]	ImgNet-21K	✓	✓	47.8	39.1	45.4	34.7	41.8
	SAM* [50]	SA-1B	✓	✗	51.6	40.4	50.1	40.1	45.6
	CLIP 🔒° [70]	WIT	✓	✗	46.1	41.8	45.8	35.1	42.2
	CLIP ° [70]	WIT	✓	✗	51.1	44.7	50.6	40.7	46.8
	EVA-02-CLIP 🔒° [86]	Merged-2B	✓	✗	48.3	42.6	46.4	37.1	43.6
	EVA-02-CLIP ° [86]	Merged-2B	✓	✗	**56.8**	**51.9**	**55.1**	**48.5**	**53.1**

size of 1024 $\times$ 1024 and batch size 8 to better compare with HGFormer [20] and all trained on 20k iterations. We refer to both setups as **VLTSeg**. For object detection all trainings were conducted with a crop size of 1024 $\times$ 1024 and batch size 14. We refer to these setups as **VLTDet**. For evaluation, we always select the last checkpoint of a training in line with previous works [62, 79].

We use the same basic augmentations for all segmentation experiments: random resizing, random cropping, horizontal flip and color jitter. For object detection we use Mosaic, RandomAffine, Mixup and RandomFlip as default augmentations with RandAugment added for VLTDet. As the optimizer, we employ AdamW [58] for all our experiments as it is a common standard for training vision transformer models [15, 37, 88] with a default learning rate of 0.0001, and the backbone learning rate is established at one-tenth of the default rate for segmentation and decayed per layer with a rate of 0.7 for object detection. A learning rate schedule with a warm-up and decay was employed. More extensive implementations details are provided in the supplement.

5 Results

In the following, we first analyze the performance of our investigated transfer learning set-up and compare it with SOTA approaches afterwards. Afterwards, we analyze important fine-tuning aspects like the combination with other DG methods, layer freezing and different encoder complexities.

5.1 Pre-Training Comparison

Our fine-tuning results for semantic segmentation training on a synthetic source datasets are shown in Tab. 1 where we compare the three major pre-training paradigms: ImageNet supervised pre-training, vision self-supervised pre-training, and vision-language pre-training.
First, we observe that the recent self-supervised vision pre-trained methods provide a stronger generalization than supervised pre-training since DEiT [90] outperforms supervised pre-training by +5.7% mIoU. Surprisingly, SAM which is pre-trained on a segmentation tasks performs slightly worse than DEiT but still outperforms ImageNet supervised training. However, SAM shows a weaker generalization than vision-language pre-training.
Notably, the CLIP [70] initialization performs even stronger on the DG mean than DEiT by +1.2% mIoU, indicating that vision-language pre-training benefits from the larger pre-training datasets and the task-agnostic image-text pair training. When employing the EVA-CLIP [86] initialization, we observe a further increase of +8.2% mIoU over CLIP. Similar effects can be observed when training on SYNTHIA as the source dataset.
Vision-language pre-training datasets contain a multitude of samples compared to self-supervised vision datasets, e.g. SAM with 11 million images has only a fraction of the 2 billion image-text pairs used for EVA-CLIP [86] pre-training. We consider this an inherent advantage of vision-language pre-training since collecting these datasets is easier, and the dataset sizes are significantly larger [70]. We attribute the clear differences between CLIP and EVA02-CLIP to three factors: architecture adjustments in EVA-02-CLIP, masked image modeling initialization and the larger size of data set. Possible ablations studies on this are left for future work.

5.2 Domain Generalized Perception Benchmarks

In Table 2, we compare VLTSeg to other state-of-the-art methods for synthetic-to-real generalization. Our simple pre-training setting outperforms the best vision DG method HRDA [38] by 7.3% mIoU and most of the significantly more complex vision-language DG methods like CLOUDS [3] and performs competitive to the more complex approach Rein [98] which needs an additional network module to reach this performance. This highlights the strong generalization capabilities of vision-language pre-trained models without including any additional domain generalization methods. We also conducted experiments on SYNTHIA and with a ResNet-101 [33] backbone (see supplement), which confirms these observations.

Table 2: Domain generalization performance of state-of-the-art approaches. Training was performed on the synthetic GTA5 ($\mathcal{D}^{\mathrm{S}} = \mathcal{D}^{\mathrm{GTA5}}_{\mathrm{train}}$). Prior work results are cited from the respective paper.

	DG Method	Enc. Params	mIoU (%) on			
			$\mathcal{D}^{\mathrm{CS}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{BDD}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{MV}}_{\mathrm{val}}$	**DG mean**
Vision DG	Baseline [101]	81.4M	46.6	45.6	50.1	47.4
	ReVT [88]	81.4M	50.0	48.0	52.8	50.3
	SHADE [112]	81.4M	53.3	48.2	55.0	52.2
	IBAFormer [84]	81.4M	56.3	49.8	58.3	54.8
	CMFormer [4]	197M	55.3	49.9	60.1	55.1
	HRDA [39]	81.4M	57.4	49.1	61.2	55.9
Vision-Language DG	DGinStyle [45]	81.4M	58.6	52.3	62.5	57.8
	DIDEX [62]	81.4M	62.0	54.3	62.0	59.7
	PromptFormer [29]	-	52.0	-	-	-
	CLOUDS [3]	198M	60.2	57.4	**67.0**	61.5
	Rein [98]	307M	**65.3**	**60.5**	64.9	**63.6**
	VLTSeg	304M	**65.3**	58.3	66.0	63.2

