Domain-Aware Fine-Tuning of Foundation Models

Uğur Ali Kaplan[† 1] **Margret Keuper**[2 3] **Anna Khoreva**[4] **Dan Zhang**[1 4] **Yumeng Li**[2 4]

Abstract

Foundation models (FMs) have revolutionized computer vision, enabling effective learning across different domains. However, their performance under domain shift is yet underexplored. This paper investigates the zero-shot domain adaptation potential of FMs by comparing different backbone architectures and introducing novel domain-aware components that leverage domain related textual embeddings. We propose **dom**ain adaptive **no**rmalization, termed as **Domino**, which explicitly leverages domain embeddings during fine-tuning, thus making the model domain aware. Ultimately, Domino enables more robust computer vision models that can adapt effectively to various unseen domains.

1. Introduction

Computer vision has made tremendous progress in recent years, thanks to the development of powerful neural network architectures and large-scale datasets (Deng et al., 2009; Simonyan & Zisserman, 2014; Long et al., 2015; Ronneberger et al., 2015; Hao et al., 2020; Dosovitskiy et al., 2020; Li et al., 2022; Schuhmann et al., 2022). However, when faced with domain shift - a common problem in real-world applications where the target domain has different characteristics than the training data - the performance of these models drops significantly. This is particularly problematic for tasks that require zero-shot domain adaptation, where during training there is no sample available in the target domain, though we might have an estimation of the potential domains (Farahani et al., 2021; Liu et al., 2022).

Foundation models (FMs), have become pivotal in various applications, from natural language processing to computer vision. Due to the large-scale pretraining, FMs have generalizable representations. Despite showing promising re-

---

[†] Work done at Bosch Center for AI [1]University of Tübingen [2]University of Mannheim [3]Max Planck Institute for Informatics [4]Bosch Center for Artificial Intelligence. Correspondence to: Yumeng Li <yumeng.li@de.bosch.com>.
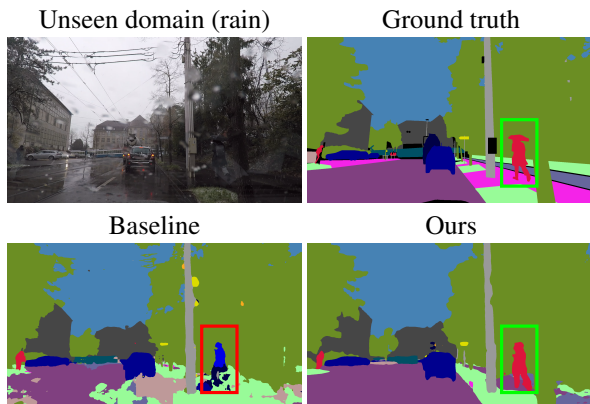
Figure 1. Traditional segmentation models trained on Cityscapes (Cordts et al., 2016) fail to make robust predictions under adverse weather conditions (Sakaridis et al., 2021), where the person is misclassified. By employing the proposed domain-aware FM fine-tuning, our model can successfully detect the person, potentially avoiding accidents.

sults on zero-shot classification tasks (Brown et al., 2020; Oquab et al., 2023), it is yet challenging to employ these models directly for dense prediction tasks, e.g., semantic segmentation (see 1st row in Table 2), necessitating the need of fine-tuning for the specific task. However, this will inevitably lead to knowledge forgetting and raise the question of how robust the adapted models are under domain shifts. Alternatively, Gong et al. (2023) proposed a prompt tuning method for test-time model adaptation.

To investigate the zero-shot domain adaptation performance of FMs, we first compare different FM backbone architectures on the challenging semantic segmentation task. We evaluate various vision backbones in Table 1, such as DINOv2 (Oquab et al., 2023; Darcet et al., 2023), ResNet-50 and ResNetRS-420 (He et al., 2016; Bello et al., 2021), CLIP-based fine-tuning methods leveraging MaskCLIP (Dong et al., 2023), and Stable Diffusion (SD) (Rombach et al., 2022) based fine-tuning, i.e., VPD (Zhao et al., 2023). We observe that these models are generally not robust to domain shifts, where there is a considerable performance drop when tested on unseen domains.

