Unified Language-driven Zero-shot Domain Adaptation

Senqiao Yang[1,2] Zhuotao Tian[2*] Li Jiang[3] Jiaya Jia[1]
[1]The Chinese University of Hong Kong
[2]Harbin Institute of Technology, Shenzhen [3]The Chinese University of Hong Kong, Shenzhen

Overview

A. ULDA on Autonomous Driving

A.1. Motivation

Autonomous driving has wide applications for intelligent transportation systems, such as reducing the labor costs, enhancing the comfortableness of customers, and so on [14, 60]. In some cases, the autonomous vehicle might work in adverse weather conditions, like night, rain, fog, and so on. These complex scenarios might take a big challenge to the autonomous driving system.

However, in practical terms, it is not always feasible to obtain comprehensive data for every possible adverse condition due to the high costs and difficulties associated with data collection. Instead, practitioners may only have a conceptual understanding or hypothetical descriptions of potential driving scenarios. In this case, the ability to augment a model's performance in such predicted scenarios without actual data collection is preferred.

Therefore, our proposed Unified Language-driven Zero-shot Domain Adaptation (ULDA) holds great potential in autonomous driving scenarios and could substantially enhance the future advancement of autonomous driving technology.

A.2. Proposed Benchmark

To further explore the potentially challenging autonomous driving scenarios, we use GPT-4 to generate several difficult situations. Our prompt is "Please describe some potential adverse driving scenarios, which pose challenges for autonomous driving, such as 'Driving in fog', 'Driving in snow'." After careful consideration of the comprehensive answers provided by GPT-4 and the scene understanding capabilities of CLIP, we choose 'sandstorm' and 'fire' as additional scenarios to augment the existing autonomous driving scenes (rain, snow, fog, night). These scenarios have been selected based on their challenging nature and the relatively limited availability of relevant data.

Due to the rarity of these scenarios, collecting images for them has been challenging. We make significant efforts to gather several images that fulfill the requirements for autonomous driving from publicly accessible websites with copyright permissions. These images have been annotated and will serve as a new benchmark. All collected data will be released.

A.3. Results

We choose Cityscapes [5] as the source domain and ACDC [42] with Fog, Night, Rain, Snow, as well as our collected Sandstorm and Fire as the target domain. All other model implementation details remain consistent with the main experiments. In order to demonstrate the effectiveness of our approach, we train the state-of-the-art (SOTA) method, PØDA [6], with six distinct segmentation heads for specific domains. In contrast, our model is exclusively trained with a unified, all-in-one head.

As illustrated in Table 1, our proposed method consistently outperforms all previous models in the Autonomous Driving scenario. Our method achieves improvements of 8.11% and 1.91% over the baseline source model and the former state-of-the-art method, PØDA*, respectively. Notably, our approach, employing a single head, even surpasses the performance of PØDA* and achieves a 1.82% mIoU improvement, which requires training separate heads

*Corresponding to: tianzhuotao@gmail.com

Scenarios	Source2Fog	Source2Night	Source2Rain	Source2Snow	Source2Sandstorm	Source2Fire	mean-mIoU
Domain Description	driving in fog	driving at night	driving under rain	driving in snow	driving in sandstorm	driving through fire	
Source	49.98	18.31	38.20	39.28	19.58	10.08	27.57
CLIPStyler	48.87	20.83	36.97	40.31	23.16	12.36	30.42
PØDA*	51.54	**25.03**	42.31	43.90	24.39	15.43	33.77
ULDA	**53.02**	24.61	**45.12**	**46.06**	**25.72**	**19.52**	**35.68**

Table 1. **Performance of ULDA on Autonomous Driving.** We use Cityscape as the source domain and ACDC and our collected data as the six target domains in this setting. Mean-mIoU represents the average mIoU value in six scenarios. PØDA* represents the model that uses different segmentation heads in specific domains with domain-id provided, while our ULDA utilizes a unified, all-in-one head.

for different scenarios. Furthermore, in challenging scenarios like 'driving through fire' and 'driving under snow,' the backgrounds are almost red and white, respectively. Our method demonstrates a significant improvement compared to the previous approach, increasing by 4.09% and 2.16%, respectively. These results demonstrate that contributed to Hierarchical Context modeling and Text-Driven Rectifier, our method can adeptly and precisely extract and utilize the multi-level correlation between images and text, thereby achieving significant improvement in complex scenarios. Additionally, our proposed domain-consistent representation learning ensures consistency across diverse domains, enabling our model to generalize effectively under a single unified segmentation head. What's more, we present the qualitative analysis in Sec. F.

B. Additional Experiments

In this section, we demonstrate the effectiveness of our method by comparing it to the One-shot Domain Adaptation method in Sec. B.1. Furthermore, we demonstrate that our method can achieve further improvement based on the domain generalization method in Sec. B.2.