In addition, we compare our fine-tuning scheme using CLIP [70] for ResNet and EVA-CLIP [86] for ViT to state-of-the-art DG object detection methods following the Single-DGOD [99] benchmark in Table 3. We include the two vision-language methods, CLIP The Gap [93] and PODA [23] and three methods with ImageNet initialization, namely S-DGOD [99], G-NAS [100] and PDDOC [55]. Our fine-tuning performs similar in-domain and has a clear advantage over all approaches with ImageNet initialization for out-of-domain by e.g. 3.4% mAP in average over G-NAS [100] with a ResNet-101 encoder. Even though both vision-language methods use feature space augmentations, our fine-tuning method also achieves significant gains. Compared to CLIP The Gap [93], we outperform by up to 9.2 % in-domain and out-of-domain with improvements up to 7.7 % in mAP. Moreover, VLTDet performs competitive with PODA [23] with a ResNet-101 backbone with an average of 36.9% mAP compared to 37.1% mAP for PODA. With the EVA02-CLIP backbone VLTDet reaches 39.1% mAP and outperforms all other methods. On the challenging night rainy conditions it outperforms PODA [23] as the second best approach by 4.5% mAP. To further analyze the effectiveness of vision-language pre-training we also benchmarked VLTSeg for in- and out-of-domain performance on real datasets as shown in Table 4. Vision-language pre-training outperforms the previous domain adaptation(!) SOTA significantly by +8.36% mIoU on the ACDC test set even though we are not accessing any ACDC samples during training different to CISS [76]. It also outperforms all existing DG approaches like Rein [98] which translates into first place on the

Table 3: Comparison of VLTDet on real-to-real object detection benchmark mAP_{50} **of S-DGOD [99]**

Method	Encoder	Init	Day Clear	Night Clear	Dusk Rainy	Night Rainy	Day Foggy	Average
S-DGOD [99]	R-101	IN1k	56.1	36.6	28.2	16.6	33.5	28.7
G-NAS [100]	R-101	IN1k	58.4	**45.0**	35.1	17.4	36.4	33.5
PDDOC [55]	R-101	IN1k	53.6	38.5	33.7	19.2	39.1	32.6
CLIP The GAP [93]	R-101	CLIP	51.3	36.9	32.3	18.7	38.5	31.6
PODA [23]	R-101	CLIP	-	43.4	40.2	20.5	**44.4**	37.1
VLTDet	R-101	CLIP	**60.5**	44.6	38.4	22.1	42.3	36.9
VLTDet	ViT-L-14	EVA02-CLIP	56.6	44.4	**43.6**	**26.6**	41.8	**39.1**

official leaderboard. For the in-domain performance on $\mathcal{D}_{\text{test}}^{\text{CS}}$, we obtain 86.4% mIoU, marking best performance on the official leaderboard. Note that this performance is achieved without an extended training dataset [7,87,95] and less than half the parameters and only $40k$ training iterations compared to the previous SOTA InternImage [95]. These results confirm the superior generalization capabilities of vision-language models in our simple fine-tuning approach compared to significantly more complex approaches.

Table 4: Test set Performance of VLTSeg on Cityscapes and ACDC compared with previous state-of-the-art methods. Detailed settings in the appendix.

Cityscapes ($\mathcal{D}_{\text{train}}^{\text{CS}}$) → Cityscapes ($\mathcal{D}_{\text{test}}^{\text{CS}}$)					**Cityscapes ($\mathcal{D}_{\text{train}}^{\text{CS}}$) → ACDC ($\mathcal{D}_{\text{test}}^{\text{ACDC}}$)**		
Method	**Rank**	**Params**	**Iter.**	**mIoU in %**	**Method**	**UDA/DG**	**mIoU in %**
ViT-Adapter-L [14]	5	571M	80k	85.2	HRDA [38]	UDA	67.96
InverseForm [7]	3	-	-	85.6	CISS [76]	UDA	69.55
HS3 [6]	4	-	-	85.8	PromptFormer [29]	DG	62.0
InternImage [95]	2	1.2B	80k	86.1	Rein [98]	DG	77.56
VLTSeg	1	304M	40k	**86.4**	VLTSeg (1024^2)	DG	**77.91**

5.3 Combination with existing DG methods

So far, most DG methods for segmentation utilize ImageNet pre-training. To analyze the effect of vision-language pre-training in combination with recent DG methods we tested several DG methods along with vision-language pre-training as shown in Tab. 5. For a fair comparison, we also evaluated those methods with an ImageNet pre-trained `ViT-L-16` encoder. The same evaluation is conducted for object detection to verify the applicability of the results to this task.
We can draw three major observations from this analysis. First, for both segmentation and object detection the vision-language pre-training shows a significantly better performance compared to ImageNet pre-training. Second, the additional gain by domain generalization methods for vision-language pre-training for segmentation is small, e.g. TLDR [48] adds 0.6% mIoU. However, for ImageNet pre-training this gain is mostly higher with RandAugment [19] providing 3.0% mIoU improvement on the DG mean. Third, we notice that this differs for object detection where e.g. RandAugment also provides an improvement for vision-language pre-training. We reason that the vision-language pre-training already provides highly generalized representations so existing DG methods only provide a small gain.

