This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

PiCaZo: Pixel-Aligned Contrastive Learning for Zero-Shot Domain Adaptation

Aniruddh Sikdar* [1] , Arya Kishore* [2], Ishika Kadam [4] , Suresh Sundaram[1,3]
[1]Robert Bosch Centre for Cyber Physical Systems, Indian Institute of Science, Bengaluru, India
[2] Indian Institute of Science, Bengaluru, India
[3]Department of Aerospace Engineering, Indian Institute of Science, Bengaluru, India
[4]Birla Institute of Technology And Science, Pilani, India
{aniruddhss, vssuresh} @iisc.ac.in

Abstract

Domain adaptation has been extensively studied in computer vision but typically requires access to target images during training, which may not always be practical. Zero-shot domain adaptation overcomes this limitation by adapting CLIP-based model trained on a source domain using only a natural language description of the target domain, i.e., a prompt, eliminating the need for target domain images. Existing approaches depend on domain IDs and domain-specific models, requiring prior knowledge to choose the correct model and train separate models for each target domain using CLIP text embeddings, limiting their flexibility. To address these limitations, we propose PiCaZo, a novel Pixel-aligned Contrastive Learning Approach for Zero-shot Adaptation. To capture multi-perspective representations in the style of potential target domains and incorporate local contextual cues across diverse domains, we introduce VOLT-PIN alongside Contrastive NightVision Learning. During model adaptation, Contrastive Domain Alignment (CDA) ensures alignment of text-rectified rain and fog features, while Fog-Adaptive Contrastive Training (FACT) improves decoder adaptation. These components operate in synergy to refine the simulation process, ensuring better alignment between simulated features and target text for more effective adaptation. Extensive semantic segmentation experiments show that PiCaZo surpasses the baseline by an average of 3.6% across three benchmark datasets, all while preserving inference-time model complexity. It also achieves superior performance compared to state-of-the-art methods.

1. Introduction

Deep learning models for semantic segmentation must be resilient to domain shifts and exhibit strong generalization capabilities, supporting applications in autonomous driving [6, 42] and robot navigation [2]. A robust segmentation model should consistently perform well across diverse lighting [36], weather [37, 38], and geographic conditions [34, 35]. However, these models often suffer substantial performance drops when target domain test data diverges from the training distribution, highlighting their limited generalization across diverse domains. In particular, when tested on out-of-distribution data, they tend to suffer substantial performance drops [27], limiting their effectiveness in real-world scenarios such as autonomous driving in dynamic and unpredictable environments.

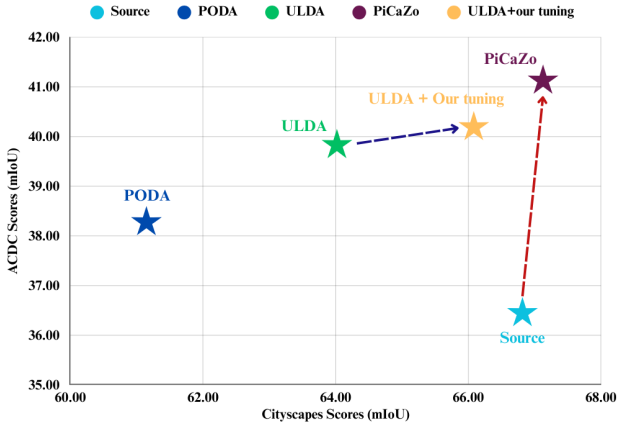


Figure 1. Comparison of PiCaZo with other models on the source domain (Cityscapes) and target domain (ACDC), showing consistent improvements. ULDA further improves with our VOLT-PIN tuning.

Unsupervised Domain Adaptation (UDA) aims to bridge the domain gap by learning domain-invariant representations that align a labeled source domain with a predefined, unlabeled target domain [15, 16, 25, 40, 41]. However, practical constraints such as privacy and security concerns, cost and logistical challenges, or data scarcity can limit di-

*Equal contribution.

rect access to target data [42]. Moreover, even when target domain data is available, it may not sufficiently capture the diverse variations encountered in real-world scenarios, thereby constraining the scalability and effectiveness of these approaches and presenting a fundamental challenge for robust generalization.

Recent progress in vision-language foundational models [3, 24, 28] has greatly improved the interaction between image-text pairs, enabling robust adaptation to novel concepts. CLIP [28], a vision-language model (VLM) trained on 400 million internet-sourced image-text pairs, exemplifies these advancements. VLMs excel in adaptability, open-set recognition, and cross-domain transfer, as highlighted by LCA [33], demonstrating superior generalization over vision-only models. This has inspired numerous studies that harness the zero-shot capabilities of VLMs to address domain adaptation challenges [15, 18, 22].

Notably, Prompt-driven Zero-shot Domain Adaptation (PØDA) [12] stands out by generating representative target domain features without requiring target domain images, while preserving pixel-wise semantics for effective adaptation to unseen domains. By using language embeddings from CLIP [28] to simulate target domain visual representations and adapting the model accordingly, PØDA eliminates the need for target images and introduces zero-shot adaptation to create novel domains using text embeddings, relaxing the constraints of UDA. However, a key limitation of PØDA is its reliance on prior knowledge, specifically the domain *ID*, to select the appropriate domain-specific model. Separate models are trained for each target domain, leveraging CLIP text embeddings for guidance. During inference, the system first determines the current target domain using a Domain ID and then selects the corresponding model. However, the reliance on domain-specific customization imposes constraints on flexibility, hindering the model's seamless adaptability across diverse and dynamic scenarios.