B.1. Effectiveness Compared One-shot Domain Adaptation

To show the effectiveness of our ULDA, We evaluate it against SM-PPM [58][1], a SOTA method in one-shot unsupervised domain adaptation (OSUDA). The OSUDA setting allows access to a single unlabeled target domain image for adapting the model to the new target domain. In SM-PPM, this image acts as an anchor for mining target styles. For a robust comparison, we adhere to the previous settings. We employ five randomly selected target images to train the SM-PPM. Additionally, we train five different models for each image, each initialized with a unique random seed. The mean Intersection over Union (mIoU) values reported represent the average across these 25 models. It's important to note that a direct comparison of the absolute results between the two models may not be entirely fair due to the differences in their backbones (ResNet-101 in SM-PPM versus ResNet-50 in ULDA) and segmentation frameworks (DeepLabv2 in SM-PPM versus DeepLabv3+ in ULDA). Therefore, our analysis focuses on the improvement each method offers over its respective naive source-only baseline, also considering the baseline's performance. As shown in Table 2, in the Cityscapes→ACDC scenario, both the absolute and relative improvements of ULDA over its source-only version surpass those of SM-PPM.

Notably, the SM-PPM utilizes three different images to adapt three domain-specific models to the three scenarios on ACDC. However, our ULDA does not have access to any target domain image and utilizes one unified model to adapt to any scenario.

[1]We use official code https://github.com/W-zx-Y/SM-PPM

Source	Target eval.	One-shot SM-PPM [58]	Zero-shot ULDA
CS	ACDC Night	13.07 / 14.60 (Δ=1.53)	18.31 / **25.40** (Δ=7.09)
	ACDC Snow	32.60 / 35.61 (Δ=3.01)	39.28 / **46.00** (Δ=6.72)
	ACDC Rain	29.78 / 32.23 (Δ=2.45)	38.20 / **44.94** (Δ=6.74)
GTA5	CS	36.60 / 42.80 (Δ=6.20)	36.38 / **42.91** (Δ=6.53)

Table 2. **Effectiveness compared to OSUDA.** Semantic segmentation performance (mIoU%) for source / adapted models, and gain provided by adaptation (Δ in mIoU). For adaptation, SM-PPM has access to one target image and adapts three specific models on ACDC, while ULDA not has access to any target domain image and utilizes one unified model.

Method	Fog	Night	Snow	Rain	Mean
Source	49.98	18.31	39.28	38.20	36.44
Source-G	51.48	21.07	42.84	42.38	39.69
PØDA*	51.54	25.03	42.31	43.90	40.65
PØDA*-G	52.87	24.86	44.34	43.17	41.31
ULDA	53.55	25.40	44.94	46.00	42.47
ULDA-G	54.21	25.94	46.02	47.15	43.33

Table 3. **Effectiveness with DG method.** '-G' means the source model is trained with the domain generalization method [7]. Source-only-G model is enhanced with a domain generalization technique. PØDA* represents the model that uses different segmentation heads in specific domains with domain-id provided.

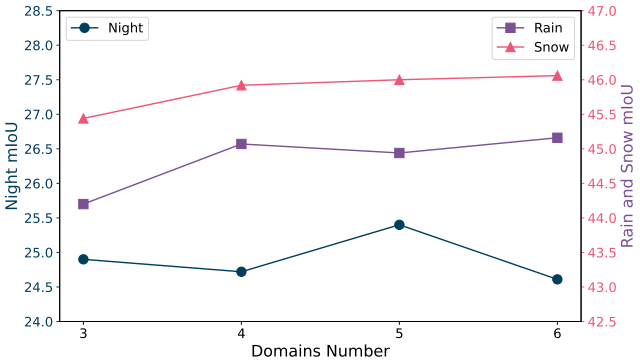

Figure 1. **Ablation Study.** The impact of the domain numbers on the model's performance.

B.2. Effectiveness on Domain Generalization

Domain generalization (DG) is a setting that aims to develop robust models that can generalize well to new, unseen domains. [7] is a current SOTA method in Domain generalization by simply perturbing feature channel statistics. Therefore, we aim to demonstrate that by incorporating the DG method, our approach can achieve further improvement. We showcase our effectiveness across four target domain scenarios in Cityscapes-ACDC. The detailed setting is followed by PØDA [6] Table 7.

First, we follow the DG sota method [7] to train the Source-G model, which augment features by shifting the per-channel $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ with Gaussian noises sampled for each batch of features. As shown in Table 3, the Source-only-G consistently outperforms the Source-only model, demonstrating a generalization capability in the Semantic Segmentation task. Moreover, when integrating the Domain Generalization technique into our proposed ULDA, significant enhancements are observed across all target domains, leading to further improvements in the mean-mIoU metric. Notably, in comparison to the PØDA*-G method, our ULDA-G approach outperforms it with a mean-mIoU improvement of 2.02%. Moreover, our ULDA method eliminates the need for Domain-ID and utilizes a single model for all scenarios.

These experiments on Domain Generalization also illustrate that our proposed ULDA setting is not a degraded version of domain generalization. Additionally, these two settings can mutually benefit each other, as discussed in Sec. 6 of the main paper.