Table 5: Vision-language pre-training in combination with different domain generalization methods. Training was performed on the synthetic GTA5 dataset ($\mathcal{D}^{\mathrm{S}} = \mathcal{D}^{\mathrm{GTA5}}_{\mathrm{train}}$) for segmentation and UrbanSyn ($\mathcal{D}^{\mathrm{S}} = \mathcal{D}^{\mathrm{US}}_{\mathrm{train}}$) for object detection.

DG Method	Init	Segmentation mIoU (%) on					Object Detection mAP_{50} (%) on				
		$\mathcal{D}^{\mathrm{CS}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{BDD}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{MV}}_{\mathrm{val}}$	**DG mean**	Δ	$\mathcal{D}^{\mathrm{CS}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{BDD}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{ACDC}}_{\mathrm{val}}$	**DG mean**	Δ
Baseline	IN-21k	46.4	47.2	52.1	48.6	-	38.0	22.4	21.9	27.4	-
	EVA-02-CLIP	65.3	58.3	66.0	63.2	-	43.1	27.1	29.0	33.1	-
RandAugment [19]	IN-21k	50.9	49.2	54.6	51.6	+3.0	40.8	22.6	21.6	28.3	+0.9
	EVA-02-CLIP	63.5	58.0	66.0	62.5	-0.7	47.3	29.4	30.6	**35.8**	+2.7
PixMix [35]	IN-21k	49.3	46.4	54.0	49.9	+1.3	41.8	22.8	23.0	29.2	+1.8
	EVA-02-CLIP	63.1	57.7	63.8	61.5	-1.7	46.7	28.0	29.9	34.9	+1.8
ReVT [88]	IN-21k	48.2	46.6	51.9	48.9	+0.3	39.3	20.7	21.5	27.2	-0.2
	EVA-02-CLIP	63.7	60.7	66.6	63.5	+0.3	42.8	23.5	26.2	30.8	-2.3
PASTA [9]	IN-21k	49.2	49.4	53.9	50.8	+2.2	38.8	22.4	22.9	28.0	+0.6
	EVA-02-CLIP	62.9	58.4	65.8	62.4	-0.8	43.8	27.4	29.2	33.5	+0.4
TLDR [48]	IN-21k	49.2	46.9	52.2	49.4	+0.8	-	-	-	-	-
	EVA-02-CLIP	64.5	61.0	66.0	**63.8**	+0.6	-	-	-	-	-

5.4 Real-to-real domain generalization

Table 6: Real-to-real generalization for segmentation and object detection. Training and evaluation conducted on different real-world datasets. The gray boxes show the oracle performance. rPD in %.

Source	Method	Segmentation mIoU (%) on				rPD	Method	Object Detection mAP_{50} (%) on			rPD
		$\mathcal{D}^{\mathrm{CS}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{BDD}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{MV}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{ACDC}}_{\mathrm{val}}$			$\mathcal{D}^{\mathrm{CS}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{BDD}}_{\mathrm{val}}$	$\mathcal{D}^{\mathrm{ACDC}}_{\mathrm{val}}$	
$\mathcal{D}^{\mathrm{CS}}_{\mathrm{train}}$	SegFormer [101]	80.8	58.0	69.1	58.6	76.6	IN21k	52.8	30.8	35.6	62.9
	SAM [50]+M2F	82.8	57.1	71.0	60.5	75.9	SAM [50]	**61.2**	37.2	**45.1**	67.2
	VLTSeg	**85.0**	**66.0**	**77.2**	**73.3**	**84.9**	VLTDet	59.7	**39.6**	44.3	**70.2**
$\mathcal{D}^{\mathrm{BDD}}_{\mathrm{train}}$	SegFormer [101]	62.4	64.4	63.0	54.2	93.0	IN21k	45.2	51.2	35.7	79.0
	SAM [50]+M2F	68.8	69.3	69.1	57.9	94.2	SAM [50]	**47.5**	**53.8**	**37.1**	78.6
	VLTSeg	**78.9**	**72.5**	**75.9**	**71.3**	**104.0**	VLTDet	46.6	49.2	34.5	**82.4**
$\mathcal{D}^{\mathrm{ACDC}}_{\mathrm{train}}$	SegFormer [101]	66.3	54.1	64.9	75.0	82.4	IN21k	22.4	22.5	32.8	68.4
	SAM [50]+M2F	68.9	57.0	68.1	77.4	83.5	SAM [50]	33.9	25.3	47.8	61.9
	VLTSeg	**78.8**	**65.0**	**74.1**	**80.0**	**90.8**	VLTDet	**44.9**	**31.0**	**48.3**	**78.6**
$\mathcal{D}^{\mathrm{MV}}_{\mathrm{train}}$	SegFormer [101]	76.1	63.0	78.9	65.7	86.5	-	-	-	-	-
	SAM [50]+M2F	76.4	62.8	78.9	63.4	85.6	-	-	-	-	-
	VLTSeg	**81.6**	**69.2**	**83.8**	**74.2**	**89.5**	-	-	-	-	-