To generalize across diverse target domains without relying on explicit cues like domain ID during inference, ULDA [45] enables a single model to adapt seamlessly, ensuring robust performance across multiple domains. During model adaptation with simulated features, addressing discrepancies with the actual target domain is crucial, as their direct use may shift the segmentation head from true target distributions. ULDA introduces the Text-Driven Rectifier (TDR) scheme, which addresses this issue by refining the simulation process to better align simulated and target features. However, the learnable parameters in TDR are optimized solely with segmentation loss, receiving limited supervision. PODA and ULDA use text embeddings for adaptation without target images, boosting generalization. However, both approaches rely on frozen CLIP encoders, without fine-tuning on downstream datasets or tasks.

CLIP's image-level supervision equips it with strong global representations, making it effective for high-level vision tasks but limiting its performance in pixel-level segmentation [11, 48]. Semantic segmentation demands a nuanced balance between global scene comprehension and fine-grained local detail [30, 43, 46]. However, CLIP's lack of localized supervision often results in dense feature representations misaligned with textual embeddings, leading to the capture of noisy and semantically inconsistent regions [23, 39, 48]. Fine-tuning CLIP for segmentation enhances adaptation to domain-specific variations but frequently leads to the loss of prior knowledge [11]. Parameter-efficient fine-tuning methods adapt with minimal parameter changes while keeping the pretrained backbone frozen, with prompt tuning methods [7, 49, 50] and adapter-based approaches [17, 47] being constrained by architectural limitations. In addition to the increased computational cost of full backpropagation and longer runtimes, these methods are specifically designed for transformer layers and do not naturally extend to convolutional networks [13].

The key challenge is effectively fine-tuning while preserving out-of-domain generalization and minimizing distortion, as performance degrades under substantial distribution shifts [20]. Recent work explores fine-tuning techniques aimed at preserving CLIP's robustness under distribution shifts. ProLIP [13] introduces a fine-tuning approach that surpasses state-of-the-art prompt learning and adapter-based methods, demonstrating the effectiveness of fine-tuning strategies. However, these methods are primarily limited to classification tasks [14]. As observed in PODA, making these encoders trainable frequently causes a performance drop that must be mitigated. DGSS [14] adopts a strategy that allows only the last convolutional block of the CLIP vision backbone to be trainable, emphasizing the significance of minimal fine-tuning. It observes a substantial performance drop with vanilla fine-tuning and DP or DP-FT strategies [31]. Fine-tuning is crucial for extending CLIP beyond image-level supervision, enhancing performance, robustness, and real-world applicability. These findings underscore the importance of strategic fine-tuning to preserve CLIP's effectiveness in semantic segmentation.

To address this, we propose PiCaZo, a novel **Pi**xel-aligned **C**ontrastive learning **A**pproach for **Z**er**O**-shot domain adaptation. During the zero-shot feature augmentation stage, to train the CLIP vision encoder E_{img} with multi-perspective views of the same objects and incorporate local contextual information across different target domain styles, we introduce VOLT-PIN strategy alongside Contrastive NightVision Learning. Robustness to distribution shifts depends on preserving domain-invariant knowledge. However, CLIP's image encoder generates input-specific representations rather than being inherently domain-invariant, with its strong generalization across

datasets and domains primarily attributed to robust text embeddings [1]. To bridge the gap between pretraining and fine-tuning on supervised pixel classification, VOLT-PIN trains only the CLIP vision encoder while keeping the text encoder frozen. VOLT-PIN operates without introducing additional trainable parameters, leveraging high-quality pretrained features to enable effective fine-tuning even under significant distribution shifts. Contrastive NightVision Learning enhances intra-domain representation learning, refining night-specific feature adaptation. During the model adaptation stage, Contrastive Domain Alignment (CDA) aligns text-rectified rain and fog features, while Fog-Adaptive Contrastive Training (FACT) further refines the decoder for fog-specific features. It further refines the simulation process to improve alignment between simulated and target features, providing additional supervision to the learnable parameters in TDR. This cultivates a more distinct embedding space, improving model adaptation as observed in Fig. 1. The key contributions of this paper are:

- We introduce PiCaZo, a novel framework for zero-shot domain adaptation. In the zero-shot feature augmentation stage, we propose VOLT-PIN with Contrastive Learning for Night-time Enhancement to tackle significant distribution shifts, enabling effective fine-tuning and enhancing intra-domain representation learning for improved night-specific feature adaptation.
- In the model adaptation stage, we propose Contrastive Domain Alignment (CDA) alongside Contrastive Learning-Based Training for Fog-Mined Decoder Adaptation, leading to improved model robustness.
- Experimental evaluations on three public datasets show that the proposed PiCaZo consistently outperforms state-of-the-art models, especially under challenging weather conditions. It achieves this while preserving the baseline model's complexity and delivering enhanced performance even in the source domain.

2. Related Works

Domain adaptation Unsupervised Domain Adaptation (UDA) seeks to minimize the domain gap by learning domain-invariant representations, aligning a labeled source domain with a predefined, unlabeled target domain [15, 25, 40]. Recently, One-Shot Unsupervised Domain Adaptation (OSUDA) has emerged [4], posing the challenge of adapting models with only a single target domain image. To tackle this, Style Mixing and Patch-Wise Prototypical Matching (SM-PPM) [44] has been introduced. For the more challenging Zero-Shot Domain Adaptation setting, no target images are available for adaptation. Recently, [12] proposed Prompt-Driven Zero-Shot Domain Adaptation (PODA), which eliminates the need for target domain data by leveraging natural language descriptions to adapt the model to new environments. ULDA [45] tackles the challenge of refining a model's embedding space to accommodate multiple domains while preserving semantic discriminability across categories.