B.3. Additional Ablation Study

Our proposed ULDA approach employs a single unified all-in-one segmentation head to adapt to various target domains. Consequently, a natural question arises: *how does the number of domains impact the performance of the model?* Therefore, we conduct a comparative analysis of the changes in mIoU across three scenarios, namely Night, Rain, and Snow, while varying the number of domains from 3 to 6. Specifically, the four-domain setting includes Night, Rain, Snow, and the addition of Fog (ACDC); the five-domain setting comprises Night, Rain, Snow, the addition of Fog (ACDC), and GTA5; the six-domain setting involves Night, Rain, Snow, the addition of Fog (ACDC), and our collected Sandstorm and Fire.

As shown in Fig. 1, the mIoU for rainy and snowy scenes exhibits an upward trend with an increase in the number of domains. And nighttime scenes exhibit a fluctuating pattern. This result demonstrates that our proposed Domain Consistent Representation Learning plays a crucial role in maintaining consistent performance across domains, preventing any decline in performance. The gradual improvement in mIoU for rainy and snowy scenarios can be attributed to the increased exposure to diverse domains, allowing the model to acquire more generalized knowledge and enhance its performance. However, the nighttime scenario, which significantly differs from daytime scenarios, faces challenges in extracting relevant knowledge from other scenarios. Nonetheless, even with an increasing number of domains, our method consistently maintains its performance without any decline.

C. Implementation Details

In this study, we follow the implementation details from the previous work, PØDA [6]. Specifically, we utilize the DeepLabv3+ framework [4] incorporating a backbone model of pre-trained CLIP-ResNet-50[2]. For the source domain, the model is trained for 200,000 iterations using randomly cropped 768x768 images. Training is performed with a polynomial learning rate schedule, starting at $lr = 10^-1$ for the classifier, and employing Stochastic Gradient Descent [2] with a momentum of 0.9 and weight decay of 10^-4. Standard color jittering and horizontal flip augmentations are applied to these crops.

During Stage-1, where the PIN is trained to simulate the target domain feature, we make use of the source feature maps after the first layer. The style parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are represented as 256-dimensional real vectors. The CLIP embeddings are 1024D vectors. To encode the target descriptions, we adapt the ImageNet templates from [41] and use them in the encoding process of the TrgPrompt.

In the Fine-tuning stage (Stage 2), we start with the pre-trained model on the source domain and focus on fine-tuning the classifier head. This process involves working with augmented PIN features, denoted as $\bar{\mathbf{f}}_{\mathrm{s}\to\mathrm{t}}$, and continuing the process for 2,000 iterations.

To evaluate the adaptation performance, we mainly utilize the mean Intersection over Union (mIoU%) metric. This metric allows us to assess the performance of the models on target images at their original resolutions.

[2] https://github.com/openai/CLIP

D. Additional Discussions

What's the difference between our and previous settings? To facilitate a more comprehensive comparison between our proposed setting and the previous settings, we further analyze the target data format and the number of models required for N target domains. As shown in Table 4, Standard Unsupervised Domain Adaptation (UDA), One-Shot Unsupervised Domain Adaptation (OSUDA), Few-Shot Unsupervised Domain Adaptation (FSUDA) all require access to the target domain image. In contrast, Prompt-driven Zero-shot Domain Adaptation (PØDA) and our proposed Unified Language-driven Zero-shot Domain Adaptation (ULDA) only need to access the target domain language description to extract the target domain knowledge, which is practical and less costly. Besides, compared to the previous settings, only ULDA requires a single model to adapt to diverse target domains without domain-IDs, instead of using domain-specific heads as in previous methods.

Why not incorporate TDR to Stage-1? We present a detailed derivation regarding the question, "Why not incorporate TDR into Stage-1?" as discussed in Section 6 of the main paper here.

For the original source domain feature, we can obtain the corresponding target domain feature $\mathbf{f}_{\mathrm{s}\to\mathrm{t}}$ through the following formula (Eq. (1) in the main paper):

$$\mathbf{f}_{\mathrm{s}\to\mathrm{t}} = \mathrm{PIN}\left(\mathbf{f}_{\mathrm{s}}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right) = \boldsymbol{\sigma}\left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right) + \boldsymbol{\mu}.$$

This normalization process transfers features from the source domain to the distribution of the target domain.

For $\mathbf{f}_{\mathrm{s}\to\mathrm{t}}$, we know $\dfrac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}$ theoretically follows the distribution $\mathcal{N}(0,1)$, thus $\mathbb{E}[\dfrac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}] = 0$, $\mathrm{Var}(\dfrac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}) = 1$. therefore, we could calculate the $\mathbb{E}(\mathbf{f}_{\mathrm{s}\to\mathrm{t}})$ and $Var(\mathbf{f}_{\mathrm{s}\to\mathrm{t}})$ as:

$$\begin{aligned}
\mathbb{E}\left(\mathbf{f}_{\mathrm{s}\to\mathrm{t}}\right) &= \mathbb{E}\left[\boldsymbol{\sigma}\left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right) + \boldsymbol{\mu}\right] \\
&= \mathbb{E}(\boldsymbol{\mu}) + \boldsymbol{\sigma}\left[\left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right)\right] \\
&= \boldsymbol{\mu} + \boldsymbol{\sigma}\cdot 0 \\
&= \boldsymbol{\mu}.
\end{aligned} \tag{1}$$

$$\begin{aligned}
\mathrm{Var}\left(\mathbf{f}_{\mathrm{s}\to\mathrm{t}}\right) &= \mathrm{Var}\left[\boldsymbol{\sigma}\left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right) + \boldsymbol{\mu}\right] \\
&= \mathrm{Var}\left[\boldsymbol{\sigma}\left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right)\right] \\
&= \boldsymbol{\sigma}^2\,\mathrm{Var}\left[\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right] \\
&= \boldsymbol{\sigma}^2
\end{aligned} \tag{2}$$

Hence, for the simulated $\mathbf{f}_{\mathrm{s}\to\mathrm{t}}$, we have $std(\mathbf{f}_{\mathrm{s}\to\mathrm{t}}) = \boldsymbol{\sigma}$, $mean(\mathbf{f}_{\mathrm{s}\to\mathrm{t}}) = \boldsymbol{\mu}$. Substituting them into main paper's Eq. (10) yields:

$$\begin{aligned}
\widetilde{\mathbf{f}}_{\mathrm{s}\to\mathrm{t}} &= \beta\left(\widetilde{\boldsymbol{\sigma}}\left(\frac{\mathbf{f}_{\mathrm{s}\to\mathrm{t}} - \mu\left(\mathbf{f}_{\mathrm{s}\to\mathrm{t}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}\to\mathrm{t}}\right)}\right) + \widetilde{\boldsymbol{\mu}}\right) + \mathbf{f}_{\mathrm{s}\to\mathrm{t}} \\
&= \beta\left(\widetilde{\boldsymbol{\sigma}}\left(\frac{\mathbf{f}_{\mathrm{s}\to\mathrm{t}} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}\right) + \widetilde{\boldsymbol{\mu}}\right) + \boldsymbol{\mu} + \boldsymbol{\sigma}\left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right). \\
&= \beta\left(\widetilde{\boldsymbol{\sigma}}\left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right) + \widetilde{\boldsymbol{\mu}}\right) + \boldsymbol{\mu} + \boldsymbol{\sigma}\left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right). \\
&= \left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right)(\beta\widetilde{\boldsymbol{\sigma}} + \boldsymbol{\sigma}) + (\beta\widetilde{\boldsymbol{\mu}} + \boldsymbol{\mu}).
\end{aligned} \tag{3}$$

As discussed in the main paper Sec. 6, $\widetilde{\boldsymbol{\sigma}}$ and $\widetilde{\boldsymbol{\mu}}$ are derived by passing text embeddings through a linear layer. The parameters $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$ are learnable and are designed to simulate features of the target domain. During Stage-1, it is necessary to optimize $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ to transform the source domain features into those of the target domain, ensuring alignment with the text embeddings. However, as the text embeddings are directly input into the linear layer to obtain $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\sigma}}$, this process results in $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ not being optimized, leading to a trivial solution. Therefore, we may not integrate rectification into Stage-1.

E. Related Work

Unsupervised Domain Adaptation (UDA) In UDA, a model trained on a labeled source domain is adapted to an unlabeled target domain. The majority of the approaches rely on discrepancy minimization [33, 34], adversarial training [8, 50] and self-training [25, 69]. These techniques primarily focus on reducing the domain gap at different levels: input [9, 63], features [30, 43, 55, 64], or output [50, 51]. However, in real-world scenarios, obtaining a substantial number of target domain images can be challenging, leading to the development of various domain adaptation settings [6, 31, 32, 37, 38, 52, 53, 62].

Recently, one challenging setting of One-Shot Unsupervised Domain Adaptation (OSUDA) has been proposed. This setting requires models to adapt to a target domain with access to only one image from that domain. To the best of our knowledge, only three studies focusing on semantic segmentation within this context have been docu-

Setting	Target Data	Domain-ID	Model Number
Standard Unsupervised Domain Adaptation	Image	Require	N
One-Shot Unsupervised Domain Adaptation [36]	Image	Require	N
Few-Shot Unsupervised Domain Adaptation [15]	Image	Require	N
Prompt-driven Zero-shot Domain Adaptation [6]	Scenario Description	Require	N
Unified Language-driven Zero-shot Domain Adaptation	Scenario Description	No Require	1

Table 4. The difference between our proposed Unified Language-driven Zero-shot Domain Adaptation and related adaptation settings. Target Data means the form of the target domain data. Domain-ID indicates whether the model requires the domain ID during testing. Model Number means for N target domains, the required segmentation head's Number.