For real-to-real domain generalization, we designed a cross-correlation experiment as shown in Table 6. We observe that VLTSeg shows a significantly better domain generalization across all datasets, both in terms of absolute mIoU and our rPD metric. The gains are significant in comparison with the SegFormer architecture [101], which is commonly used in domain adaptation and generalization and therefore included as a baseline. When training on Cityscapes, VLTSeg improves the rPD by +8.3% and by +14.7% mIoU on ACDC. Similarly, when training on Mapillary, we improve significantly on ACDC by +8.5% mIoU.
We fine-tuned a ViT-backbone initialized with SAM [50] as it is the largest

segmentation vision pre-training dataset currently available. The generalization performance of SAM pre-trained weights is dataset-dependent but overall not significantly higher than for the ImageNet pre-trained SegFormer [101]. VLTSeg in contrast reaches a significantly better generalization than SAM, which shows that vision-language pre-training generalizes better than large-scale vision pre-training.

For real-to-real generalization for object detection we obtain similar results. For both training on Cityscapes and ACDC significant improvements of +7.3% and +10.2% rPD can be observed indicating a better generalization of vision-language pre-training. For training on BDD the improvement is smaller and ImageNet pre-training performs slightly better when evaluating on ACDC. This is most likely related to the much larger training dataset of 70k images (e.g. Cityscapes 2975 images) which causes a diminishing influence of the pre-training as previous work have already shown [32,118]. Surprisingly, the SAM pre-training [50] performs different and achieves a better absolute performance on Cityscapes and BDD but on ACDC it performs signifcantly worse than VLTDet. Also the rPD of SAM is lower for all datasets.

5.5 Analysis

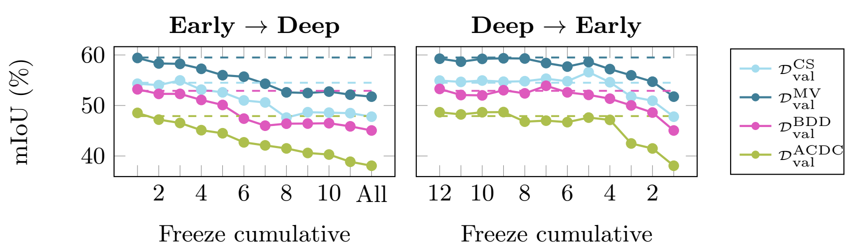

Fig. 2: Ablation of freezing encoder layers for the syn-to-real DG performance. Training on $\mathcal{D}_{\text{train}}^{\text{GTA5}}$) with `EVA-02-ViT-B-16`. Dashed lines indicate full fine-tuning.

Layer Freezing We study the effect of layer freezing in Fig. 2 to discover the best setting for fine-tuning the CLIP weights. We perform freezing cumulatively, starting either from the beginning or the end of the network. Fully fine-tuning results in the best performance while freezing the last layers performs competitively. Interestingly, only freezing the early layers degrades performance significantly indicating their importance for adapting to the perception task.

Robustness Analysis In Fig. 3, we show the EVA-CLIP and ImageNet-trained models' robustness with and without augmentations against common natural corruptions. We employ the benchmark corruptions from [60] imitating real-world shifts like blurs, noise, or weather in 5 different severities. While EVA-CLIP is more robust, augmentations improve the robustness of object detection for both ImageNet and EVA-CLIP initialization, with RandAugment being the best policy. For semantic segmentation however, the robustness is only significantly enhanced with additional methods for ImageNet pre-training.

Encoder Complexity We also researched the impact of the encoder complexity

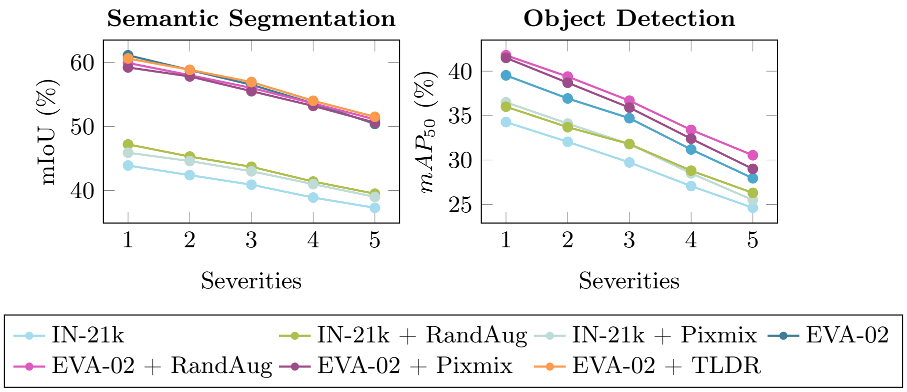

Fig. 3: Robustness over increasing severities of corruptions from [60] for segmentation and object detection. Training on $\mathcal{D}_{\mathrm{train}}^{\mathrm{GTA5}}$ with `EVA-02-L-16` and `ViT-L-16`.

on the domain generalization performance as shown in Tab. 7. We observe that for CLIP and EVA02-CLIP the generalization gets better with an increasing encoder complexity increasing by 16.5% mIoU from `EVA-02-S-16` to `EVA-02-L-14`. A larger complexity enables the encoder to better utilize the large vision-language pre-training datasets for generalized representations.

