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

Abstract

This paper introduces Unified Language-driven Zero-shot Domain Adaptation (ULDA), a novel task setting that enables a single model to adapt to diverse target domains without explicit domain-ID knowledge. We identify the constraints in the existing language-driven zero-shot domain adaptation task, particularly the requirement for domain IDs and domain-specific models, which may restrict flexibility and scalability. To overcome these issues, we propose a new framework for ULDA, consisting of Hierarchical Context Alignment (HCA), Domain Consistent Representation Learning (DCRL), and Text-Driven Rectifier (TDR). These components work synergistically to align simulated features with target text across multiple visual levels, retain semantic correlations between different regional representations, and rectify biases between simulated and real target visual features, respectively. Our extensive empirical evaluations demonstrate that this framework achieves competitive performance in both settings, surpassing even the model that requires domain-ID, showcasing its superiority and generalization ability. The proposed method is not only effective but also maintains practicality and efficiency, as it does not introduce additional computational costs during inference. The code is available on the project website[1].

1. Introduction

Being robust to the domain shift is a critical concept in machine learning, as it enables models trained on a source domain to be effectively applied to a new target domain [8, 9, 22, 28]. The domain adaptation (DA) task [5, 10, 11, 16, 23] may assume the availability of target domain data for fine-tuning the model. However, this assumption may potentially hinder real-world applications [2, 6, 17, 27]. For example, privacy concerns or data scarcity may prevent direct access to target data. Therefore, the underlying challenges caused by the absence of direct access to target domain data in developing the domain-versatile models should be considered and decently addressed, ensuring the applicability of DA techniques in practical situations.

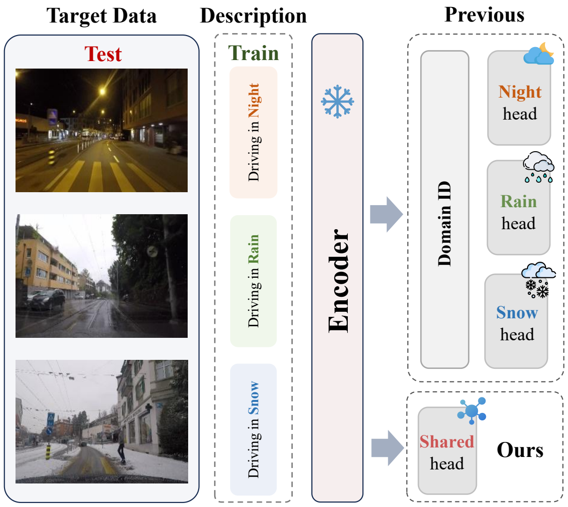

Figure 1. Our proposed Unified Language-driven Domain Adaptation (ULDA) task focuses on real-world practical scenarios. In the training phase, ULDA does not allow access to images of the target domain and only provides source domain images along with the textual descriptions. During testing, ULDA requires a single model to adapt to diverse target domains without domain-IDs, instead of using domain-specific heads as in previous methods.

Recently, the development of vision-language foundational models [1, 15, 18, 19, 26] has greatly advanced the alignment of image-text pairs, enabling effective generalization to novel concepts. This has paved the way for numerous studies that leverage the zero-shot capabilities of these models to tackle domain adaptation challenges [7, 8, 12, 14, 18]. Notably, PØDA [6] stands out for leveraging language embeddings obtained from CLIP [18] to simulate target domain visual representations, and PØDA tunes the model to fit the simulated features, such that the target images are not needed.

However, while PØDA demonstrates an impressive ability to achieve zero-shot domain adaptation without relying on target domain images, we observe that it introduces constraints that should be taken into account in practical contexts: *prior knowledge, i.e., domain-ID, is required to se-*

[*] Corresponding Author (tianzhuotao@hit.edu.cn).
[1] Senqiaoyang.com/project/ULDA

lect the domain-specific model. For example, for "driving in rain" and "driving in snow", two individual models are trained to fit these two task domains with the help of CLIP text embeddings. During inference, when an image comes, the system processes in two steps: 1) know what the current domain is via Domain ID, and then 2) select the corresponding model. These domain-specific customization requirements may hinder the model's flexibility and scalability, thereby limiting its broader applications.

To address the aforementioned issue, we propose a novel and practical task called Unified Language-driven Zero-shot Domain Adaptation (ULDA), as shown in Fig. 1. Following previous literature [6], ULDA also does not allow access to images of the target domain, only providing source domain images along with the textual descriptions regarding target domains to accomplish the adaptation training process. However, ULDA takes a step further by requiring a single model to adapt to diverse target domains without explicit hints, *i.e.*, the domain-ID, during testing. Nevertheless, this new requirement also presents a significant challenge: *how to adapt a single model's embedding space to accommodate multiple domains while still maintaining semantic discriminative capabilities for different categories?*

To address this challenge, we propose a new framework for ULDA. It has three essential components: Hierarchical Context Alignment (HCA), Domain Consistent Representation Learning (DCRL), and Text-Driven Rectifier (TDR). Specifically, HCA aligns simulated features with target text at multiple visual levels to mitigate the semantic loss caused by the vanilla scene-text alignment. Then, DCRL retains the semantic correlations between different regional representations to that of the text embeddings across diverse domains, ensuring structural consistency. Additionally, we incorporate TDR to rectify simulated features, mitigating the bias between the simulated and real target visual features.