We hypothesized that visual embeddings can vary considerably with different domains, making the model vulnerable to changes such as weather and lighting conditions. To mitigate this issue, we propose to incorporate textual domain

embeddings and introduce **dom**ain adaptive **no**rmalization, termed as **Domino** during fine-tuning. We employed CLIP (Radford et al., 2021) to automatically extract the domain embedding, eliminating the need for manual definition of the domain concept. Further, we explicitly leverage the domain embedding through Domino, making the model domain aware. Domino can effectively improve the model's generalization ability across different domains without requiring data from the target domain during training (see Figure 1).

Among all FM backbones, we observe that Stable Diffusion delivers the best generalization performance. Therefore, we employ it as the default backbone to test the proposed Domino, where we observed a significant enhancement on the model generalization capability with domain-aware fine-tuning. Meanwhile, SD can not only serve as the backbone, its generative nature benefits us to explore diverse synthetic data potentially resembling the target domain, which helps to close the domain gap. We discovered that with a proper mixing ratio of real and synthesized data (see Table 3), the model's generalization performance can be further boosted.

In summary, we empirically study the generalization capability of different FM backbones. Further, we propose a domain-aware fine-tuning strategy to explicitly leverage the domain information, leading to more robust models, especially when combined with diverse synthetic data.

2. Method

2.1. Preliminary

Stable Diffusion has demonstrated astonishing text-to-image synthesis capability, thanks to their large-scale pretraining. Naturally, it has learned rich multi-modal representations. Recent work VPD (Zhao et al., 2023) has explored its potential for downstream applications, e.g., depth estimation and semantic segmentation. More specifically, they fine-tune the denoising UNet and extract intermediate features and cross-attention maps from it, which can provide semantically meaningful features (Hertz et al., 2023; Li et al., 2023a). Note that semantic classes are used as prompts for semantic segmentation task, and a text adapter is introduced as well. Further, a lightweight task-specific decoder taking extracted features is incorporated.

Despite showing promising results on the in-domain evaluation, it remains unclear how this model can perform under domain shift. In other words, it's not yet explored how the prior knowledge of powerful Stable Diffusion can help the downstream applications, which is of greater interest, and the focus of this work. Additionally, we propose a novel domain-aware fine-tuning strategy, where we extract the domain embeddings with CLIP (see Section 2.2), and incorporate them into the fine-tuning pipe to enhance domain awareness (see Section 2.3).
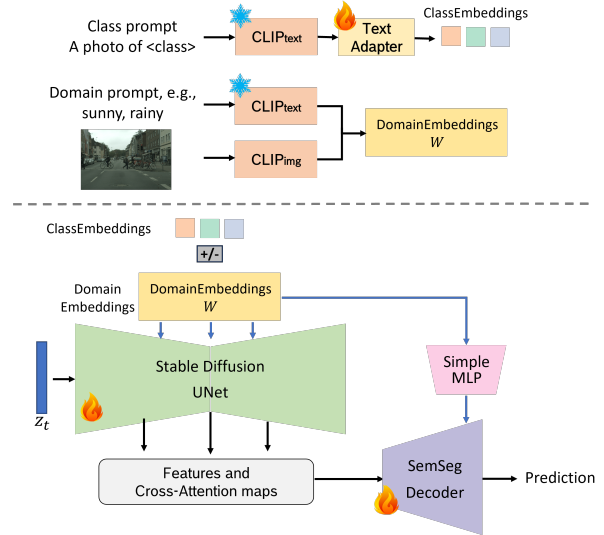

Figure 2. Trainable CLIP image encoder integrated into VPD architecture. Domain embedding is computed from input image and added/subtracted from text adapter output to manipulate domain cues. Cross-attention calculation with strengthened/weakened cues improves segmentation performance in both in-domain and domain shift scenarios.

2.2. Automatic Domain Embedding Extraction

The domain concept is quite often assigned manually. Instead, we seek a way to automatically obtain the domain information. Prior work (Wang et al., 2023) has shown that CLIP (Radford et al., 2021) is capable of assessing the look and feel of images. Thus, we propose to leverage CLIP to automatically extract the domain embeddings, which can be further utilized to enhance domain-awareness during FM fine-tuning.

