2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Language-guided Attribute Alignment and Semantic Consistency for Zero-shot Domain Adaptation

Junhong Pan, Chenyi Jiang, Minxian Li, and Haofeng Zhang

***Abstract*— In cross-domain visual understanding tasks, models often achieve strong performance on the source domain but suffer severe degradation when applied to target domains with substantial distribution shifts. This challenge is particularly prominent under the zero-shot domain adaptation setting, where adaptation must be achieved without access to target-domain samples and instead relies on language guidance to bridge the gap. However, existing approaches typically depend on fixed class names or handcrafted prompt templates, which fail to capture fine-grained semantic attributes present in the target domain. Moreover, the insufficient alignment between visual and linguistic modalities further constrains the transferability of semantic knowledge. To address these issues, we propose an attribute-driven cross-modal feature modulation framework, termed Language-guided Attribute alignment and Semantic Consistency (LASC). On the semantic side, we introduce an attribute-driven prompt generation module that dynamically combines category information with domain-relevant attributes to construct adaptive text prompts, which are aligned with visual features through cross-modal attention for enhanced semantic stability. Furthermore, we incorporate a semantic consistency constraint, where a memory bank enforces intra-class compactness and inter-class separation, ensuring robust discriminability across domains. Extensive experiments demonstrate that our approach achieves significant improvements over state-of-the-art baselines on multiple cross-domain benchmarks, and maintains strong adaptation ability without requiring any target-domain data. The code is available at `https://github.com/JHP-3/LASC`.**

I. Introduction

Cross-domain visual understanding, particularly in semantic segmentation, has become a central research direction in computer vision. Its goal is to enable models trained in a source domain to maintain good generalization performance in an unseen target domain. This capability is crucial for safety-critical applications such as autonomous driving, medical imaging, and remote sensing, where collecting labeled data in every possible environment is impractical or even impossible. However, due to the significant distribution differences between the source and target domains, models often suffer from significant performance degradation when directly applied to novel environments, highlighting the necessity of robust domain generalization techniques [1].

Recent research on domain adaptation and cross-domain semantic segmentation has explored a variety of strategies to mitigate distributional shifts. Early studies primarily focused on feature alignment, such as moment matching and feature normalization, which aimed to reduce low-level statistical discrepancies across domains [2]. Later, adversarial training methods introduced domain discriminators to learn domain-invariant features, showing effectiveness but often suffering from training instability and limited semantic preservation [3]. In parallel, augmentation-based strategies diversified training data through style mixing and randomized transformations, improving robustness but leaving semantic misalignment unresolved [4]. More recently, the success of vision–language models such as CLIP [5] has spurred a line of language-guided approaches for cross-domain adaptation. These methods generally construct prompts or semantic embeddings to bridge the gap between visual and textual modalities, showing promising improvements in transferability across domains. However, most existing designs still rely on fixed class names or handcrafted templates, which limits their ability to capture domain-specific attributes and adapt to diverse target conditions.

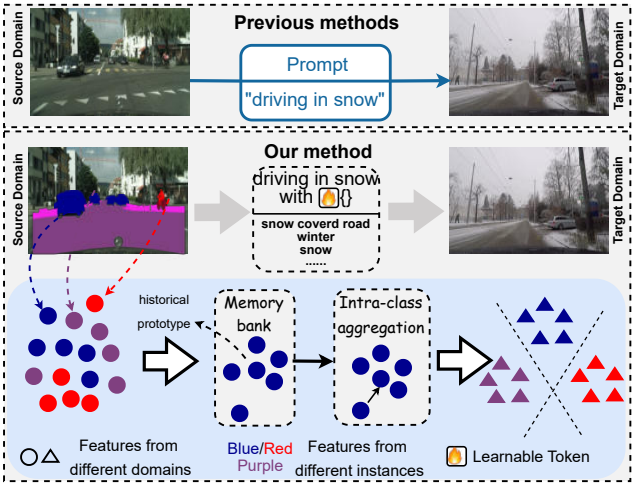

Fig. 1. Motivation of our work. Previous methods rely on fixed prompts (e.g., "driving in snow"), which fail to capture domain-specific attributes and lead to weak transfer across domains. Our method (LASC) generates attribute-driven prompts (e.g., snow-covered road, winter), enabling richer semantic representations, and incorporates a memory-based consistency mechanism that aggregates intra-class features and separates inter-class features, thus enhancing cross-domain robustness.