We validate the effectiveness of our proposed method through extensive empirical evaluations conducted in both the previous classic setting [6] and the proposed ULDA. The results consistently demonstrate that our approach achieves competitive performance in both settings, highlighting its superiority and efficacy. In summary, our contribution can be summarized in three key aspects:

- Unlike existing literature, we go beyond existing approaches by examining the limitations that hinder further applications. To this end, we propose a more practical setting called Unified Language-driven Zero-shot Domain Adaptation (ULDA).
- To address the new challenge posed by ULDA, we propose a new framework, and it comprises three key components, namely Hierarchical Context Alignment (HCA), Domain Consistent Representation Learning (DCRL), and Text-Driven Rectifier (TDR), for achieving better alignment to the text embedding space, ensuring a better adaptation performance.
- Despite its simplicity, our proposed method's effectiveness has been verified in both settings. Furthermore, it does not introduce any additional computational costs during model inference, ensuring its practicality.

2. Preliminary

In this section, we introduce a closely related work PØDA [6], which proposes a paradigm for prompt-driven zero-shot domain adaptation in computer vision, by only leveraging a natural language description of the target domain, thus eliminating the need for target domain images during training. A more detailed introduction regarding related works is shown in the supplementary file.

Specifically, PØDA undergoes two stages of training to leverage the pretrained CLIP encoder for optimizing source feature transformations and aligning them with the text embedding of the target domain. In Stage-1, it learns to simulate target features. In Stage-2, it fine-tunes the segmentation head with the simulated ones. Details are as follows.

Stage-1: PØDA introduces Prompt-driven Instance Normalization (PIN), as in Eq. (1), in which $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are learnable variables guided by a text domain prompt to simulate the knowledge of the target domain, while $\mu(\mathbf{f}_{\mathrm{s}})$ and $\sigma(\mathbf{f}_{\mathrm{s}})$ represent the mean and standard deviation of the source input features $\mathbf{f}_{\mathrm{s}}$.

$$\mathbf{f}_{\mathrm{s}\to\mathrm{t}} = \mathrm{PIN}\left(\mathbf{f}_{\mathrm{s}}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right) = \boldsymbol{\sigma}\left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right) + \boldsymbol{\mu}. \quad (1)$$

PIN is adopted to transform source domain features into the target domain, *i.e.*, $\mathbf{f}_{\mathrm{s}\to\mathrm{t}}$. This operation is followed by an attention-based pooling operation, resulting in the output denoted as $\bar{\mathbf{f}}_{\mathrm{s}\to\mathrm{t}}$. To ensure a proper shift from the source to the target domain, it is necessary to promote similarity between $\bar{\mathbf{f}}_{\mathrm{s}\to\mathrm{t}}$ and CLIP text embeddings TrgEmb, by applying the loss function presented in Equation (2), which encourages alignment between the transformed features and the textual representations, facilitating decent adaptation from the source to target domain.

$$\mathcal{L}_{\mu,\sigma}\left(\bar{\mathbf{f}}_{\mathrm{s}\to\mathrm{t}}, \mathrm{TrgEmb}\right) = 1 - \frac{\bar{\mathbf{f}}_{\mathrm{s}\to\mathrm{t}} \cdot \mathrm{TrgEmb}}{\left\|\bar{\mathbf{f}}_{\mathrm{s}\to\mathrm{t}}\right\| \left\| \mathrm{TrgEmb} \right\|}. \quad (2)$$

Stage-2: With the simulated features obtained via Eq. (1), PØDA fine-tunes the pre-trained segmentation head to enable the model to better adapt to the target domain for accomplishing the downstream task. This stage's training is supervised by the cross-entropy loss between the segmentation predictions and the ground-truth masks.

It is worth noting that, for both two training stages, only the images from the source domain and text descriptions are available. After two phases of training, the model is evaluated on the images of target domains. More details can be found in [6].

3. ULDA: Unified Language-driven Zero-shot Domain Adaptation

Motivation. Traditional domain adaptation methods often depend on having access to data from the target domain in order to align the models. However, this dependence on target domain data can lead to overfitting to specific domains and subsequently undermine the generalization performance of the models. In real-world applications, especially in dynamic settings like autonomous driving, acquiring comprehensive data for every possible adverse condition (e.g., rain, snow, fog, night) is not always feasible. Instead, practitioners may only have a conceptual understanding or hypothetical descriptions of potential downstream tasks. In this case, the ability to augment a model's performance in such predicted scenarios without actual data collection is preferred.

To achieve this, PØDA [6] tunes different models to tackle individual scenarios separately. However, the prior knowledge of each upcoming domain for selecting the corresponding model may not always be accessible in practice. Therefore, we believe it is necessary to adopt an adaptation approach such that a single model can be scaled to simultaneously fit a broad spectrum of domain conditions.

Considering these practical constraints, we propose a new task setting, namely Unified Language-driven Zero-shot Domain Adaptation (ULDA), which encourages the model to be capable of adapting to a variety of conditions without access to real data and domain prior knowledge.

Task setting. The model $\mathcal{M}$ is limited to training with data $\mathcal{I}_s$ from the source domain $\mathcal{D}_s$ and has no access to target domain data $\mathcal{I}_t$, where $t = 1, 2...n$ represent the n target domains $\mathcal{D}_t$. $\mathcal{M}$ can only utilize natural descriptions $\mathcal{T}_t$ to understand the characteristics of target domain scenarios.

One of the challenges in this context is to effectively extract sufficient information from textual descriptions alone for adapting source visual features to different domains. Another crucial challenge is to enable a single model $\mathcal{M}$ to adapt to multiple target domains $\mathcal{D}_t$ without relying on specific domain IDs. This would allow the model to achieve robustness across diverse scenarios while still maintaining a strong ability to discriminate between different classes.