CLIP Fine-Tuning for Downstream Tasks Parameter-efficient fine-tuning adapts models with minimal parameter changes while keeping the pretrained backbone frozen. However, these methods are tailored for transformer layers and do not seamlessly extend to convolutional networks [13]. [20] studies lightweight fine-tuning, updating a small model portion for better robustness, though it may distort pretrained features and hurt OOD generalization. ProLIP [13] uses CLIP's text-image compatibility without introducing learnable parameters or altering the model architecture. It trains a linear classifier on frozen features, outperforming state-of-the-art prompt tuning and adapters, showcasing fine-tuning effectiveness. Recent fine-tuning methods maintain CLIP's robustness under shifts but are limited to classification [14]. These fine-tuning approaches depend on the specific downstream task [14]. PODA enables zero-shot segmentation with frozen CLIP encoders, but fine-tuning deeper vision layers degrades performance. DGSS [14] allows only the last convolutional block of the CLIP vision backbone to be trainable, emphasizing the value of minimal fine-tuning. It observes a drastic drop in performance when following DP, DP-FT strategies [31]. The results highlight the need for strategic fine-tuning strategies to preserve CLIP's robustness in semantic segmentation.

3. Preliminaries

We review PØDA [12], a zero-shot domain adaptation method that uses textual prompts, *TrgPrompt*, eliminating the need for target domain images. In Stage-1, the model undergoes Zero-shot Feature Augmentation to generate target domain features, while the Model Adaptation stage fine-tunes the segmentation head using these features. The segmentation model *M* consists of an encoder M_{feat} followed by a decoder and classification head M_{cls}, i.e., M=(M_{feat},M_{cls}). The encoder is the CLIP image encoder E_{img} (ResNet-50 backbone), while the textual prompts are encoded using the CLIP text encoder E_{text}. The target domain embedding, denoted as *TrgEmb*, is defined as *TrgEmb* = E_{text}*(TrgPrompt)*, representing the CLIP-generated text embedding for the target domain prompt.

Zero-shot Feature Augmentation introduces Prompt-driven Instance Normalization (PIN), as defined in Eq. 1, where learnable variables μ and σ are guided by a text domain prompt to simulate target domain characteristics, extracting style statistics from the imaginary target feature map f_t, given by,

$$f_{s\to t} = PIN(f_s, \mu, \sigma) = \sigma\left(\frac{f_s - \mu(f_s)}{\sigma(f_s)}\right) + \mu \quad (1)$$

where, $\mu(f_s)$ and $\sigma(f_s)$ denote the mean and standard

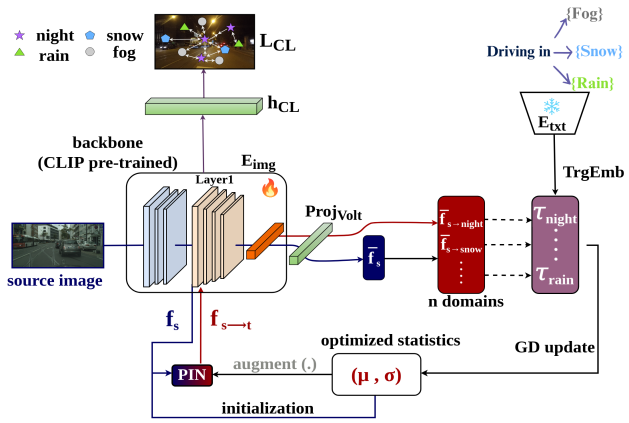


Figure 2. Target extraction from the source image through zero-shot feature augmentation. The source image is processed through the CLIP image encoder to obtain $\bar{f}s \to t$ CLIP embedding. $PROJ_{VOLT}$ projects the features, enabling the concurrent extraction of multiple target style variations, aligning with corresponding *TrgEmb* for different domains $\{\tau_{night}, ..\tau_{rain}\}$. Contrastive NightVision Learning (CNL) applies L_{CL} for pixel-level supervised contrastive learning, enhancing night-specific feature adaptation.

deviation of the source input features, (f_s). This is depicted as the *augment*$(\cdot)$ operation in Fig. 2, which integrates target domain style into f_s, generating $f_{s\to t}$ while preserving its dimensions and semantic integrity. PIN transforms source domain features into the target domain $f_{s\to t}$. Finally, these features are projected into the CLIP latent space, resulting in $\bar{f}_{s\to t}$. For style mining, the loss function in Equation 2 is optimized for (μ, σ) to enhance the similarity between $\bar{f}_{s\to t}$ and the CLIP text embeddings *TrgEmb*, given by,

$$L_{\mu,\sigma}(\bar{f}_{s\to t}, TrgEmb) = 1 - \frac{\bar{f}_{s\to t}.TrgEmb}{\| \bar{f}_{s\to t} \| \| TrgEmb \|} \quad (2)$$

This promotes alignment between the transformed features and textual representations, supporting successful adaptation from the source to the target domain. In order to preserve the CLIP latent space, M_{feat} is frozen in PODA.

Model Adaptation PODA fine-tunes the segmentation head to improve the model's adaptation to target domain. During this stage, only source domain images and the mined target style (μ, σ) are used, selected based on the specific domain. The simulated target domain features are leveraged to fine-tune the segmentation head L_{seg}, refining the classifier M_{cls} for effective adaptation to the target domain. The training in this stage is guided by the segmentation loss, i.e, cross-entropy loss between the predictions and the ground-truth labels. During the backward pass, only the weights of M_{cls} are updated using the loss gradients, resulting in the adapted model, $M^{'} = (M_{feat}, M^{'}_{cls})$.