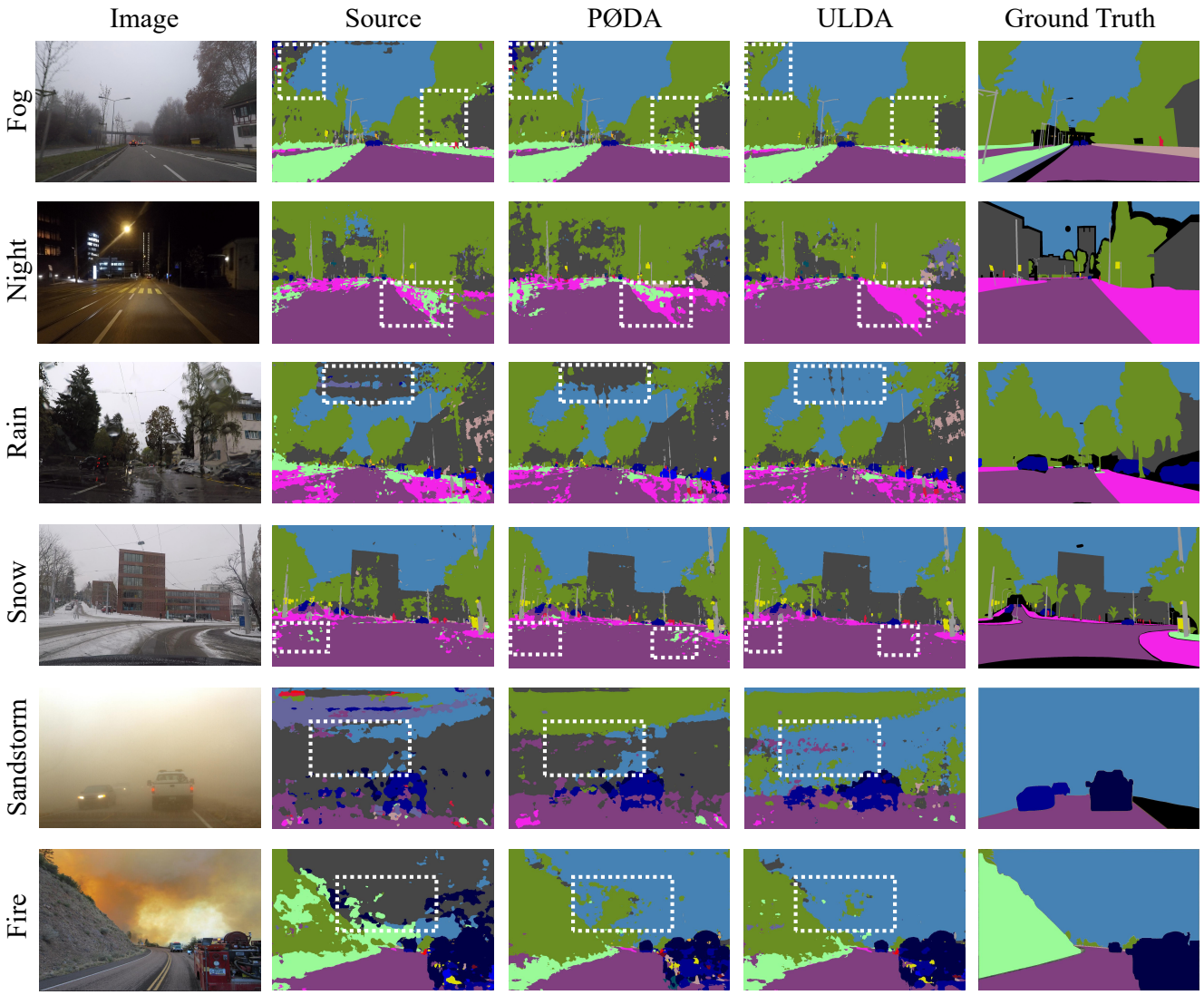

Figure 2. **Qualitative Analysis.** We compare the qualitative results of Autonomous Driving scenarios. Only our method employs a single segmentation head across all scenarios, other methods train a segmentation head specifically for each domain.

mented [1, 36, 58]. Luo et al. [36] highlight the limitations of traditional UDA methods when limited to a single unlabeled target image. They propose a style mining algorithm that combines a stylized image generator with a task-specific module to prevent overfitting. In contrast, Wu et al. [58] introduce a novel approach named style mixing and patch-wise prototypical matching (SM-PPM). This method involves blending the features of a source image with those of the target linearly, and employing patch-wise prototypical matching to mitigate negative adaptation [21]. Benig-