Table 7: Domain generalization of different encoder initializations and complexities. Training on GTA5. * denotes training over 20k iterations.

Init	Encoder	mIoU in %				
		$\mathcal{D}_{\mathrm{val}}^{\mathrm{CS}}$	$\mathcal{D}_{\mathrm{val}}^{\mathrm{BDD}}$	$\mathcal{D}_{\mathrm{val}}^{\mathrm{MV}}$	$\mathcal{D}_{\mathrm{val}}^{\mathrm{ACDC}}$	**DG mean**
CLIP	`ViT-B-16`	50.0	45.0	54.3	42.5	48.0
CLIP	`ViT-L-14`	55.6	52.5	59.9	51.5	54.9
EVA-02-CLIP*	`EVA-02-S-16`	45.9	47.5	51.7	41.3	46.6
EVA-02-CLIP	`EVA-02-B-16`	54.9	50.3	57.2	46.6	52.3
EVA-02-CLIP	`EVA-02-L-14`	65.3	58.3	66.0	62.6	63.1

Limitations Our standard encoder is the `EVA-02-L-14` with 304M parameters which is 3.7× the parameters of the SegFormer-MiT-B5 encoder [101] which was used for previous state-of-the-art works. Therefore, our transfer learning requires more training resources in terms of GPU memory and time and is, similar to other DG approaches, not directly applicable to real-time settings.

6 Conclusion

We demonstrate that simple fine-tuning with vision-language representations for domain generalized segmentation and object detection performs competitively or superior to much more complex approaches. Hereby, we provide a strong but simple baseline, which we extensively evaluate on various DG benchmarks. We compare SOTA vision and vision-language pre-training strategies. Our results underpin the strong vision-language generalization and evolve the established strategy for domain generalization of initializing models with ImageNet pre-training to the utilization of vision-language transfer learning.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Bińkowski, M.a., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Proc. of NeurIPS (2022) 5
2. Bao, H., Wang, W., Dong, L., Wei, F.: Vl-beit: Generative vision-language pretraining (2022) 5
3. Benigmim, Y., Roy, S., Essid, S., Kalogeiton, V., Lathuilière, S.: Collaborating foundation models for domain generalized semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3108–3119 (2024) 2, 3, 4, 10
4. Bi, Q., You, S., Gevers, T.: Learning Content-enhanced Mask Transformer for Domain Generalized Urban-Scene Segmentation. arXiv:2307.00371 pp. 1–18 (2023) 4
5. Bi, Q., Zhou, B., Yi, J., Ji, W., Zhan, H., Xia, G.S.: Good: Towards domain generalized orientated object detection. arXiv preprint arXiv:2402.12765 (2024) 2, 3
6. Borse, S., Cai, H., Zhang, Y., Porikli, F.: Hs3: Learning with proper task complexity in hierarchically supervised semantic segmentation. arXiv preprint arXiv:2111.02333 (2021) 11
7. Borse, S., Wang, Y., Zhang, Y., Porikli, F.: Inverseform: A loss function for structured boundary-aware segmentation. In: Proc. of CVPR. pp. 5901–5911 (2021) 11
8. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proc. of ICCV. pp. 9650–9660 (2021) 4, 6
9. Chattopadhyay*, P., Sarangmath*, K., Vijaykumar, V., Hoffman, J.: Pasta: Proportional amplitude spectrum training augmentation for syn-to-real domain generalization. In: Proc. of ICCV (2023) 3, 4, 7
10. Chen, C., Zheng, Z., Ding, X., Huang, Y., Dou, Q.: Harmonizing transferability and discriminability for adapting object detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8869–8878 (2020) 2
11. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019) 8
12. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: Proc. of ICML. pp. 1597–1607. PMLR (2020) 4
13. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. in 2021 ieee. In: Proc. of ICCV. pp. 9620–9629 4
14. Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. arXiv preprint arXiv:2205.08534 (2022) 11
15. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proc. of CVPR. pp. 1290–1299 (2022) 3, 6, 7, 8, 9