4. Methodology

Problem Formulation UDA requires target domain data, which may not always be feasible. Zero-shot domain adaptation enhances a model's generalization without the need for actual target data. Model *M* is trained on the source domain D_s using x_s training samples with pixel-level labels for semantic segmentation task, without access to target domain data D_t, across *n* such use unseen target domains. *M* uses pre-defined textual prompts *TrgPrompt*, to interpret target domain scenarios. The objective is to extract relevant information from textual descriptions to adapt source visual features across different domains while preserving the meaningful semantic consistency of the source domain. Unlike PODA, which uses separate models for individual scenarios, our goal is to train a single model capable of adapting to a broad spectrum of domain conditions without relying on specific domain *IDs*.

4.1. PiCaZo: Pixel-aligned Contrastive learning for Zero-shot adaptation

Zero-shot Feature Augmentation To train the CLIP vision encoder E_{img}, with multi-perspective about same objects, incorporating local contextual information for different target domains, we propose the VOLT-PIN along with Contrastive NightVision Learning (CNL).

(i) VOLT-PIN (Vision-Only, Locked-Text with PIN): Semantic segmentation relies on the sophisticated understanding of both global context and local context for fine-grained details [30, 43, 46]. However, CLIP's image-level supervision often results in dense local features that capture noisy regions for a given concept [23, 39, 48]. To refine the encoder for the downstream task, we fine-tune it on the source training data using the VOLT-PIN (Vision-Only, Locked-Text with PIN) strategy.

Since CLIP's strong generalization ability is largely attributed to its robust text embeddings [1], we keep the CLIP text encoder E_{text} frozen while fine-tuning all layers of the CLIP vision encoder E_{img} on the source training data. Unlike PODA [12] and ULDA [45], which use frozen encoders relying on noisy local features within the CLIP embedding space for segmentation, we fine-tune the entire CLIP vision encoder, including the first layer used for style mining by PIN, to enhance local context perception. During each training iteration, *n* target style variations are mined concurrently, and the features for each potential target domain, obtained after the attention pooling layer of E_{img}, are projected into the text embedding space via a projection layer, denoted by

$$\bar{f}_{s\to t} = Proj_{VOLT}(\bar{f}_{s\to t}) \quad (3)$$

For style mining, the loss function in Equation 2 optimizes (μ, σ) to align *TrgEmb* with the projected features.

Unlike PODA [12], we project all features simultaneously to align with their respective *TrgEmb*, as shown in Fig. 2. As all potential target styles are mined concurrently in each iteration, each concept is observed from multiple target perspectives, integrating local information from the vision encoder with text features for a richer understanding of visual concepts. To align $\bar{f}_{s\to t}$ with target text embeddings while maintaining distinct semantic characteristics across categories, we use the regional alignment L_r and domain consistency (L_{DC}) losses from ULDA [45]. These losses are useful for training E_{img} and enhancing style mining.

(ii) Contrastive NightVision Learning (CNL): To refine intra-domain representation learning for night-specific features, we propose pixel-level supervised contrastive learning framework for enhanced night-time feature discrimination. This process fine-tunes the E_{img} to simultaneously learn night-domain-specific style representations, which are iteratively mined using PIN augmentation. When mining only the night domain style specifically for the CLIP visual features $\bar{f}_{s\to night}$, we employ pixel-level contrastive learning L_{CL}, as shown Fig. 2.

Features from first four layers are upsampled and mapped into an embedding space via a projection head h_{CL}. Supervised contrastive learning [19] is employed to cluster pixel embeddings of the same category while separating those of different categories. Using pixel-wise labels, pixels of the same class serve as positive samples, whereas pixels from different classes are treated as negative samples. The pixel-wise contrastive loss is given by,

$$L_{CL} = -\sum_{C(i)=C(j)} log\frac{r(e_i, e_j)}{\sum_{k=1}^{Np} r(e_i, e_j)} \tag{4}$$

Here, e_i denotes the i^{th} feature map derived from the projection head, while N_p represents the total number of pixels. The similarity measure, *r*, is computed using an exponential similarity function: r(e_i, e_j) = exp(s(e_i, e_j) / τ), where *s* denotes cosine similarity and τ is the temperature parameter. While various sampling strategies can be applied, we empirically find that the semi-hard example sampling strategy [9] delivers the best performance. Given that night is the most challenging domain for model adaptation, we prioritize training the encoder layers specifically for this setting. Emphasizing night-domain adaptation through contrastive learning refines semantic representation, fostering enhanced generalization across diverse domains.

Model Adaptation Discrepancies between simulated and actual target features can misalign the segmentation head, degrading performance. To address this, ULDA introduces a *text-driven rectifier (TDR)* in stage 2 of the model adaptation, defined as

$$\tilde{f}_{s\to t} = \beta(\tilde{\sigma}(\frac{\bar{f}_{s\to t} - \mu(\bar{f}_{s\to t})}{\sigma(\bar{f}_{s\to t})}) + \tilde{\mu}) + \bar{f}_{s\to t} \tag{5}$$

where β is a learnable parameter. $\tilde{\sigma}$ and $\tilde{\mu}$ are derived by passing CLIP text embeddings through a projection layer to capture the target domain features. $\tilde{f}_{s\to t}$ are obtained from multiple domains to fine-tune the head. TDR preserves domain uniqueness, enhancing generalization and multi-domain adaptation. However, the learnable parameters in TDR are optimized solely using segmentation loss. To further enhance training, we introduce an additional contrastive learning objective. Specifically, we propose Contrastive Domain Alignment (CDA), which comprises class-wise contrastive learning for Text-rectified rain & fog features and Fog-Adaptive Contrastive Training (FACT) for fog-mined decoder adaptation.