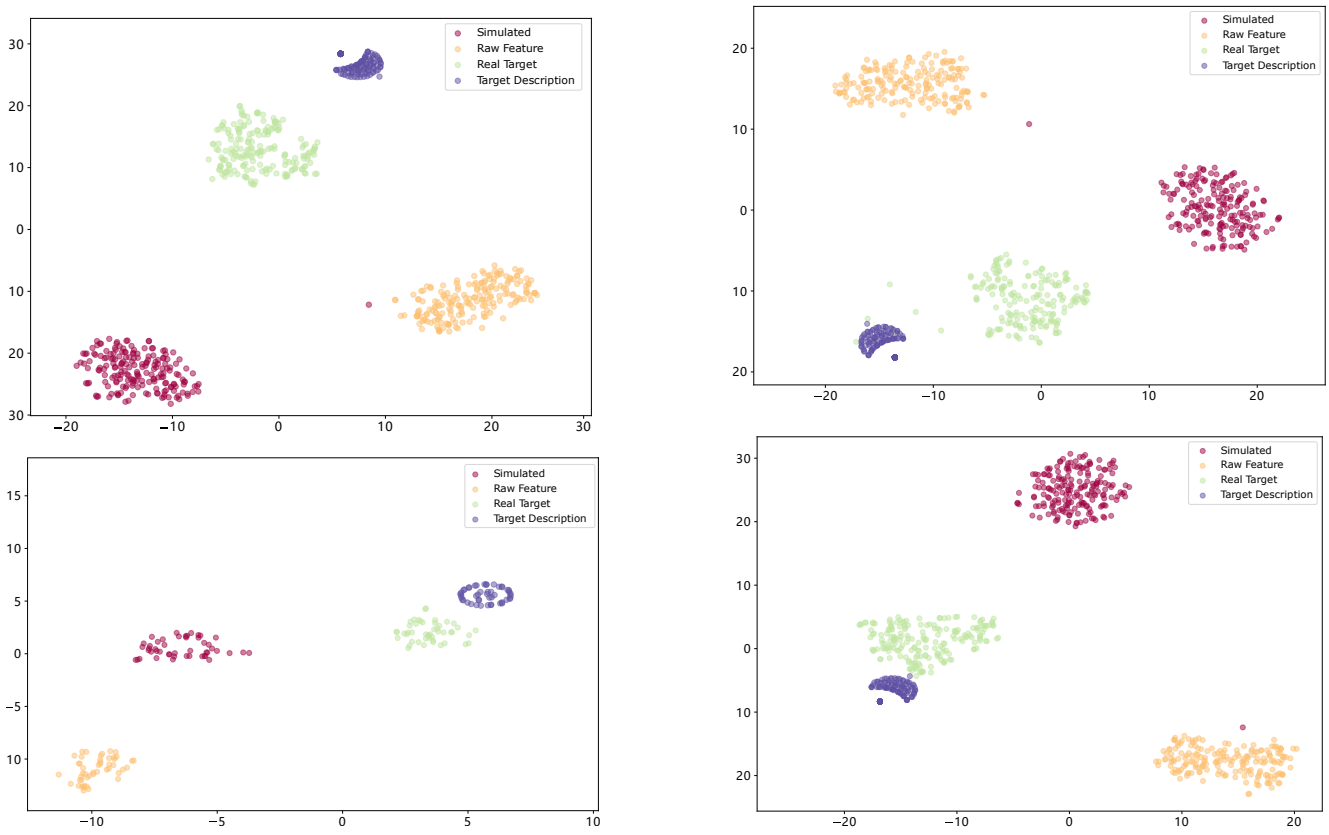

Figure 3. Real data of the Text-Driven Rectifier

mim et al. [1] advance the field by introducing Stable Diffusion and DreamBooth techniques. They extract domain knowledge from the pre-trained model, enabling the transfer of source images to the target image.

For the more challenging setting Zero-Shot Unsupervised Domain Adaptation, where no target image is available, Lengyel et al. [20] explore day-to-night domain adaptation. They introduce the Color Invariant Convolution Layer (CIConv) to achieve network invariance under varying lighting conditions. However, this approach heavily relies on the physics prior knowledge and is specifically tailored to a certain type of domain gap.

More recently, Fahes et al. [6] introduced a new setting named Prompt-driven Zero-shot Domain Adaptation. This setting, which does not allow access to the target domain data, leverages natural language descriptions of the target domain to adapt the model to new environments. PØDA utilizes CLIP to extract the target domain knowledge embedded within these natural descriptions. It employs a two-stage process to simulate target domain features, effectively addressing the domain gap.

Text-driven Vision Models. Recent advancements in contrastive image-language pretraining have led to significant achievements in multimodal learning across various tasks such as zero-shot classification [41], multi-modal retrieval [12], visual question answering [22], and have facilitated extensive work on Multimodal Large Language Models [19, 23, 24, 26, 29, 61, 68]. These developments have paved the way for modifying images using textual descriptions, bridging the previously challenging gap between visual and linguistic representations. For text-guided style transfer, CLIPstyler [16] diverges from relying on a generative process. This methodology offers a more realistic approach as it is not restricted to a specific training distribution, yet it simultaneously poses challenges due to the necessity of utilizing the encapsulated information within the CLIP latent space. The absence of a direct mapping between image and text representations necessitates regularization to effectively extract useful information from text embeddings. In this context, CLIPstyler optimizes a U-net autoencoder to preserve content, while varying the output image embedding in the CLIP latent space during the optimization process.

Semantic Segmentation which involves the classification of each pixel in an image, is a crucial task in computer vision. Several notable contributions in this domain have been

introduced [3, 18, 27, 28, 44, 47, 48, 54, 56, 59, 67]. Although these methods achieve impressive results, they often require considerable amounts of pixel-level annotated data, which can be a laborious and time-consuming task to collect and annotate. Additionally, they may face difficulties in effectively generalizing when deployed in new domains.