Comparison with other settings. Different from Unsupervised Domain Adaptation (UDA), ULDA offers the advantage of generalizing to target domains without the need for target domain images. Instead, it only relies on a concise one-sentence description for each domain. This leads to a significant reduction in resource overhead. Additionally, unlike prompt-driven zero-shot domain adaptation proposed in PØDA that requires domain IDs to invoke domain-specific models, the proposed ULDA enables a single model to adapt to multiple downstream scenarios without the requirement for separate tuning for each scenario.

4. Method

The proposed ULDA brings a challenge in representation learning as a single model needs to adapt to multiple domains. This challenge arises from the fact that aligning the model towards target domains, such as "driving in rain" and "driving in snow," may potentially compromise semantic discrimination for precise segmentation.

For better accomplishing ULDA, we propose a framework that is composed of three components. The overview is shown in Fig. 2, and the respective details are as follows.

4.1. Hierarchical Context Alignment

Vanilla scene-text alignment causes semantic loss. PØDA achieved vision-language alignment at a scene level by directly aligning the pooled feature $\bar{\mathbf{f}}_{\mathrm{s}\to\mathrm{t}}$ with the text embedding TrgEmb via Eq. (2). However, it is challenging for the model to achieve a decent alignment with the target domain by only adapting the global context to fit the target domain, because this may cause potential semantic loss when aligning features of different objects in a scene to a single shared target text domain embedding, causing a deviation from their respective real semantic distributions.

To alleviate this issue, we propose a Hierarchical Context Alignment (HCA) strategy, which enables intricate alignments on the feature $\mathbf{f}_{\mathrm{s}\to\mathrm{t}}$ at multiple levels, including 1) the entire scene, 2) regions in the scene, and 3) pixels in the scene. The scene-text alignment follows that of Eq. (2), while the proposed region- and pixel-text alignments are elaborated as follows.

Regional alignment. During the adaptation process, it is essential for regions belonging to different categories to retain their unique semantic characteristics. To achieve this, by leveraging the class names existing in the ground truth, and the target domain description, we can get the more fine-grained d-dimensional text embedding $\mathcal{T} \in \mathcal{R}^{[n\times d]}$ of n classes contained in the image. By doing so, we can align different regions with more suitable counterparts, ensuring that their individual semantic characteristics are preserved. For example, in a rainy scenario with n classes, we can get descriptions such as "the bus in rain," "the road in rain," "the rider in rain," and so on. Then, the corresponding text embeddings $\mathcal{T} \in \mathcal{R}^{[n\times d]}$ for these descriptions can be obtained from the pre-trained CLIP text encoder.

After that, given the image feature map $\mathbf{f}_{\mathrm{s}\to\mathrm{t}} \in \mathcal{R}^{[HW\times d]}$ and text embedding $\mathcal{T} \in \mathcal{R}^{[n\times d]}$, the pixel-wise ground-truth annotation $\boldsymbol{y} \in \mathcal{R}^{[HW]}$ can be accordingly transformed into n binary masks $\boldsymbol{y}_* \in \mathcal{R}^{[n\times HW]}$ indicating the existence of n classes in $\boldsymbol{y}$. Then, we can obtain the regional prototypes $\mathcal{C} \in \mathcal{R}^{[n\times d]}$ by applying masked average pooling (MAP) with $\boldsymbol{y}_*$ and $\mathbf{f}_{\mathrm{s}\to\mathrm{t}}$ as Eq. (3):

$$\mathcal{C} = \frac{\boldsymbol{y}_* \times \mathbf{f}_{\mathrm{s}\to\mathrm{t}}}{\sum_{j=1}^{HW} \boldsymbol{y}_*(\cdot, j)}. \tag{3}$$

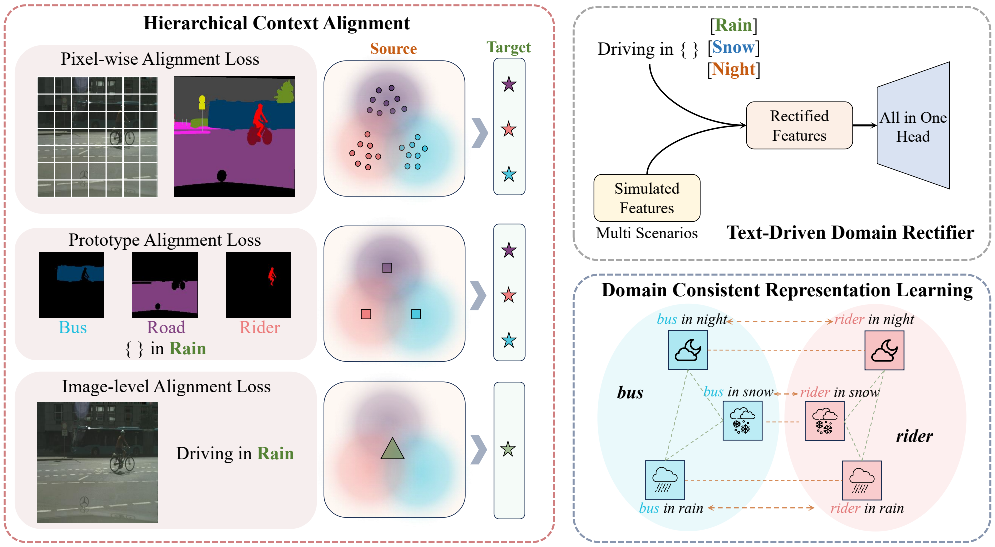

Figure 2. **Illustration of the three key components for our ULDA framework.** The ULDA's pipeline follows [6]. Our proposed Hierarchical Context Alignment operates across Pixel-level, Regional-level, and Scene-level to align features with text embeddings. The circles, squares, and triangles represent the hierarchical features, respectively. The Domain Consistent Representation Learning ensures a consistent correlation between prototypes and text embeddings across multiple target domains. Text-Driven Rectifier incorporates text embeddings to rectify the simulated PIN features during the fine-tuning phase.