(iii) Contrastive Domain Alignment (CDA): We aim to construct a discriminative embedding space for each segmentation class, with a stronger emphasis on fog and rain-mined styles for improved learning. To achieve this, we optimize the text-rectified features $\tilde{f}_{s\to fog}$ and $\tilde{f}_{s\to rain}$ using a supervised contrastive loss. The feature $\tilde{f}_{s\to fog}$, denoted as $F_{(m,n)}$, is paired with $\tilde{f}_{s\to rain}$ as an anchor, represented by $F_{a(m,n)}$, where (m, n) indicates the pixel position in the feature map, as both retain the same spatial content. Negative samples are determined using a resized segmentation label y_D. The features are projected into a shared embedding space via the projection head π, as shown in Fig. 3 and trained using the loss defined below,

$$L_{CDA} = \sum L_{IN}(F_{(m,n)}, F_{a(m,n)}, F_{(p,q)}) \tag{6}$$

where, L_{IN} represents the InfoNCE loss [26], where *(p,q)* $\in$ { *(p,q)* $\in$ *P* | $y_{(p,q)} \neq y_{(m,n)}$ }, with *P* denoting all pixel positions in the feature map.

(iv) Fog-Adaptive Contrastive Training (FACT): We employ a unique approach to fine-tune M_{cls} using $\tilde{f}_{s\to fog}$. This involves optimizing the text-rectified features $\tilde{f}_{s\to fog}$ alongside the layer preceding the prediction head in M_{cls}, denoted as X^q through a supervised contrastive loss. The feature $\tilde{f}_{s\to fog}$, denoted as $F_{(m,n)}$, is paired with X^q as an anchor, represented by$X^q_{a(m,n)}$, where (m, n) indicates the pixel position in the feature map. Negative samples are identified using a resized segmentation label. The features are then mapped into a shared embedding space through projection head, as shown in Fig. 3 and optimized using the loss function defined below,

$$L_{FACT} = \sum L_{IN}(F_{(m,n)}, X^q_{a(m,n)}, X^q_{a(p,q)}) \tag{7}$$

where *(p,q)* $\in$ { *(p,q)* $\in$ *P* | $y_{(p,q)} \neq y_{(m,n)}$ }, with *P* denoting all pixel positions in the feature map.

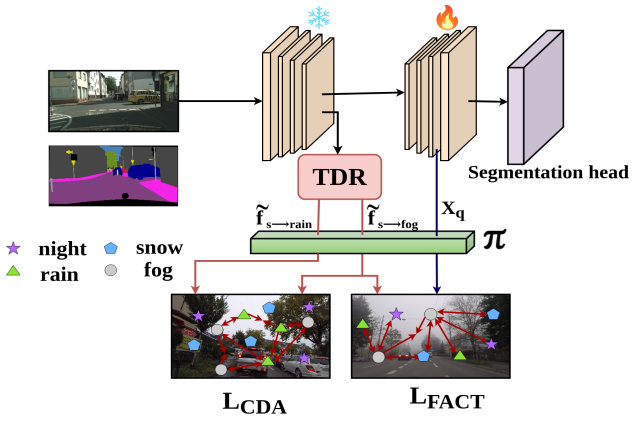


Figure 3. Model adaptation with simulated features. Contrastive Domain Alignment (CDA) aligns fog and rain features from TDR, while Fog-Adaptive Contrastive Training (FACT) fine-tunes M_{CLS} using fog-mined features.

Overall loss function: During the zero-shot feature augmentation stage, E_{img} remains fully trainable, and the overall loss function L for this stage is defined as follows,

$$L = L_{\mu,\sigma}(\bar{f}_{s\to t}, TrgEmb) + L_r + L_{DC} + L_{CL} \quad (8)$$

Once the styles are effectively mined, we adopt PODA's strategy for model adaptation, training only M_{cls} to ensure optimal adaptation. The overall loss function during this stage is,

$$L = L_{seg} + L_{CDA} + L_{FACT} \quad (9)$$

where, L_{seg} represents the segmentation loss, which is computed using the cross-entropy loss. The proposed framework generates an adapted model $M^{'}$, which doesn't require prior knowledge of upcoming domains and maintains the same complexity as the baseline model.

5. Experiments

5.1. Experimental Setup

Datasets and Evaluation For the experimental evaluation, Cityscapes [10] is used as the source domain, containing 2,975 training images and 500 validation images. The evaluation is conducted on three datasets: ACDC [32], GTA5 [29], and MUSES [5]. The evaluation is conducted using the official validation sets, with ACDC comprising 406 samples. Evaluation on GTA5 is conducted using 1000 images, adhering to the protocol in [12, 45]. Additionally, we validate it on the MUSES dataset using the official validation set, which consists of 250 images. Both ACDC and MUSES comprise urban scenes captured under challenging weather conditions, including snow, rain, fog, and nighttime settings. Consistent with previous works, all datasets utilize a shared set of 19 semantic classes. We evaluate all models on target images at their original resolutions. For PiCaZo and other reproduced results, we report the mean performance across three models trained with different random seeds. Mean Intersection over Union (mIoU %) is used to evaluate adaptation performance on target images. *