To calculate the domain embedding, we begin by defining base cases that describe the aspects of interest. For instance, in autonomous driving, related domains involve different weather conditions and time of the day. Each of these domain descriptions are encoded using the CLIP text encoder:

$$d_i = \text{CLIP}_{\text{Text}}(k_i) \ \forall i = 1, \dots, N,$$

where $K = \{k_1, \dots, k_N\}$ are defined domain descriptions and $D = \{d_1, \dots, d_N\}$ is the resulting description embeddings. Given an image, we can obtain the image embedding through the CLIP image encoder and compute the similarity with each individual description embedding. Finally, the domain embedding of the given image is computed via a weighted sum of all description embeddings:

$$I = \text{CLIP}_{\text{Image}}(x)$$
$$\alpha_i = \text{Softmax}\left(\frac{I \cdot d_i}{||I|| \cdot ||d_i||}\right),$$
$$W = \sum_i \alpha_i \cdot d_i,$$

2

Table 1. Comparison of the different foundation models as backbone for semantic segmentation.

BACKBONE	PRE-TRAINING	PARAMETER SIZE	CITYSCAPES mIoU ↑	ACDC mIoU ↑	MIOU% ↑
VIT-L/14	DINOv2 WITH REGISTERS	304M	49.90	22.35	47.79
VIT-B/16	CLIP (MASKCLIP+)	147M	73.51	36.98	50.31
RESNET-101	CLIP (MASKCLIP+)	156M	73.95	35.27	47.69
RESNET-50	SUPERVISED	28.5M	69.68	38.28	54.94
RESNETRS-420 UNET	SUPERVISED	205M	73.74	50.85	68.96
VPD	STABLE DIFFUSION	868M	**76.82**	**62.50**	**81.36**

where x represents the input image, α_i and W denotes the similarity and final domain embedding, respectively. By doing so, we eliminate the need for manual domain assignment and mitigate the ill-defined one-to-one mapping issue by taking a weighted average.

2.3. Domain Aware Fine-tuning

Stable Diffusion naturally is capable of utilizing textual embedding, and thus we can simply combine the domain embedding with the original prompt embedding, i.e., class embedding, as illustrated in Figure 2. There are two possible ways of combination, namely addition or subtraction. When adding the domain embedding, we encourage the model to explore the domain information for prediction. While for subtraction, we essentially try to remove domain-related information, thus enforcing domain-invariant learning. As experimented in Table 2, we found the latter is more effective in improving the generalization performance.

In addition to being utilized by the Stable Diffusion backbone, we propose to incorporate the domain embedding in the segmentation decoder head to further enhance domain awareness. Inspired by SPADE (Park et al., 2019), we introduce **dom**ain adaptive **no**rmalization, termed as **Domino**. Specifically, we map the domain embedding through simple MLP layers into modulation parameters γ and β.

$$f_{adp} = \frac{f - \mu_f}{\gamma_f}\gamma(W) + \mu(W), \quad (1)$$

where f are intermediate features from the segmentation decoder, μ_f and γ_f denote their mean and standard deviation. $\mu(W)$ and $\gamma(W)$ are the learned affine transformation parameters, conditioned on the domain embedding W. In doing so, we effectively inject domain information into the segmentation decoder, making it domain-aware.

3. Experiments

Experimental Setup. In this study, we focus on the challenging task of semantic segmentation, which requires per-pixel semantic class prediction. We train the models on the Cityscapes (Cordts et al., 2016) training set, and evaluate the model's generalization performance on ACDC (Sakaridis et al., 2021). Cityscapes is an urban scene dataset with 19 semantic classes, which is collected mainly in Germany

under good weather conditions during daytime. In contrast, ACDC contains more adverse weather conditions, such as fog, snow, rain and nighttime. Compared to Cityscapes, there is a significant domain shift, making it a challenging test dataset meanwhile a perfect testbed for assessing the model's generalization capability.