To address these challenges, numerous few-shot [39, 40, 45, 46] and semi-supervised [13, 17, 35, 65, 66] methods have been proposed. Recent research has primarily focused on addressing these challenges by employing domain adaptation strategies. For instance, in [63], a method is proposed that swaps the low-frequency spectrum to align the source and target domains. Another approach [49] involves mixing images from both domains along with their corresponding labels and pseudo-labels. Besides, [57] utilizes adversarial learning to train a domain adaptation network specifically for nighttime semantic segmentation. Furthermore, [10] introduces a novel model and training strategies to enhance training stability and mitigate overfitting to the source domain. Lastly, [11] employs masking of the target images to enable the model to learn spatial context relations of the target domain, providing additional clues for robust visual recognition. However, these methods both require access to the image of the target domain, which may not be feasible in some reality scenarios. Thus, we propose the Unified Language-driven Zero-shot Domain Adaptation to address this problem.

F. Qualitative Analysis

Qualitative Analysis on Autonomous Driving scenarios As shown in Fig. 2, we compare our proposed ULDA with the previous method on Autonomous Driving scenarios. Specifically, our visualization results are generated using a comprehensive all-in-one head, while the visualization results of the previous state-of-the-art (SOTA) method are generated using specific heads in each domain. The figures demonstrate that our method performs well on various objects, such as sidewalks and sky cars. This further reinforces the effectiveness of our approach.

Real data of the Text-Driven Rectifier In Section 4.3 of the main paper, we assert that using simulated target domain features to fine-tune the segmentation head may result in persistent discrepancies between the simulated features and the actual target domain features. To support this claim, we present real data in Figure 3. The relationships among the simulated feature, raw feature, real target feature, and the target description cluster demonstrate that due to the limitations of the simulating process, access to target domain data is restricted, resulting in the failure to capture the simulated features accurately. These findings highlight the significance of incorporating the Text-Driven Rectifier into the fine-tuning process. By doing so, it encourages closer alignment between the simulated features and the target features.

References

[1] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. One-shot unsupervised domain adaptation with personalized diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 698–708, 2023. 5, 6

[2] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, 2010. 3

[3] 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*, 40(4):834–848, 2017. 7

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 3

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1

[6] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. Pøda: Prompt-driven zero-shot domain adaptation. In *ICCV*, 2023. 1, 3, 4, 5, 6

[7] Qi Fan, Mattia Segu, Yu-Wing Tai, Fisher Yu, Chi-Keung Tang, Bernt Schiele, and Dengxin Dai. Towards robust object detection invariant to real-world domain shifts. In *ICLR*, 2023. 2, 3

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. 4

[9] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 4

[10] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022. 7

[11] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023. 7

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 6

[13] Li Jiang, Shaoshuai Shi, Zhuotao Tian, Xin Lai, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In *ICCV*, 2021. 7

[14] Kichun Jo, Junsoo Kim, Dongchul Kim, Chulhoon Jang, and Myoungho Sunwoo. Development of autonomous car—part

i: Distributed system architecture and development process. *IEEE Transactions on Industrial Electronics*, 61(12):7131–7140, 2014. 1

[15] Tarun Kalluri and Manmohan Chandraker. Cluster-to-adapt: Few shot domain adaptation for semantic segmentation across disjoint labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4121–4131, 2022. 5

[16] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *CVPR*, 2022. 6

[17] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *CVPR*, 2021. 7

[18] Xin Lai, Zhuotao Tian, Xiaogang Xu, Ying-Cong Chen, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Decouplenet: Decoupled network for domain adaptive semantic segmentation. In *ECCV*, 2022. 7

[19] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 6

[20] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *ICCV*, 2021. 6

[21] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *ECCV*, 2020. 5

[22] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021. 6

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 6

[24] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. *arXiv preprint arXiv:2312.16217*, 2023. 6

[25] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 2019. 4

[26] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023. 6

[27] Fangjian Lin, Sitong Wu, Yizhe Ma, and Shengwei Tian. Full-scale selective transformer for semantic segmentation. In *Proceedings of the Asian Conference on Computer Vision*, pages 2663–2679, 2022. 7

[28] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 7

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 6

[30] Jiaming Liu, Qizhe Zhang, Jianing Li, Ming Lu, Tiejun Huang, and Shanghang Zhang. Unsupervised spike depth estimation via cross-modality cross-domain knowledge transfer. *arXiv preprint arXiv:2208.12527*, 2022. 4

[31] Jiaming Liu, Ran Xu, Senqiao Yang, Renrui Zhang, Qizhe Zhang, Zehui Chen, Yandong Guo, and Shanghang Zhang. Adaptive distribution masked autoencoders for continual test-time adaptation. *arXiv preprint arXiv:2312.12480*, 2023. 4