Implementation Details. The base segmentation model (*source model*), DeepLabv3+ [8], with a pre-trained CLIP-ResNet-50 backbone [28] , is adopted and thoroughly trained on the source domain following the configuration of [12]. ImageNet templates from [28] are used to encode the target descriptions in *TrgPrompt*. For zero-shot feature augmentation, we adopt the training procedure of PODA [12] for PIN, using a batch size of 4. During the model adaptation stage, we fine-tune the pre-trained source model with augmented PIN features, focusing on M_{cls} for 2000 iterations with a batch size of 4, following the approach in [12]. The SDG optimizer is used to optimize different learnable layers in both stages. The projection layer $Proj_{VOLT}$ projects the feature maps to $\mathcal{R}^{256\times1\times1}$, while h_{CL} maps them to $\mathcal{R}^{256\times192\times192}$. Meanwhile, the projection layer π further projects them to $\mathcal{R}^{19\times10\times512}$.

5.2. Comparison with state of art

Comparison with 0-shot adaptation methods Tables 1 and 2 present a comparison of state-of-the-art models with PiCaZo on MUSES, ACDC, and GTA5, including the source-only model, CLIPstyler [21] for zero-shot style transfer, as well as zero-shot adaptation methods such as PODA [12] and ULDA [45]. *TrgPrompt* denotes prompts that describe each target domain, as shown in both tables. CLIPstyler generates images reflecting the input text's style but often introduces artifacts that hinder their usability for segmentation tasks. PiCaZo surpasses PODA by an average of 2.965 % and ULDA by 1.825 % on the ACDC dataset, highlighting its effectiveness. This highlights the model's adaptability to challenging rain and fog scenarios.

Comparison to one-shot UDA PiCaZo is evaluated against the state-of-the-art OSUDA method, SM-PPM [44], which uses a single unlabeled target domain image as an anchor point for target style mining, as shown in Table 7. SM-PPM requires separate images for each domain to adapt three domain-specific models. Due to differences in backbone (ResNet-101 vs. ResNet-50) and segmentation framework (DeepLabv2 vs. DeepLabv3+), the models' absolute results are not directly comparable. Instead, we assess each method's improvement over its source-only baseline. PiCaZo surpasses SM-PPM in both absolute and relative gains over the source-only baseline, achieving superior performance despite being purely zero-shot.

Effects of Irrelevant Prompts Following PODA [12], we query six irrelevant prompts using ChatGPT, such as *"mes-*

*Reported results have been reproduced using open-source code.

Scenarios		Source2Fog	Source2clear	Source2Rain	Source2Snow	Mean (mIoU)
TrgPrompt		driving in fog	driving in clear	driving under rain	driving in snow	
Method	Ref	mIoU	mIoU	mIoU	mIoU	
Source*	OpenAI	19.87	41.06	28.89	31.77	30.40
PODA*	ICCV 2023	24.38	**42.22**	31.18	34.29	33.02
ULDA*	CVPR 2024	**26.98**	39.38	30.69	34.05	32.78
PiCaZo	Ours	25.70 (↓ 1.27)	41.52 (↓ 0.7))	**32.29** (↑ +1.11)	**36.56** (↑ +2.27)	**34.023** (↑ +1)

Table 1. Performance comparison (mIoU%) of zero-shot domain adaptation for semantic segmentation against state-of-the-art models and the source-only baseline. Cityscapes is used as the source domain, while MUSES serves as the four target domains. Mean-mIoU represents the average mIoU across all four target domains. * indicates results reproduced using official code.

merizing northern lights display", *"playful dolphins in the ocean"*, *"breathtaking view from mountaintop"*, *"cheerful sunflower field in bloom"*, *"dramatic cliff overlooking the ocean"* and *"majestic eagle in flight over mountains"*, to assess potential negative transfer. The reported PODA performance represents the average across all domains. Both PODA and PiCaZo primarily gain from generalization via randomization rather than adaptation when using irrelevant prompts, as seen in Table 3, with PiCaZo exhibiting a more substantial improvement over the source-only model. This could be due to the enhanced capability of the CLIP vision encoder for downstream datasets. However, using meaningful prompts still yields the best performance.

Qualitative results Figure 4 presents qualitative results, with training on Cityscapes and validation on the ACDC dataset.

5.3. Ablation Study

Partial freezing of backbones We evaluate the impact of freezing different encoder layers in table 6 as part of the VOLT-PIN strategy. To ensure a systematic analysis, we perform ablations without incorporating any contrastive learning losses. As observed, fine-tuning all layers along with the projection layer delivers the best overall performance.

Contrastive learning with different domains We conduct an ablation study on Contrastive NightVision Learning (CNL) during zero-shot feature augmentation across different domain styles, as presented in table 5. Notably, the best performance is achieved when prioritizing the night domain, where the model initially struggles the most.

Ablation with key components Table 4 showcases the impact of different losses in PiCaZo. As observed, the combination of all losses results in the highest performance.