For evaluation, we use the mean intersection-over-union (mIoU) (Everingham et al., 2015). Similar to (Fahes et al., 2023), we also report the domain adaptation performance as the percentage of target domain mIoU (ACDC) over source domain mIoU (Cityscapes):

$$\text{mIoU\%} = 100 \times \frac{\text{ACDC mIoU}}{\text{Cityscapes mIoU}}$$

We train all models on A100 GPUs using single-GPU training, using weighted cross-entropy loss. For a fair comparison, we train all models for 80,000 iterations. We use the AdamW optimizer (Loshchilov & Hutter, 2018) with a learning rate of $8e-5$ and weight decay $1e-3$. Diffusion transformer is using a different learning rate of 0.1 while the text encoder is frozen as in the VPD paper. We use poly learning rate scheduling to decrease the learning rate to 0 as training progresses.

Comparison of FM Backbones. In this comparison, we include a wide range of FMs, such as self-supervised DINOv2 (Oquab et al., 2023; Darcet et al., 2023), CLIP (Radford et al., 2021) and Stable Diffusion (Rombach et al., 2022).

Additionally, to provide a comparison against a similar model structure, we combine pre-trained ResNetRS-420's (Bello et al., 2021) in UNet form and fine-tune it on the Cityscapes dataset. We use CLIP-based methods with ViT and ResNet backbones, based on MaskCLIP+ (Dong et al., 2023).

In Table 1, we present the evaluation results of all models finetuned with cross entropy loss. We can see that VPD (built upon Stable Diffusion) performs the best on the Cityscapes validation set, as well as on the unseen ACDC, which presents a considerable domain shift. Notably, CLIP backbones and supervised models come close to the VPD performance on Cityscapes, but all of them show a significant performance degradation under domain shift. A surprising observation is the evident under-performance

Table 2. Comparison of architectural modifications. Using a frozen backbone leads to the worst results. Our proposed Domino can effectively boost the generalization performance, especially when subtracting the domain information to encourage domain-invariant learning.

VARIANT	CITYSCAPES mIoU ↑	ACDC mIoU ↑	mIoU % ↑
VPD (FROZEN)	69.19	50.48	72.96
VPD	76.87	62.61	81.45
DOMINO-ADD	**77.20**	63.53	82.29
DOMINO-SUB	76.13	**65.00**	**85.38**

Table 3. Comparison of varying mixing ratio of real and synthetic data. An even split provides the best generalization performance.

REAL / SYNTHETIC	CITYSCAPES mIoU ↑	ACDC mIoU ↑	mIoU % ↑
100/0	76.87	62.61	81.45
75/25	**77.45**	63.52	82.01
50/50	77.07	**64.49**	**83.68**
0/100	71.24	56.84	79.79

of the DINOv2 model. We attribute this to its need for a longer training time. In our experiments, we observed that while the validation performance of the other models kept increasing slowly nearing the end of our training schedule, DINOv2's validation scores kept increasing quickly, and it could have benefited from more training time. However, for a fair comparison, we have trained it using the same training schedule as the other models, which led to its underfitting on Cityscapes.

Another observation is that despite CLIP-based models also being pre-trained on large-scale image-text pairs, they have a substantial degradation on ACDC compared to Stable Diffusion. We hypothesize this is due to CLIP's tendency of catastrophic forgetting when fine-tuned fully, while being trained with a generative objective, Stable Diffusion is more tolerant in this regard. Ultimately, this validates our selection of Stable Diffusion as the FM backbone for further improvement of zero-shot domain adaptation.

Effect of Domino. In Table 2, we present the comparison of the proposed domain-aware fine-tuning Domino with the original VPD (Zhao et al., 2023). We first examine the impact of architectural changes on performance, taking into account prior work suggesting that some foundational models perform better with frozen weights. To explore this phenomenon, we experimented using a frozen Stable Diffusion backbone (1st row in Table 2), which has the worst performance. This indicates the necessity of fine-tuning the entire backbone for the specific downstream task. We implemented both domain addition and subtraction, denoted as Domino-Add and Domino-Sub in Table 2. We observe that both variants outperform the baseline VPD in generalization

performance on ACDC, which indicates the effectiveness of domain-aware fine-tuning. Notably, the Domino-Sub version achieves the best generalization results. This suggests that by encouraging domain-invariant representation learning during training, the model becomes more robust under domain shifts. We also see that Domino-Add can improve in-domain performance. We hypothesize that encouraging the usage of domain embedding can ease pattern learning as Cityscapes mostly consist of clear day images.