[32] Jiaming Liu, Senqiao Yang, Peidong Jia, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. Vida: Homeostatic visual domain adapter for continual test time adaptation. *arXiv preprint arXiv:2306.04344*, 2023. 4

[33] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 4

[34] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 4

[35] Xiaoliu Luo, Zhuotao Tian, Taiping Zhang, Bei Yu, Yuan Yan Tang, and Jiaya Jia. Pfenet++: Boosting few-shot semantic segmentation with the noise-filtered context-aware prior mask. *TPAMI*, 2024. 7

[36] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. In *NeurIPS*, 2020. 5

[37] Jiayi Ni, Senqiao Yang, Jiaming Liu, Xiaoqi Li, Wenyu Jiao, Ran Xu, Zehui Chen, Yi Liu, and Shanghang Zhang. Distribution-aware continual test time adaptation for semantic segmentation. *arXiv preprint arXiv:2309.13604*, 2023. 4

[38] Theodoros Panagiotakopoulos, Pier Luigi Dovesi, Linus Härenstam-Nielsen, and Matteo Poggi. Online domain adaptation for semantic segmentation in ever-changing conditions. In *European Conference on Computer Vision*, pages 128–146. Springer, 2022. 4

[39] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23641–23651, 2023. 7

[40] Bohao Peng, Xiaoyang Wu, Li Jiang, Yukang Chen, Hengshuang Zhao, Zhuotao Tian, and Jiaya Jia. Oa-cnns: Omni-adaptive sparse cnns for 3d semantic segmentation. *arXiv preprint arXiv:2403.14418*, 2024. 7

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 6

[42] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021. 1

[43] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016. 4

[44] Haoru Tan, Sitong Wu, and Jimin Pi. Semantic diffusion network for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:8702–8716, 2022. 7

[45] Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *CVPR*, 2022. 7

[46] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. *TPAMI*, 2022. 7

[47] Zhuotao Tian, Pengguang Chen, Xin Lai, Li Jiang, Shu Liu, Hengshuang Zhao, Bei Yu, Ming-Chang Yang, and Jiaya Jia. Adaptive perspective distillation for semantic segmentation. *TPAMI*, 2023. 7

[48] Zhuotao Tian, Jiequan Cui, Li Jiang, Xiaojuan Qi, Xin Lai, Yixin Chen, Shu Liu, and Jiaya Jia. Learning context-aware classifier for semantic segmentation. In *AAAI*, 2023. 7

[49] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1379–1389, 2021. 7

[50] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 4

[51] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 4

[52] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 4

[53] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 4

[54] Yifeng Wang and Yi Zhao. Handwriting recognition under natural writing habits based on a low-cost inertial sensor. *IEEE Sensors Journal*, 2023. 7

[55] Yifei Wang, Wen Li, Dengxin Dai, and Luc Van Gool. Deep domain adaptation by geodesic distance minimization. In *ICCV Workshops*, 2017. 4

[56] Sitong Wu, Tianyi Wu, Fangjian Lin, Shengwei Tian, and Guodong Guo. Fully transformer networks for semantic image segmentation. *arXiv preprint arXiv:2106.04108*, 2021. 7

[57] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778, 2021. 7

[58] Xinyi Wu, Zhenyao Wu, Yuhang Lu, Lili Ju, and Song Wang. Style mixing and patchwise prototypical matching for one-shot unsupervised domain adaptive semantic segmentation. In *AAAI*, 2022. 2, 5

[59] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 7

[60] Runsheng Xu, Yi Guo, Xu Han, Xin Xia, Hao Xiang, and Jiaqi Ma. Opencda: an open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 1155–1162. IEEE, 2021. 1

[61] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *arXiv preprint arXiv:2312.14074*, 2023. 6

[62] Senqiao Yang, Jiarui Wu, Jiaming Liu, Xiaoqi Li, Qizhe Zhang, Mingjie Pan, and Shanghang Zhang. Exploring sparse visual prompt for cross-domain semantic segmentation. *arXiv preprint arXiv:2303.09792*, 2023. 4

[63] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020. 4, 7

[64] Zelin Zang, Lei Shang, Senqiao Yang, Fei Wang, Baigui Sun, Xuansong Xie, and Stan Z Li. Boosting novel category discovery over domains with soft contrastive learning and all in one classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11858–11867, 2023. 4

[65] Ye Zhang, Linghan Cai, Ziyue Wang, and Yongbing Zhang. Seine: Structure encoding and interaction network for nuclei instance segmentation. *arXiv preprint arXiv:2401.09773*, 2024. 7

[66] Ye Zhang, Ziyue Wang, Yifeng Wang, Hao Bian, Linghan Cai, Hengrui Li, Lingbo Zhang, and Yongbing Zhang. Boundary-aware contrastive learning for semi-supervised nuclei instance segmentation. *arXiv preprint arXiv:2402.04756*, 2024. 7

[67] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 7

[68] Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. Regionblip: A unified multimodal pre-training framework for holistic and regional comprehension. *arXiv preprint arXiv:2308.02299*, 2023. 6

[69] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, 2019. 4