6. Conclusions.

This paper presents PiCaZo, a novel Zero-Shot Domain Adaptation framework for semantic segmentation. In the zero-shot feature augmentation stage, we introduce the

Source		Target Eval.	Method	mIoU
TrgPrompt = "driving at night"				
CS	↦	ACDC - Night	source-only	18.31
			CLIPStyler	21.38
			PODA*	**23.97**
			ULDA*	21.21
			PiCaZo	22.75
TrgPrompt = "driving in snow"				
CS	↦	ACDC - Snow	source-only	39.28
			CLIPStyler	41.09
			PODA*	40.59
			ULDA*	43.01
			PiCaZo	**43.94**
TrgPrompt = "driving under rain"				
CS	↦	ACDC - Rain	source-only	38.20
			CLIPStyler	37.17
			PODA*	39.20
			ULDA*	41.96
			PiCaZo	**44.47**
TrgPrompt = "driving in fog"				
CS	↦	ACDC - Fog	PODA*	49.34
			ULDA*	51.53
			PiCaZo	**53.82**
TrgPrompt = "driving in a game"				
CS	↦	GTA5	source-only	39.59
			CLIPStyler	38.73
			PODA*	40.91
			ULDA*	40.93
			PiCaZo	**42.33**

Table 2. Performance comparison (mIoU%) of zero-shot domain adaptation for semantic segmentation against state-of-the-art models and the source-only baseline. Cityscapes serves as the source domain, while ACDC and GTA5 are the target domains. PiCaZo surpasses PODA by **+2.965** and **+1.42** mIoU on ACDC and GTA5, respectively, and outperforms ULDA by an average of **+1.825** and **+1.4** mIoU. TrgPrompt represents the target prompt. * denotes results reproduced using official code.

Method	ACDC (Night)	ACDC (Snow)	ACDC (Rain)	Mean IoU
Source Only	18.31	39.28	38.20	36.44
Relevant Prompts (TrgPrompt)				
PODA	23.97	40.59	39.20	38.28
PiCaZo	22.75	43.94	44.47	41.25
Irrelevant Prompts				
PODA	20.45	39.95	39.33	33.24
PiCaZo	21.52	43.54	43.25	36.10

Table 3. Performance comparison between relevant prompts (i.e., TrgPrompt) and irrelevant prompts. PiCaZo outperforms PODA on average across both prompt types.

Method	CNL	CDA	FACT	Mean (mIoU)
PiCaZo	✔	✗	✗	40.54
	✔	✔	✗	41.12
	✔	✔	✔	**41.245**

Table 4. Ablation study of various losses, including Contrastive NightVision Learning (CNL), Contrastive Domain Alignment (CDA), and Fog-Adaptive Contrastive Training (FACT).

Method	CNL	CDA	FACT	Mean (mIoU)
PiCaZo	Snow	✗	✗	40.36
	Rain	✗	✗	39.99
	Fog	✗	✗	40.05
	Night	✗	✗	40.54
	Night	✔	✔	**41.245**

Table 5. Ablation study of Contrastive NightVision Learning (CNL) during zero-shot feature augmentation across different domain styles. Notably, applying CNL to the night domain yields the highest overall performance on the ACDC dataset.

Method	Encoder				Mean (mIoU)
	$L1$	$L2$	$L3$	$L4$	
PiCaZo	✗	✗	✗	✗	39.41
w\o	✗	✗	✗	✔	39.63
C.L	✔	✔	✔	✔	40.10
PiCaZo	✔	✔	✔	✔	**41.24**

Table 6. Impact of freezing encoder layers during zero-shot augmentation on the ACDC dataset (excluding contrastive learning losses), where ✗ indicates a frozen layer and ✔ denotes a trainable layer. Performance is reported as the average across all weather conditions, highlighting the benefits of keeping all encoder layers (L) trainable. (Average across all domains is reported. $L1$ denotes Layer 1, and subsequent layers follow accordingly.)

VOLT-PIN strategy combined with Contrastive NightVision Learning. The VOLT-PIN strategy enhances fine-tuning for

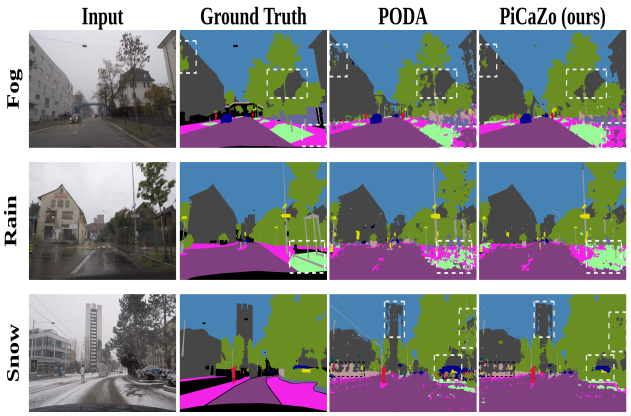


Figure 4. Qualitative comparison of PiCaZo and PODA on the ACDC dataset.

Source		Target Eval.	One-shot SM-PPM	Zero-shot PiCaZo
CS	$\mapsto$	ACDC Night	13.07 / 14.60 (Δ=1.53)	18.31 / 22.75 (Δ=4.44)
CS	$\mapsto$	ACDC Snow	32.60 / 35.61 (Δ=3.01)	39.28 / 43.94 (Δ=4.66)
CS	$\mapsto$	ACDC Rain	29.78 / 32.23 (Δ=2.45)	38.20 / 44.47 (Δ=6.27)

Table 7. Segmentation performance (mIoU %) for source/adapted models, with Δ indicating the improvement from adaptation. SM-PPM uses a single target domain image to adapt three domain-specific models, whereas PiCaZo operates without any target domain access.

segmentation task, enabling effective adaptation even under substantial distribution shifts. During the model adaptation stage, Contrastive Domain Alignment (CDA) ensures alignment of text-rectified rain and fog features, while Fog-Adaptive Contrastive Training (FACT) further enhances the decoder for fog-specific features. This cultivates a more discriminative embedding space, enhancing model adaptation and optimizing decoder performance. PiCaZo consistently delivers top performance across three open-source datasets, outperforming the source model by an average of 3.6%, surpassing ULDA by 1.416%, and exceeding PODA by 1.796% across all datasets. It demonstrates enhanced overall robustness while maintaining the same model complexity as the baseline segmentation model.