Then, we can calculate the cosine similarity matrix $\mathcal{S} \in \mathcal{R}^{[n \times n]}$ between the $\mathcal{T}$ and $\mathcal{C}$ in Eq. (4).

$$\mathcal{S} = \frac{\mathcal{C} \times \mathcal{T}^T}{\|\mathcal{C}\| \, \|\mathcal{T}\|^T} \tag{4}$$

Therefore, we could use the categorical prototypes $\mathcal{C}_y$ and text embeddings $\mathcal{T}$ to accomplish the regional alignment as:

$$\mathcal{L}_r = -\sum_{i=1}^{n} \log \left(\frac{\exp\left(\mathcal{S}_{ii}/\tau\right)}{\sum_{k=1}^{n} \exp\left(\mathcal{S}_{ik}/\tau\right)} \right) \tag{5}$$

where τ is the temperature parameter, and we empirically set it to 0.1. Eq. (5) encourages the regional prototypes to be similar to the corresponding text embeddings in the target domain while pushing away negative pairs.

Pixel-wise alignment. Building upon the regional alignment, we further enhance the alignment between the source and the unseen target domains by incorporating a pixel alignment loss $\mathcal{L}_p$. Compared to the alignments at the scene and regional levels, $\mathcal{L}_p$ serves to narrow the distance at a more intricate level, enabling more precise alignment between the two domains. Similar to the regional alignment, we begin the pixel-level alignment by computing the class probability $\mathcal{P} \in \mathcal{R}^{[HW \times n]}$ for each pixel, as in Eq. (6):

$$\mathcal{P} = \frac{\mathbf{f}_{\mathrm{s}\rightarrow\mathrm{t}} \times \mathcal{T}^T}{\|\mathbf{f}_{\mathrm{s}\rightarrow\mathrm{t}}\| \, \|\mathcal{T}\|^T} \tag{6}$$

Then, we use $\mathcal{P}$ to calculate the cross-entropy loss with the ground truth $\boldsymbol{y} \in \mathcal{R}^{[HW]}$ using Eq. (7):

$$\mathcal{L}_p = -\frac{1}{HW} \sum_{i=1}^{HW} \boldsymbol{y}_i \log\left(\mathcal{P}_i\right) \tag{7}$$

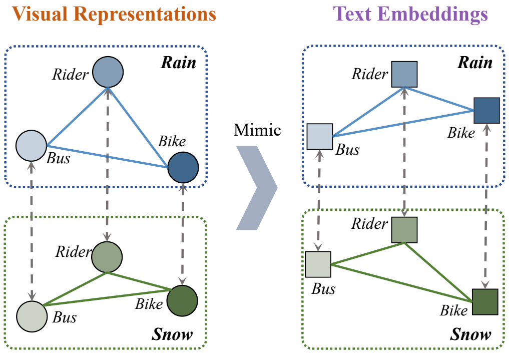

Figure 3. **Domain Consistent Representation Learning.** We ensure the visual regional representations have similar correlations with that of text embeddings, both within the same domain and across different domains.

The overall objective. To this end, the training objective $\mathcal{L}_{HC}$ for hierarchical context alignment is formulated as:

$$\mathcal{L}_{HC} = \lambda_r \mathcal{L}_r + \lambda_p \mathcal{L}_p + \mathcal{L}_{\mu,\sigma}\left(\mathbf{f}_{\mathrm{s}\rightarrow\mathrm{t}}, \mathrm{TrgEmb}\right) \tag{8}$$

4.2. Domain Consistent Representation Learning

Despite successfully bringing the feature $\mathbf{f}_{\mathrm{s}\rightarrow\mathrm{t}}$ closer to the target domain, this process may unintentionally interfere with the relational information among different prototypes. As depicted in Fig. 3, although the entities 'rider,' 'bus,' and 'bike' from the source domain have been individually aligned with their target domain text embeddings, the inherent correlations between these classes might be disrupted due to the domain shift. Furthermore, the same category in different contexts may exhibit distinct correlations between the visual and text representations.

For instance, the text embeddings of “a bus in the snow,” “a bus in the rain,” and “a bus at night” may have different correlations compared to their visual counterparts in the contexts of ‘snow’, ‘rain’ and ‘night’ respectively. This discrepancy in relational consistency between the simulated domain features and text embeddings can lead the model to erroneously diverge from the true distributions represented by the text embeddings of the target domain. To tackle this problem, we propose the domain consistency loss $\mathcal{L}_{DC}$.