16. Choi, S., Jung, S., Yun, H., Kim, J.T., Kim, S., Choo, J.: RobustNet: Improving Domain Generalization in Urban-Scene Segmentation via Instance Selective Whitening. In: Proc. of CVPR. pp. 11580–11590 (Jun 2021) 2, 3
17. Contributors, M.: Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark (2020) 8
18. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding (2016) 7
19. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proc. of NeurIPS. vol. 33 (2020) 7, 11
20. Ding, J., Xue, N., Xia, G.S., Schiele, B., Dai, D.: Hgformer: Hierarchical grouping transformer for domain generalized semantic segmentation. In: Proc. of CVPR. pp. 15413–15423 (2023) 4, 8
21. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: Proc. of ICLR. pp. 1–21 (May 2021) 6, 7
22. Fahes, M., Vu, T.H., Bursuc, A., Pérez, P., de Charette, R.: A simple recipe for language-guided domain generalized segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23428–23437 (2024) 2, 3, 4
23. Fahes, M., Vu, T.H., Bursuc, A., Pérez, P., de Charette, R.: Pøda: Prompt-driven zero-shot domain adaptation (2023) 10, 11
24. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. arXiv preprint arXiv:2303.11331 (2023) 6
25. Feng, Y., Jiang, J., Tang, M., Jin, R., Gao, Y.: Rethinking supervised pre-training for better downstream transferring. In: Proc. of ICLR (2021) 4
26. Geng, X., Liu, H., Lee, L., Schuurmans, D., Levine, S., Abbeel, P.: Multimodal masked autoencoders learn transferable representations (2022) 5
27. Goldblum, M., Souri, H., Ni, R., Shu, M., Prabhu, V.U., Somepalli, G., Chattopadhyay, P., Ibrahim, M., Bardes, A., Hoffman, J., Chellappa, R., Wilson, A.G., Goldstein, T.: Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. In: Proc. of NeurIPS Datasets and Benchmarks Track (2023), https://openreview.net/forum?id=1yOnfDpkVe 6
28. Gómez, J.L., Silva, M., Seoane, A., Borrás, A., Noriega, M., Ros, G., Iglesias-Guitian, J.A., López, A.M.: All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes. arXiv preprint arXiv:2312.12176 (2023) 3, 7
29. Gong, R., Danelljan, M., Sun, H., Mangas, J.D., Van Gool, L.: Prompting Diffusion Representations for Cross-Domain Semantic Segmentation. arXiv:2307.02138 pp. 1–17 (2023) 4, 6, 11
30. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. In Proc. of NeurIPS **33**, 21271–21284 (2020) 4
31. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proc. of CVPR. pp. 9729–9738 (2020) 4
32. He, K., Girshick, R., Dollár, P.: Rethinking imagenet pre-training. In: Proc. of ICCV. pp. 4918–4927 (2019) 4, 13

33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. of CVPR. pp. 770–778 (2016) 7, 10
34. He, W., Jamonnak, S., Gou, L., Ren, L.: Clip-s4: Language-guided self-supervised semantic segmentation. In: Proc. of CVPR. pp. 11207–11216 (June 2023) 5
35. Hendrycks, D., Zou, A., Mazeika, M., Tang, L., Li, B., Song, D., Steinhardt, J.: PixMix: Dreamlike Pictures Comprehensively Improve Safety Measures. In: Proc. of CVPR. pp. 16783–16792. New Orleans, LA, USA (Jun 2022) 7
36. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A.A., Darrell, T.: CyCADA: Cycle-Consistent Adversarial Domain Adaptation. In: Proc. of ICML. pp. 1989–1998 (Jul 2018) 2
37. Hoyer, L., Dai, D., Van Gool, L.: DAFormer: Improving Network Architectures and Training Strategies for Domain-Adaptive Semantic Segmentation. In: Proc. of CVPR. pp. 9924–9935 (Jun 2022) 2, 7, 9
38. Hoyer, L., Dai, D., Van Gool, L.: HRDA: Context-Aware High-Resolution Domain-Adaptive Semantic Segmentation. In: Proc. of ECCV. pp. 372–391 (Oct 2022) 2, 10, 11
39. Hoyer, L., Dai, D., Van Gool, L.: Domain Adaptive and Generalizable Network Architectures and Training Strategies for Semantic Image Segmentation. arXiv:2304.13615 pp. 1–15 (Apr 2023)
40. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: MIC: Masked Image Consistency for Context-Enhanced Domain Adaptation. In: Proc. of CVPR. pp. 11721–11732 (Jun 2023) 2
41. Huang, J., Guan, D., Xiao, A., Lu, S.: FSDR: Frequency Space Domain Randomization for Domain Generalization. In: Proc. of CVPR. pp. 6891–6902 (Jun 2021) 2, 3
42. Huang, J., Guan, D., Xiao, A., Lu, S.: Fsdr: Frequency space domain randomization for domain generalization. In: Proc. of CVPR. pp. 6891–6902 (2021) 7
43. Huh, M., Agrawal, P., Efros, A.A.: What makes imagenet good for transfer learning? arXiv preprint arXiv:1608.08614 (2016) 4
44. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Proc. of ICML. pp. 4904–4916. PMLR (2021) 4
45. Jia, Y., Hoyer, L., Huang, S., Wang, T., Van Gool, L., Schindler, K., Obukhov, A.: Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In: Synthetic Data for Computer Vision Workshop@ CVPR 2024 (2023) 4
46. Kerssies, T., De Geus, D., Dubbelman, G.: How to benchmark vision foundation models for semantic segmentation? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1162–1171 (2024) 4
47. Kim, N., Son, T., Lan, C., Zeng, W., Kwak, S.: WEDGE: Web-Image Assisted Domain Generalization for Semantic Segmentation. arXiv:2109.14196 pp. 1–14 (Sep 2021) 3, 7
48. Kim, S., Kim, D.h., Kim, H.: Texture learning domain randomization for domain generalized segmentation. Proc. of ICCV (2023) 3, 6, 7, 11
49. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proc. of CVPR. pp. 6399–6408 (2019) 7
50. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023) 1, 3, 12, 13