References

[1] Sravanti Addepalli, Ashish Ramayee Asokan, Lakshay Sharma, and R Venkatesh Babu. Leveraging vision-language models for improving domain generalization in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23922–23932, 2024. 3, 4

[2] Aayush Agrawal, Aniruddh Sikdar, Rajini Makam, Suresh Sundaram, Suresh Kumar Besai, and Mahesh Gopi. Syn2real domain generalization for underwater mine-like object detection using side-scan sonar. *IEEE Geoscience and Remote Sensing Letters*, 2025. 1

[3] 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. 2

[4] 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. 3

[5] Tim Brödermann, David Bruggemann, Christos Sakaridis, Kevin Ta, Odysseas Liagouris, Jason Corkill, and Luc Van Gool. Muses: The multi-sensor semantic perception dataset for driving under uncertainty. *arXiv preprint arXiv:2401.12761*, 2024. 6

[6] Prithvijit Chattopadhyay, Kartik Sarangmath, Vivek Vijaykumar, and Judy Hoffman. Pasta: Proportional amplitude spectrum training augmentation for syn-to-real domain generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19288–19300, 2023. 1

[7] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. 2

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 6

[9] Mu Chen, Zhedong Zheng, Yi Yang, and Tat-Seng Chua. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1905–1914, 2023. 5

[10] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 6

[11] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 2

[12] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. Poda: Prompt-driven zero-shot domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18623–18633, 2023. 2, 3, 4, 5, 6

[13] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul de Charette. Fine-tuning clip's last visual projector: A few-shot cornucopia. *arXiv preprint arXiv:2410.05270*, 2024. 2, 3

[14] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. A simple recipe for language-guided domain generalized segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23428–23437, 2024. 2, 3

[15] 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, 2, 3

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 1

[17] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 2

[18] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[19] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 5

[20] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022. 2, 3

[21] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18062–18071, 2022. 6

[22] Gihyun Kwon and Jong Chul Ye. One-shot adaptation of gan in just one clip. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12179–12191, 2023. 2

[23] Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. In *European Conference on Computer Vision*, pages 264–282. Springer, 2024. 2, 4

[24] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2

[25] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4399–4409, 2021. 1, 3

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[27] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019. 1

[28] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 6

[29] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016. 6

[30] Darwin Saire and Adin Ramirez Rivera. Global and local features through gaussian mixture models on image semantic segmentation. *IEEE Access*, 10:77323–77336, 2022. 2, 4

[31] Kuniaki Saito, Donghyun Kim, Piotr Teterwak, Rogerio Feris, and Kate Saenko. Mind the backbone: Minimizing backbone distortion for robust object detection. *arXiv preprint arXiv:2303.14744*, 2023. 2, 3

[32] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 6

[33] Jia Shi, Gautam Gare, Jinjin Tian, Siqi Chai, Zhiqiu Lin, Arun Vasudevan, Di Feng, Francesco Ferroni, and Shu Kong. Lca-on-the-line: Benchmarking out-of-distribution generalization with class taxonomies. *arXiv preprint arXiv:2407.16067*, 2024. 2

[34] Aniruddh Sikdar, Sumanth Udupa, Suresh Sundaram, and Narasimhan Sundararajan. Fully complex-valued fully convolutional multi-feature fusion network (fc 2 mfn) for building segmentation of insar images. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 581–587. IEEE, 2022. 1

[35] Aniruddh Sikdar, Sumanth Udupa, Prajwal Gurunath, and Suresh Sundaram. Deepmao: Deep multi-scale aware overcomplete network for building segmentation in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 487–496, 2023. 1

[36] Aniruddh Sikdar, Qiranul Saadiyean, Prahlad Anand, and Suresh Sundaram. Ssl-rgb2ir: Semi-supervised rgb-to-ir image-to-image translation for enhancing visual task training in semantic segmentation and object detection. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5017–5023. IEEE, 2024. 1

[37] Aniruddh Sikdar, Jayant Teotia, and Suresh Sundaram. Skdnet: Spectral-based knowledge distillation in low-light thermal imagery for robotic perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9041–9047. IEEE, 2024. 1

[38] Aniruddh Sikdar, Jayant Teotia, and Suresh Sundaram. Ogpnet: Optical guidance meets pixel-level contrastive distillation for robust multi-modal and missing modality segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):6922–6930, 2025. 1

[39] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 2, 4

[40] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016. 1, 3

[41] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 1

[42] Sumanth Udupa, Prajwal Gurunath, Aniruddh Sikdar, and Suresh Sundaram. Mrfp: Learning generalizable semantic segmentation from sim-2-real with multi-resolution feature perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5904–5914, 2024. 1, 2

[43] Chunyan Wang, Dong Zhang, Liyan Zhang, and Jinhui Tang. Coupling global context and local contents for weakly-supervised semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2, 4

[44] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2740–2749, 2022. 3, 6

[45] Senqiao Yang, Zhuotao Tian, Li Jiang, and Jiaya Jia. Unified language-driven zero-shot domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23407–23415, 2024. 2, 3, 4, 5, 6

[46] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19456–19465, 2023. 2, 4

[47] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022. 2

[48] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 2, 4

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[50] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15659–15669, 2023. 2