Specifically, for n categories in m target domains, with Eq. (3), we can obtain m prototypes $\mathcal{C} \in \mathcal{R}^{[n\times d]}$ we group it into the $\widetilde{\mathcal{C}} \in \mathcal{R}^{[mn\times d]}$. Similarly, we obtain the m text embedding $\mathcal{T} \in \mathcal{R}^{[n\times d]}$, grouped into the $\widetilde{\mathcal{T}} \in \mathcal{R}^{[mn\times d]}$. Lastly, we adopt Eq. (9) as $\mathcal{L}_{DC}$ to enforce representation consistency across multiple domains by preserving the correlation between the prototypes and the corresponding text embeddings in different scenes.

$$\mathcal{L}_{DC} = MSE(\frac{\widetilde{\mathcal{C}}\widetilde{\mathcal{C}}^T}{\|\widetilde{\mathcal{C}}\|^2}, \frac{\widetilde{\mathcal{T}}\widetilde{\mathcal{T}}^T}{\|\widetilde{\mathcal{T}}\|^2}) \tag{9}$$

4.3. Text-Driven Rectifier

Evils in the simulated features. During the second stage introduced in Sec. 2, the model utilizes the simulated target domain features to fine-tune the segmentation head, enabling the model to be effectively adapted to the target domain. However, as shown in Fig. 4, discrepancies may persist between the simulated features and the actual target domain features. It is crucial to consider these discrepancies as using simulated features directly may lead to a deviation of the segmentation head from the true target distributions, yielding worse segmentation performance after tuning.

Therefore, we propose to address this issue by leveraging the text embeddings obtained from CLIP, which effectively resemble the distributions in the real target domain. By adopting these text embeddings as a prior, we may rectify the simulation process, thereby encouraging the simulated features to align more closely with the target features.

Rectification benefits adaptation. Specifically, we denote the features simulated by PIN as $\mathbf{f}_{\mathrm{s}\rightarrow\mathrm{t}}$, *i.e.* $\mathbf{f}_{\mathrm{s}\rightarrow\mathrm{t}} = \mathrm{PIN}\left(\mathbf{f}_{\mathrm{s}}, \boldsymbol{\mu}, \boldsymbol{\sigma}\right)$. Then, we get the rectified feature $\widetilde{\mathbf{f}}_{\mathrm{s}\rightarrow\mathrm{t}}$ by following Eq. (10):

$$\widetilde{\mathbf{f}}_{\mathrm{s}\rightarrow\mathrm{t}} = \beta\left(\widetilde{\boldsymbol{\sigma}}\left(\frac{\mathbf{f}_{\mathrm{s}\rightarrow\mathrm{t}} - \mu\left(\mathbf{f}_{\mathrm{s}\rightarrow\mathrm{t}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}\rightarrow\mathrm{t}}\right)}\right) + \widetilde{\boldsymbol{\mu}}\right) + \mathbf{f}_{\mathrm{s}\rightarrow\mathrm{t}}, \tag{10}$$

where β is a learnable factor, initialized as 0.1, controlling the extent of the rectification applied to $\mathbf{f}_{\mathrm{s}\rightarrow\mathrm{t}}$. $\widetilde{\sigma}$ and $\widetilde{\mu}$ are obtained by passing text embedding through a linear layer to represent the mean and standard deviation of the target domain features, respectively. Then we utilize the $\widetilde{\mathbf{f}}_{\mathrm{s}\rightarrow\mathrm{t}}$ obtained from multiple domains to fine-tune the head.

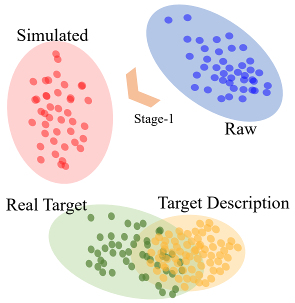

Figure 4. **Text-Driven Rectifier.** Despite the simulated features exhibiting a closer alignment with the real targets in comparison to raw features, a gap still exists. To address this disparity, we employ the text descriptions from the target domain to rectify the simulated ones when tuning the segmentation head in Stage-2.

Consequently, through text-driven rectification (TDR), we are able to correct the features initially simulated by PIN to preserve the distinctive attributes of each domain. This enhancement improves the overall generalization capability of the shared final head, enabling it to effectively adapt to multiple domains simultaneously.

It is worth noting that TDR is applied to bridge the gap between simulated features and actual target domain features, exclusively during Stage-2 mentioned in Sec. 2 for head tuning. We do not apply TDR to Stage-1 as it would lead to a trivial solution. The theoretical proof is in Sec. 6.

4.4. Overall Loss Function

With the above strategies, the overall loss function $\mathcal{L}$ for Stage-1’s training becomes:

$$\mathcal{L} = \lambda_{HC}\mathcal{L}_{HC} + \lambda_{DC}\mathcal{L}_{DC} + \lambda_{seg}\mathcal{L}_{seg} \tag{11}$$

where the λ_{HC}, λ_{DC}, λ_{seg}, are weighting coefficients balancing the respective loss components. We observe the segmentation loss $\mathcal{L}_{seg}$ benefits the Stage-1 training, as illustrated in the ablation study. For Stage-2, which involves fine-tuning the model for segmentation, only the vanilla segmentation loss $\mathcal{L}_{seg}$ is adopted.

5. Experiments

In Sec. 5.1, we present the details of our experiment setup, encompassing the datasets, task settings, and implementation specifics. We showcase the effectiveness of our method in both traditional settings, demonstrated in Sec. 5.2, and in new practical settings, elaborated upon in Sec. 5.3. Additionally, we utilize GPT-4 to generate multiple natural descriptions of autonomous driving scenarios, including some uncommon situations such as sandstorms and forest fires. These scenarios are often challenging for model training due to their limited data availability. Due to space limitations, we detail this interesting experiment in the supplementary.