51. Klingner, M., Termöhlen, J.A., Ritterbach, J., Fingscheidt, T.: Unsupervised batchnorm adaptation (ubna): A domain adaptation method for semantic segmentation without using source domain representations. In: Proc. of WACV. pp. 210–220 (2022) 7
52. Kornblith, S., Shlens, J., Le, Q.V.: Do better imagenet models transfer better? In: Proc. of CVPR. pp. 2661–2671 (2019) 4
53. Lee, S., Seong, H., Lee, S., Kim, E.: WildNet: Learning Domain Generalized Semantic Segmentation from the Wild. In: Proc. of CVPR. pp. 9936–9946 (Jun 2022) 2, 3, 6
54. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023) 5
55. Li, D., Wu, A., Wang, Y., Han, Y.: Prompt-driven dynamic object-centric learning for single domain generalization. In: Proc. of CVPR. pp. 17606–17615 (2024) 10, 11
56. Li, Y., Mao, H., Girshick, R.B., He, K.: Exploring plain vision transformer backbones for object detection. In: Proc. of ECCV (2022) 6, 7
57. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proc. of CVPR. pp. 7061–7070 (June 2023) 5
58. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. of ICLR (2018) 9
59. Mattolin, G., Zanella, L., Ricci, E., Wang, Y.: Confmix: Unsupervised domain adaptation for object detection via confidence-based mixing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 423–433 (2023) 2
60. Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A.S., Bethge, M., Brendel, W.: Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484 (2019) 6, 13, 14
61. Neuhold, G., Ollmann, T., Rota Bulo, S., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: Proc. of ICCV. pp. 4990–4999 (2017) 7
62. Niemeijer, J., Schwonberg, M., Termöhlen, J.A., Schmidt, N.M., Fingscheidt, T.: 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. pp. 2830–2840 (2024) 4, 8
63. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 4, 6
64. Pan, X., Luo, P., Shi, J., Tang, X.: Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net. In: Proc. of ECCV. pp. 464–479 (Sep 2018) 2, 3
65. Pan, X., Zhan, X., Shi, J., Tang, X., Luo, P.: Switchable Whitening for Deep Representation Learning. In: Proc. of ICCV. pp. 1863–1871 (Oct 2019) 3
66. Peng, D., Lei, Y., Hayat, M., Guo, Y., Li, W.: Semantic-Aware Domain Generalized Segmentation. In: Proc. of CVPR. pp. 2594–2605 (Jun 2022) 2, 3, 6
67. Peng, D., Lei, Y., Liu, L., Zhang, P., Liu, J.: Global and Local Texture Randomization for Synthetic-to-Real Semantic Segmentation. In IEEE TIP **30**, 6594–6608 (2021) 2, 3, 6

68. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F.: Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366 (2022) 4
69. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. ArXiv **abs/2306** (2023) 5
70. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proc. of ICML. pp. 8748–8763. PMLR (2021) 2, 3, 4, 5, 6, 9, 10
71. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proc. of CVPR. pp. 18082–18091 (2022) 7
72. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Proc. of ECCV. pp. 102–118. Springer (2016) 7
73. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972 (2021) 4
74. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2021) 1
75. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proc. of CVPR. pp. 3234–3243 (2016) 6
76. Sakaridis, C., Bruggemann, D., Yu, F., Van Gool, L.: Condition-invariant semantic segmentation. arXiv preprint arXiv:2305.17349 (2023) 10, 11
77. Sakaridis, C., Dai, D., Van Gool, L.: ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding. In: Proc. of ICCV. pp. 10765–10775 (Oct 2021) 7
78. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: Proc. of NeurIPS Datasets and Benchmarks Track (2022), https://openreview.net/forum?id=M3Y74vmsMcY 4
79. Schwonberg, M., El Bouazati, F., Schmidt, N.M., Gottschalk, H.: Augmentation-Based Domain Generalization for Semantic Segmentation. In: Proc. of IV - Workshops. pp. 1–8 (Jun 2023) 2, 3, 4, 8
80. Schwonberg, M., Niemeijer, J., Termöhlen, J.A., Schäfer, J.P., Schmidt, N.M., Gottschalk, H., Fingscheidt, T.: Survey on Unsupervised Domain Adaptation for Semantic Segmentation for Visual Perception in Automated Driving. IEEE Access **11**, 54296–54336 (May 2023) 2
81. Shen, S., Li, C., Hu, X., Xie, Y., Yang, J., Zhang, P., Rohrbach, A., Gan, Z., Wang, L., Yuan, L., et al.: K-lite: Learning transferable visual models with external knowledge. In: Proc. of NeurIPS (2022) 4
82. Shi, C., Yang, S.: Edadet: Open-vocabulary object detection using early dense alignment. In: Proc. of ICCV (October 2023) 5
83. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: FLAVA: A foundational language and vision alignment model. In: Proc. of CVPR (2022) 5
84. Sun, Q., Chen, H., Zheng, M., Wu, Z., Felsberg, M., Tang, Y.: IBAFormer: Intra-batch Attention Transformer for Domain Generalized Semantic Segmentation. arXiv:2309.06282 pp. 1–10 (Sep 2023) 4, 7