Despite these advances, existing prompt-based approaches still face two major limitations. First, **fixed prompts are rigid and context-insensitive**.They fail to capture variations such as adverse weather or illumination changes, and this rigidity often produces prompts that are too coarse to represent fine-grained scene properties. As a result, textual descriptions align weakly with visual features and models show poor cross-domain generalization [6]. Second, **static**

This work was supported in part by National Natural Science Foundation of China under the Grants No. 62371235 and No. U25A20444, and in part by Key Research and Development Plan of Jiangsu Province under the Grant BE2023008-2. (Corresponding author: Haofeng Zhang.)

All authors are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210094, China. zhanghf@njust.edu.cn

categories ignore attributes critical for semantic alignment. Attributes—fine-grained properties such as texture, material, or environmental conditions—are crucial in shaping visual semantics [7]. Since attributes interact with categories in complex ways (e.g., a "snow-covered road" alters both the appearance and the interpretation of the scene), overlooking them causes unstable semantic correspondences and hinders effective knowledge transfer across modalities [5], [8].

To address these challenges, we propose an attribute-driven cross-modal feature modulation framework, namely Language-guided Attribute alignment and Semantic Consistency (LASC), for domain adaptation in semantic segmentation. Specifically, our method tackles the rigidity of fixed prompts by introducing attribute-driven semantic modeling, and the instability of cross-modal alignment by enforcing semantic consistency. To further stabilize semantic representations under distribution shifts, we introduce a memory-based consistency constraint that enforces intra-class compactness and inter-class separability [9]. Together, these two components form a unified framework that enhances both the expressiveness and robustness of cross-modal representations across domains. The main contributions of this paper can be summarized as follows:

- To address the rigidity of fixed prompts, we introduce a mechanism that directly encodes domain-sensitive attributes and aligns them with visual features. This design provides fine-grained semantic cues that adapt to environmental variations, enabling more robust and transferable cross-domain representations.
- To mitigate the instability of cross-modal alignment, we propose a memory-based constraint that enforces intra-class compactness and inter-class separability. This regularization stabilizes semantic embeddings across domains and preserves clear decision boundaries under large distribution shifts.
- Our approach is instantiated on prompt-guided zero-shot adaptation, complemented by statistics-based modulation [10] to enhance the robustness of visual features. Extensive experiments on multiple cross-domain benchmarks demonstrate consistent improvements over state-of-the-art methods and confirm strong generalization ability without target-domain supervision.

II. Related Works

A. Domain Adaptation and Style Transfer

Domain adaptation focuses on alleviating distribution shifts between source and target domains. Early approaches explored feature transformation or subspace alignment, such as transfer component analysis [1], to map data from different domains into a shared space. Later methods emphasized learning transferable representations through tailored architectures and training strategies [11]. Adversarial learning has become a cornerstone of domain adaptation. Gradient reversal layers enable adversarial objectives to align features across domains [12], while contrastive learning provides more robust representations by optimizing pairwise similarity [13]. Category-aware or frequency-based transformations further enhance adaptation in complex scenarios such as semantic segmentation [14].

Style transfer, in parallel, aims to decouple content from style. Methods based on GANs have achieved strong results by transferring stylistic patterns while preserving content structures. Recent works have exploited semantic guidance from pretrained vision-language models like CLIP to steer generative models across domains [15]. In addition, data-mixing strategies have also been proposed to enhance robustness by exposing models to diverse distributions [4].

B. Leveraging CLIP for Zero-shot Segmentation

Recent advances in prompt learning have reshaped Zero-shot Segmentation. Works such as CoOp [6] and CoCoOp [16] design learnable or adaptive prompts to guide pretrained models, improving transferability across domains. Extensions like PØDA [10] and ATPrompt [7] introduce attribute-aware prompts, which better capture fine-grained semantics and enhance generalization to unseen classes [17]. These approaches highlight the importance of prompt design in vision-language generalization.

CLIP [5] has become a cornerstone in vision-language modeling, offering powerful visual-semantic alignment. Recent studies leverage its embeddings to refine visual features via semantic supervision, improving representation alignment across modalities [18], [19]. By utilizing the semantic structures encoded within CLIP, knowledge distillation frameworks can enhance model generalization [20], [21].

In generation tasks, methods such as StyleCLIP [22] and CLIPstyler [8] demonstrate that CLIP can provide fine-grained semantic guidance for style transfer and text-to-image synthesis. Spatially guided alignment techniques further extend this guidance to pixel-level control [23]. Unlike prior approaches, our method incorporate semantic consistency constraints into the alignment process, leading to more robust domain generalization.