5.1. Experiment Setup

Datasets. We primarily use the Cityscapes [4] as the source domain dataset. Following PØDA, we report the main results using ACDC [21]. To demonstrate the generalization of our method, we also investigate two extra adaptation scenarios: real to synthetic (source: Cityscapes; target: GTA5 [20]) and synthetic to real (source: GTA5; target: Cityscapes). The evaluation configuration follows [6].

Implentation Details . In this study, we conduct comparisons on both the setting of PØDA and our proposed setting. As for the base segmentation model, DeepLabv3+ [3] with a backbone model of pre-trained CLIP-ResNet-50[2] is adopted. The base model is sufficiently trained on the source domain following the configuration of [6]. In the fine-tuning stage (Stage 2), we begin with the source pre-trained model and only fine-tune the classifier head, also following the configurations of [6] for a fair comparison. All models are tested on the original images without resizing, and more details are in the supplementary file.

5.2. Effectiveness on Traditional Settings.

Effectiveness on prompt-driven zero shot adaptation. Following the previous benchmark, we explore various adaptation scenarios, including: day→night, clear→snow, clear→rain, real→synthetic, and synthetic→real. We compare our approach with two state-of-the-art baselines: CLIPstyler [13] for zero-shot style transfer and PØDA [6] for prompt-driven zero-shot adaptation. Notably, PØDA, CLIPstyler and our approach, do not utilize target images during training. Following the previous setting, we only select simple prompts for each domain to demonstrate the effectiveness of our method.

As shown in Table 1, our proposed method consistently outperforms all baseline models in zero-shot domain adaptation, using mean Intersection over Union (mIoU) as the comparative metric. Our method surpasses previous approaches, achieving improvements in all the scenarios. It is noteworthy that the previous SOTA method, PØDA, requires training a separate head for each scenario. We believe that while using distinct heads for individual scenarios simplifies the task, it also compromises the method's generalizability, limiting its practical application in real-world settings. In contrast, our method surpasses previous approaches by using only a single head, further demonstrating our method's effectiveness. Furthermore, our proposed method, without altering the original framework or requiring any additional information, achieves significant improvements by deeply exploring the relationships between multi-level images and texts. This further validates the effectiveness of our approach in the traditional zero-shot domain adaptation task.

Source	Target eval.	Method	mIoU[%]
CS		Prompt = “driving at night”	
	ACDC Night	source-only	18.31
		CLIPstyler	21.38
		PØDA	25.03
		ULDA	**25.40**
		Prompt = “driving in snow”	
	ACDC Snow	source-only	39.28
		CLIPstyler	41.09
		PØDA	43.90
		ULDA	**46.00**
		Prompt = “driving under rain”	
	ACDC Rain	source-only	38.20
		CLIPstyler	37.17
		PØDA	42.31
		ULDA	**44.94**
		Prompt = “driving in a game”	
	GTA5	source-only	39.59
		CLIPstyler	38.73
		PØDA	40.77
		ULDA	**42.91**
GTA5		Prompt = “driving in real”	
	Cityscapes	source-only	36.38
		CLIPstyler	32.40
		PØDA	40.02
		ULDA	**41.73**

Table 1. **Performance on classic prompt driven zero shot domain adaptation in semantic segmentation.** Performance (mIoU%) of ULDA framework compared against previous methods and source-only baseline. CS stands for Cityscapes [4].

Besides, in supplementary, we compare our proposed method with the one-shot SOTA method SM-PPM [25] to demonstrate the effectiveness of our method.

5.3. Effectiveness on Unified language-driven zero shot adaptation

Quantitative Results. For the newly introduced task of Unified Language-driven Zero-Shot Adaptation, our aim is to propose a benchmark that is more aligned with real-world scenarios and holds practical value. Accordingly, following the setting described in Sec. 3, we are limited to only having natural descriptions of potential target domains. And this setting also requires our model to be versatile enough to be tested across all downstream target domains using just a singular model architecture. We set two practical adaptation scenarios as our benchmark, including: clear-to-adverse-weather adaptation on Cityscapes→ACDC and synthetic-to-real adaptation on GTA5→Cityscape and ACDC.

We establish the mean mIoU as the comparative metric for our study. This mean mIoU is derived by calculating the average of mIoU values across various domains. Additionally, we also report the mIoU and mean Accuracy (mAcc) for each individual domain. To demonstrate the generalizability of our method, we just utilize simple prompt descriptions to reflect the target domain knowledge.

Clear-to-Adverse weather. In Table 2, we compare our

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

Scenarios		Source2Fog		Source2Night		Source2Rain		Source2Snow		Mean-mIoU
Domain Description		driving in fog		driving at night		driving under rain		driving in snow		
Method	REF	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	
Source	OpenAI	49.98	65.42	18.31	34.16	38.20	58.97	39.28	54.64	36.44
CLIPStyler	CVPR2022 [12]	48.87	64.31	20.83	35.32	36.97	57.46	40.31	54.42	36.75
PØDA*	ICCV2023 [6]	51.54	64.51	25.03	55.5	42.31	75.4	43.90	70.7	40.65
ULDA	ours	**53.55**	**80.2**	**25.40**	**55.8**	**44.94**	**74.4**	**46.00**	**70.0**	**42.47**

Table 2. **Performance comparison of clear-to-adverse weather in ULDA.** We use Cityscape as the source domain and ACDC as the four target domains in this setting. Mean-mIoU represents the average mIoU value in four scenarios. PØDA* represents the model that uses different segmentation heads in specific domains, while the others adopt the shared head.