85. Sun, Q., Melnyk, P., Felsberg, M., Tang, Y.: Augment Features Beyond Color for Domain Generalized Segmentation. arXiv:2307.01703 pp. 1–10 (Jul 2023) 3, 4, 6
86. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023) 3, 4, 6, 7, 9, 10
87. Tao, A., Sapra, K., Catanzaro, B.: Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821 (2020) 11
88. Termöhlen, J.A., Bartels, T., Fingscheidt, T.: A Re-Parameterized Vision Transformer (ReVT) for Domain-Generalized Semantic Segmentation. arXiv:2308.13331 pp. 1–18 (Aug 2023) 7, 9
89. Termöhlen, J.A., Bartels, T., Fingscheidt, T.: A Re-Parameterized Vision Transformer (ReVT) for Domain-Generalized Semantic Segmentation. In: Proc. of ICCV - Workshops. pp. 1–10. Paris, France (Oct 2023) 7
90. Touvron, H., Cord, M., Jégou, H.: Deit iii: Revenge of the vit. In: Proc. of ECCV (2022) 9
91. Tranheden, W., Olsson, V., Pinto, J., Svensson, L.: DACS: Domain Adaptation via Cross-Domain Mixed Sampling. In: Proc. of WACV. pp. 1379–1389 (Jan 2021) 2
92. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to Adapt Structured Output Space for Semantic Segmentation. In: Proc. of CVPR. pp. 7472–7481 (Jun 2018) 2
93. Vidit, V., Engilberge, M., Salzmann, M.: Clip the gap: A single domain generalization approach for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3219–3229 (2023) 2, 3, 10, 11
94. Wang, T.: Learning to detect and segment for open vocabulary object detection. In: Proc. of CVPR. pp. 7051–7060 (June 2023) 5
95. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: Proc. of CVPR. pp. 14408–14419 (2023) 11
96. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., Wei, F.: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proc. of CVPR. pp. 19175–19186 (June 2023) 5
97. Wang, Z., Yu, M., Wei, Y., Feris, R., Xiong, J., Hwu, W., Huang, T.S., Shi, H.: Differential Treatment for Stuff and Things: A Simple Unsupervised Domain Adaptation Method for Semantic Segmentation. In: Proc. of CVPR. pp. 12635–12644 (Jun 2020) 6
98. Wei, Z., Chen, L., Jin, Y., Ma, X., Liu, T., Ling, P., Wang, B., Chen, H., Zheng, J.: Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28619–28630 (2024) 2, 4, 10, 11
99. Wu, A., Deng, C.: Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 847–856 (2022) 2, 10, 11
100. Wu, F., Gao, J., Hong, L., Wang, X., Zhou, C., Ye, N.: G-nas: Generalizable neural architecture search for single domain generalization object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5958–5966 (2024) 10, 11

101. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In: Proc. of NeurIPS. pp. 12077–12090 (Dec 2021) 7, 12, 13, 14
102. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proc. of CVPR. pp. 9653–9663 (2022) 4, 6
103. Xu, Q., Yao, L., Jiang, Z., Jiang, G., Chu, W., Han, W., Zhang, W., Wang, C., Tai, Y.: DIRL: Domain-Invariant Representation Learning for Generalizable Semantic Segmentation. In: Proc. of AAAI. pp. 2884–2892 (Jun 2022) 3
104. Yamada, Y., Otani, M.: Does robustness on imagenet transfer to downstream tasks? In: Proc. of CVPR. pp. 9215–9224 (2022) 4
105. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proc. of CVPR. pp. 2636–2645 (2020) 7
106. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In: Proc. of NeurIPS (2023) 5
107. Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., Gong, B.: Domain Randomization and Pyramid Consistency: Simulation-to-Real Generalization Without Accessing Target Domain Data. In: Proc. of ICCV. pp. 2100–2110 (Oct 2019) 3, 6
108. Yue, X., Zhang, Y., Zhao, S., Sangiovanni-Vincentelli, A., Keutzer, K., Gong, B.: Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In: Proc. of ICCV. pp. 2100–2110 (2019) 7
109. Zeng, Y., Zhang, X., Li, H., Wang, J., Zhang, J., Zhou, W.: X 2-vlm: All-in-one pre-trained model for vision-language tasks. arXiv preprint arXiv:2211.12402 (2022) 5
110. Zhang, K., Sun, Y., Wang, R., Li, H., Hu, X.: Multiple fusion adaptation: A strong framework for unsupervised semantic segmentation adaptation. arXiv preprint arXiv:2112.00295 (2021) 2
111. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical Pseudo Label Denoising and Target Structure Learning for Domain Adaptive Semantic Segmentation. In: Proc. of CVPR. pp. 12414–12424 (2021) 2
112. Zhao, Y., Zhong, Z., Zhao, N., Sebe, N., Lee, G.H.: Style-Hallucinated Dual Consistency Learning for Domain Generalized Semantic Segmentation. In: Proc. of ECCV. pp. 535–552 (2022) 3, 4
113. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proc. of CVPR. pp. 16793–16803 (2022) 5
114. Zhong, Z., Zhao, Y., Lee, G.H., Sebe, N.: Adversarial Style Augmentation for Domain Generalized Urban-Scene Segmentation. In: Proc. of NeurIPS. pp. 338–350 (Dec 2022) 3
115. Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., Kong, T.: ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832 (2021) 4, 6
116. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In Proc. of CVPR pp. 16795–16804 (2022), https://api.semanticscholar.org/CorpusID:247363011 4
117. Zhou, Z., Lei, Y., Zhang, B., Liu, L., Liu, Y.: Zegclip: Towards adapting clip for zero-shot semantic segmentation. In: Proc. of CVPR 5

118. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.: Rethinking pre-training and self-training. In Proc. of NeurIPS **33**, 3833–3845 (2020) 4, 13