III. Method

A. Problem Formulation

The objective of our task is to enable models to generalize across unseen target domains while preserving performance on the source domain, by exploiting both visual data from the source domain and attribute-driven semantic descriptions as auxiliary knowledge. Formally, let $x \in \mathcal{X}_{src}$ denote an image from source domain and $y \in \mathcal{Y}_{src}$ its corresponding label. The source-domain training set is defined as $\mathcal{D}_{src} = (x, y)$, which provides paired supervision. In addition, we introduce an attribute set $\mathcal{A} = \{a_1, a_2, \ldots, a_M\}$, where each attribute represents a domain-relevant semantic property (*e.g.*, lighting, weather, material). These attributes are combined with category labels to construct text descriptions $t^{(c,a)}$, which serve as semantic priors and domain knowledge. During training, the model has no access to images from unseen target domains $\mathcal{D}_{trg}$ but can rely on the attribute-augmented prompts as transferable knowledge. The objective is therefore to learn a joint vision–language representation that can achieve robust adaptation to the target domain.

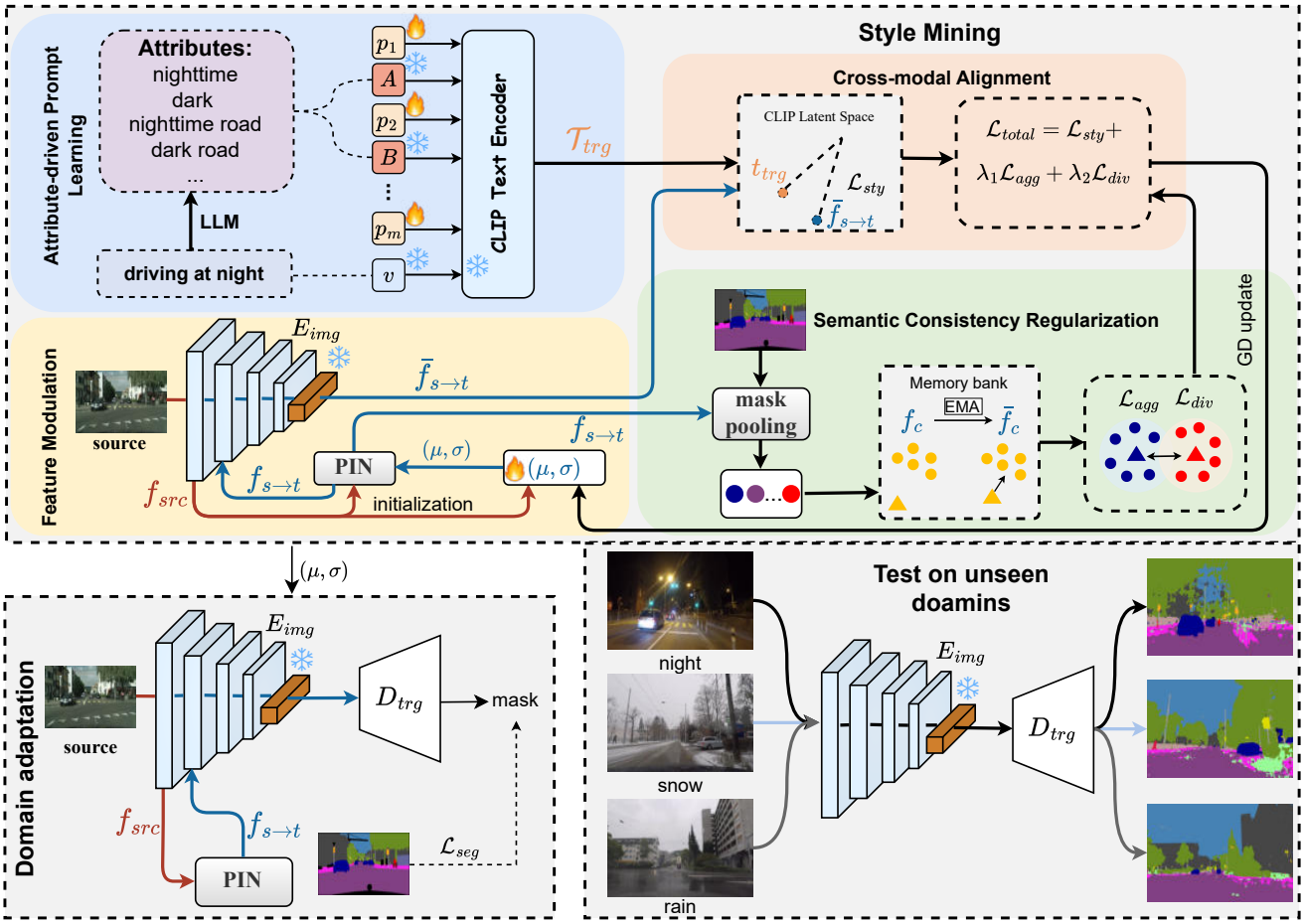

Fig. 2. Overview of LASC. Our framework consists of three stages. The first is **Style Mining**, including four modules: **Feature Modulation** for extracting normalized domain-robust features via the encoder and PIN; **Attribute-driven Prompt Learning**, which encodes target-domain conditions and attributes into CLIP embeddings; **Cross-modal Alignment**, aligning class-level visual features with text in a shared space; and **Semantic Consistency Regularization**, where a memory bank with EMA-updated prototypes enforces intra-class compactness and inter-class separation, jointly optimized with the alignment loss. The second stage is **Domain Adaptation**, which augments source features with learned statistics and fine-tunes the segmentation head using source labels. Finally, in **Evaluation**, the adapted model is directly applied to unseen domains without additional supervision.

B. Method Overview

The training part of LASC contains two main modules: Style Mining and Domain Adaptation, where Style Mining is responsible for mining the style for a specific domain, which is transferred to Domain Adaptation for learning the segmentation decoder. Concretely, the Style Mining module first uses a Feature Modulation unit to extract deep features of the source domain image through a visual encoder, and re-normalize its statistics using learnable modulation parameters to obtain a more stable feature distribution. Subsequently, by leveraging Semantic Consistency Regularization unit, it aggregates features within a class and maintain discriminability between classes, thereby mitigating the uncertainty caused by cross-domain distribution differences. Simultaneously, an Attribute-driven Prompt Learning unit is introduced to embed