Scenarios	Source2CS		Source2Fog		Source2Night		Source2Rain		Source2Snow		Mean-mIoU
Domain Description	driving in real		driving in fog		driving at night		driving under rain		driving in snow		
Method	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	
Source	36.38	46.19	33.20	42.51	12.22	22.56	33.32	43.15	32.33	40.60	29.49
CLIPStyler	32.20	41.64	30.79	40.37	11.12	20.18	31.17	40.06	30.65	38.97	27.19
PØDA*	40.05	48.95	35.76	44.98	13.35	25.24	34.19	45.93	33.81	42.10	31.43
ULDA	**41.73**	**51.98**	**36.98**	**46.56**	**15.72**	**28.99**	**35.84**	**47.39**	**35.77**	**43.74**	**33.21**

Table 3. **Performance comparison of Synthetic-to-Real in ULDA** We use GTA5 as the source domain, Cityscapes and ACDC as the five target domains in this setting. Mean-mIoU represents the average mIoU value in five scenarios. PØDA* represents the model used different segmentation heads in specific domains, while the others adopt the shared head.

proposed ULDA with the SOTA method in Zero-shot domain adaptation. PØDA* employs four distinct heads, selecting a specific head tailored to each scenario. Our proposed method consistently surpasses all previous models in Unified Language-driven Domain Adaptation. It achieves improvements of 6.03% over the baseline source model. Remarkably, our approach, which utilizes a single head, even exceeds the performance of PØDA*, and achieves improvements of 1.82% mIoU, which necessitates training separate heads for different scenarios. This highlights the strength of our method, particularly in its ability to employ Hierarchical Context modeling. Such modeling adeptly extracts and leverages the multi-level correlation between image and text, facilitating more effective domain transfer. Moreover, our approach of domain-consistent representation learning ensures consistency across various domains. This enables our model to generalize effectively to various domains within a single unified model architecture.

Synthetic-to-Real. In traditional settings, the capability of methods in synthetic-to-real scenarios is typically validated first on the GTA5→Cityscapes, and then further verified through additional experiments to demonstrate their applicability to adverse weather conditions, such as adaptation from Cityscapes to ACDC. However, the practical utility and generalizability of these experiments are somewhat limited. In autonomous driving scenarios, it is more desirable to learn from a diverse range of source domain virtual datasets like GTA5 and to achieve direct generalization to various complex real-world scenarios, like ACDC. Therefore, our experimental setting, termed "Synthetic-to-Real," specifically focuses on the GTA5→Cityscapes+ACDC dataset.

In Table 3, we present a comprehensive comparison of our proposed ULDA method against the current SOTA methods in Zero-shot domain adaptation. Our method consistently outperforms all previous models in Unified Language-driven Domain Adaptation, achieving significant improvements of 3.72% and 1.78% over the baseline source model and the former SOTA method, PØDA*, respectively. Notably, our method shows remarkable performance enhancement in complex scenarios. For instance, in GTA5-to-Cityscapes where the source model only achieves 36.38% mIoU and 46.19% mAcc, our proposed method achieves a 5.35% mIoU and 5.79% mAcc increase. This highlights that our method can effectively learn extensive knowledge directly from GTA5 through language alone and generalize to real and complex scenarios like Cityscapes ACDC. Such results underscore the effectiveness and practicality of our approach. It demonstrates our method's capability to extract and apply knowledge from the language description of the target domains, proving its utility and adaptability in real-world complex environments.

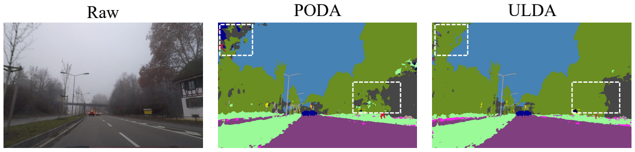

Figure 5. **Qualitative Results** of our models and previous SOTA method on ACDC-Foggy. More results are in the Appendix.

Qualitative Results To demonstrate the effectiveness of our proposed method, we show the qualitative analysis, shown in Fig. 5. Other visualization results are shown in the supplementary.

5.4. Effectiveness of each component

We conduct the ablation study on Cityscapes-to-ACDC in the Language-driven zero shot adaptation setting and eval-

	HCA	DCRL	TDR	$\mathcal{L}_{seg}$	Mean-mIoU
Ex_1					39.67
Ex_2	✓				40.52
Ex_3	✓	✓			41.49
Ex_4	✓		✓		41.35
Ex_5	✓	✓		✓	41.68
Ex_6	✓	✓	✓	✓	42.47

Table 4. **Ablation: Contribution of each component.**

uate the contribution of each component in our method, including Hierarchical Context Alignment (HCA), Domain Consistent Representation Learning (DCRL) and Text-Driven Domain Rectify (TDR). Mean-mIoU and Mean-mAcc are used as metrics, representing the average mIoU and mAcc values across four scenarios, respectively.

Effects of HCA and DCRL. As shown in Table 4, Ex_1 represents the baseline method, PØDA, which only leverages the scene-level alignment. In comparison, Ex_2 introduces HCA to align image embeddings and natural descriptions on a multi-level basis. This approach results in improvements of 0.85% in mean-mIoU. In Ex_3, the integration of DCRL leads to further enhancements, with an increase of 0.97% in mean-mIoU. This signifies that incorporating Domain Consistent Representation Learning effectively addresses domain discrepancies. For a detailed examination of the impact of DCRL, refer to the further ablation study provided in the Appendix.