Effect of Synthetic Data. We investigate the impact of incorporating synthetic data into our framework for zero-shot domain adaptation. Prior work has demonstrated that utilizing synthetic data during training can enhance performance for traditional models (Beery et al., 2020; Li et al., 2023b; 2024; Wang et al., 2024). However, the effectiveness of synthetic data with Foundation Models (FMs) is a topic of discussion, particularly as these models have been trained on massive real data. Yet, we hypothesized for improving domain generalization, it is crucial for the model to see diverse data from different domains, e.g., under different weather conditions. To explore this, we employ ALDM (Li et al., 2024) to generate synthetic data based on the labels of Cityscapes training set and diverse prompts.

In Table 3, we present the results of fine-tuning the base VPD model with varying ratios between real and synthetic data. Notably, combining synthetic data with real data leads to improved performance in both domains. As we increase the proportion of synthetic data, target domain performance also increases. However, when using synthetic data only, we observe a performance decrease in both domains. This finding suggests that relying solely on synthetic data is not beneficial, as the quality of synthetic might not be as good as real ones. This also highlights the importance of incorporating real data alongside synthetic data for foundational models during fine-tuning.

4. Conclusion & Discussion

In this study, we conducted extensive experiments comparing different foundation models as the backbone for zero-shot domain adaptation in semantic segmentation. We empirically observe that Stable Diffusion based VPD model achieves better generalization performance. We then demonstrated the proposed Domino with explicit usage of the domain information can significantly boost the model's generalization further. Our results indicate that domain embedding addition encourages the use of domain cues, which can be beneficial for improving the in-domain segmentation performance. In contrast, domain embedding subtraction encourages the use of more domain-invariant features which can enhance the generalization performance. Furthermore, we have found that incorporating synthetic data with a proper ratio during the foundational model fine-tuning is also beneficial for domain generalization performance.

Acknowledgement

References

Beery, S., Liu, Y., Morris, D., Piavis, J., Kapoor, A., Joshi, N., Meister, M., and Perona, P. Synthetic examples improve generalization for rare classes. In *WACV*, 2020.

Bello, I., Fedus, W., Du, X., Cubuk, E. D., Srinivas, A., Lin, T.-Y., Shlens, J., and Zoph, B. Revisiting resnets: Improved training and scaling strategies. *NeurIPS*, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *NeurIPS*, 2020.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. In *ICLR*, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *CVPR*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015.

Fahes, M., Vu, T.-H., Bursuc, A., Pérez, P., and De Charette, R. Poda: Prompt-driven zero-shot domain adaptation. In *ICCV*, 2023.

Farahani, A., Voghoei, S., Rasheed, K., and Arabnia, H. R. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, 2021.

Gong, R., Danelljan, M., Sun, H., Mangas, J. D., and Van Gool, L. Prompting diffusion representations for cross-domain semantic segmentation. *arXiv preprint arXiv:2307.02138*, 2023.

Hao, S., Zhou, Y., and Guo, Y. A brief survey on semantic segmentation with deep learning. *Neurocomputing*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-or, D. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023.

Li, Y., Mao, H., Girshick, R., and He, K. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022.

Li, Y., Keuper, M., Zhang, D., and Khoreva, A. Divide & bind your attention for improved generative semantic nursing. In *BMVC*, 2023a.

Li, Y., Zhang, D., Keuper, M., and Khoreva, A. Intra-& extra-source exemplar-based style synthesis for improved domain generalization. *IJCV*, 2023b.

Li, Y., Keuper, M., Zhang, D., and Khoreva, A. Adversarial supervision makes layout-to-image diffusion models thrive. In *ICLR*, 2024.

Liu, X., Yoo, C., Xing, F., Oh, H., El Fakhri, G., Kang, J.-W., Woo, J., et al. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*, 2022.

Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *ICLR*, 2018.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.

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 *ICML*, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

Sakaridis, C., Dai, D., and Van Gool, L. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, October 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Wang, J., Chan, K. C., and Loy, C. C. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.

Wang, J., Laube, K. A., Li, Y., Metzen, J. H., Cheng, S.-I., Borges, J., and Khoreva, A. Label-free neural semantic image synthesis. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2024.

Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., and Lu, J. Unleashing text-to-image diffusion models for visual perception. *ICCV*, 2023.