Effects of TDR and segmentation loss. Building on Ex_2, Ex_4 introduces Text-Driven Domain Rectify (TDR). TDR, by making rectify during the fine-tuning phase, bridges the gap between the simulated features and the real features of the target domain. This results in a performance improvement of 0.83% in mean-mIoU. In Ex_5, based on the foundation laid by Ex_3, we introduce a downstream task loss: $\mathcal{L}_{seg}$, during the fine-tuning phase. This loss function helps prevent the overfitting of features to text embeddings and ensures the retention of downstream task capabilities. It effectively maintains a balance between domain adaptation and task-specific performance. By incorporating this element, Ex_5 achieves an additional performance improvement of 0.33% in mean-mIoU.

Overall, Ex_6 shows the complete combination of all components, achieving 42.47% mean-mIoU in total. This demonstrates that all components compensate each other and jointly address the challenge in Language-driven zero-shot domain adaptation. Due to the limitation of the space, the additional ablation studies are shown in the Appendix.

6. Further Discussions

Why not incorporate TDR to Stage-1? 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:

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

For the simulated $\mathbf{f}_{\mathrm{s}\to\mathrm{t}}$, we have $std(\mathbf{f}_{\mathrm{s}\to\mathrm{t}}) = \sigma$, $mean(\mathbf{f}_{\mathrm{s}\to\mathrm{t}}) = \mu$. Substituting them into Eq. (10) yields:

$$\begin{aligned}\widetilde{\mathbf{f}}_{\mathrm{s}\to\mathrm{t}} &= \beta\left(\widetilde{\boldsymbol{\sigma}}\left(\frac{\mathbf{f}_{\mathrm{s}\to\mathrm{t}} - \mu\left(\mathbf{f}_{\mathrm{s}\to\mathrm{t}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}\to\mathrm{t}}\right)}\right) + \widetilde{\boldsymbol{\mu}}\right) + \mathbf{f}_{\mathrm{s}\to\mathrm{t}} \\ &= \beta\left(\widetilde{\boldsymbol{\sigma}}\left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{s}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right) + \widetilde{\boldsymbol{\mu}}\right) + u + \sigma\left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right). \\ &= \left(\frac{\mathbf{f}_{\mathrm{s}} - \mu\left(\mathbf{f}_{\mathrm{s}}\right)}{\sigma\left(\mathbf{f}_{\mathrm{s}}\right)}\right)\left(\beta\widetilde{\boldsymbol{\sigma}} + \sigma\right) + \left(\beta\widetilde{\boldsymbol{\mu}} + u\right).\end{aligned} \tag{12}$$

$\widetilde{\boldsymbol{\sigma}}$ and $\widetilde{\boldsymbol{\mu}}$ are derived by passing text embeddings through a linear layer. The parameters σ and μ are learnable and are designed to simulate features of the target domain. During Stage-1, it is necessary to optimize μ and σ 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 $\widetilde{\mu}$ and $\widetilde{\sigma}$, this process results in μ and σ not being optimized, leading to a trivial solution. Therefore, we may not integrate rectification into Stage-1. A more detailed simplification process is shown in the Appendix.

Is ULDA a degraded form of domain generalization? Our propose ULDA is not a degraded form of domain generalization (DG) [24]. Because the ULDA does not conflict with DG; rather, it complements them. DG methods typically utilize meta-learning or feature alignment techniques to incorporate domain-invariant information from source data during the training phase within the source domain. In contrast, ULDA focuses on enhancing a pre-trained model, enabling it to generalize more efficiently and effectively across a wider range of target domains. Subsequent experiments in the supplementary demonstrate that our proposed method can also yield benefits for DG methods.

7. Concluding Remarks

This work spots issues in the literature and presents a new setting named Unified Language-driven Zero-shot Domain Adaptation (ULDA) with three simple yet effective strategies Hierarchical Context Alignment (HCA), Domain Consistent Representation Learning (DCRL), and Text-Driven Rectifier (TDR). The effectiveness and practical merits of our method have been verified by the decent performance achieved by challenging benchmarks without imposing any additional inference burdens.

Acknowledgements. This work receives partial support from the Shenzhen Science and Technology Program under No. KQTD20210811090149095.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1

[2] 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. 1

[3] 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. 6

[4] 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. 6

[5] Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7212–7222, 2022. 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, 2, 3, 4, 6, 7

[7] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 1

[8] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022. 1

[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. 1

[10] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1

[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. 1

[12] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. *arXiv preprint arXiv:2112.00374*, 2021. 1, 7

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

[14] Gihyun Kwon and Jong Chul Ye. One-shot adaptation of gan in just one clip. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[15] 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. 1

[16] Hongbin Lin, Yifan Zhang, Zhen Qiu, Shuaicheng Niu, Chuang Gan, Yanxia Liu, and Mingkui Tan. Prototype-guided continual adaptation for class-incremental unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 351–368. Springer, 2022. 1

[17] 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. 1

[18] 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. 1

[19] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[20] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 6

[21] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021. 6

[22] 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. 1

[23] Fan Wang, Zhongyi Han, Yongshun Gong, and Yilong Yin. Exploring domain-invariant parameters for source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2022. 1

[24] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 8

[25] 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. 6

[26] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023. 1

[27] Senqiao Yang, Jiarui Wu, Jiaming Liu, Xiaoqi Li, Qizhe Zhang, Mingjie Pan, Yulu Gan, Zehui Chen, and Shanghang Zhang. Exploring sparse visual prompt for domain adaptive dense prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16334–16342, 2024. 1

[28] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, 2020. 1