task-related attributes into the text representation for learning style information. Using the CLIP pre-trained model, it can achieve visual-semantic cross-modal alignment. Building on this foundation, it further introduces a Cross-modal Alignment unit to enhance the interaction between visual features and attribute semantics, thereby achieving more discriminative semantic representations. The learned style information is transferred to the Domain Adaptation module to further optimize the decoder of segmentation network, which is finally used as the segmenter for target domain in the last evaluation module. The overall framework is shown in Fig. 2.

C. Style Mining

1) Feature Modulation: To incorporate channel-wise feature modulation based on learnable statistics, we first introduce the Prompt-driven Instance Normalization (PIN) structure [10]. Given an input image x from the source domain, the visual encoder E_I produces a feature map from the first layer,

$$\mathbf{f}_{src} = \mathrm{E}_{\ell 1}(x) \in \mathbb{R}^{C \times H \times W}, \tag{1}$$

where C, W, H are the dimensions of channel, width and height respectively of the feature. For each channel $i \in [1, C]$, we compute the first- and second-order statistics,

$$\mu_{src}(i) = \frac{1}{HW} \sum_{h,w} \mathbf{f}_{src}(i, w, h), \tag{2}$$

$$\sigma_{src}(i) = \sqrt{\frac{1}{HW}\sum_{h,w}(\mathbf{f}_{src}(i,w,h) - \mu_{src}(i))^2}. \quad (3)$$

Instead of directly relying on these domain-specific statistics, we introduce learnable parameters (μ_t, σ_t) that re-normalize the features into a more robust distribution:

$$\mathbf{f}_{s\to t}(i) = \sigma_t(i)\frac{\mathbf{f}_{src}(i) - \mu_{src}(i)}{\sigma_{src}(i)} + \mu_t(i). \quad (4)$$

This $\mathbf{f}_{s\to t}$ is subsequently used to replace the original second layer feature. This modulation scheme serves two purposes simultaneously: (1) it removes the bias of raw domain-dependent statistics, and (2) it allows the model to adaptively learn new feature distributions that are more resilient to unseen domains. As a result, the modulated features $\mathbf{f}_{s\to t}$ maintain stability across varying visual conditions, providing a stronger foundation for subsequent semantic alignment.

2) Attribute-driven Prompt Learning (APL): Relying solely on the category label y to generate textual prompts is insufficient to capture cross-domain variations. For instance, the visual appearance of a "car" differs significantly between sunny and rainy conditions, and a single category label cannot account for such differences. To overcome this limitation, we introduce an attribute set $\mathcal{A}$ that supplements contextual and environmental information relevant to the target domain.

Specifically, we construct task-related natural language descriptions by combining learnable tokens, domain-sensitive attributes, and target-domain descriptors as follows:

$$\mathbf{P}_T = [\mathbf{p}_1, \mathbf{a}_1, \mathbf{p}_2, \mathbf{a}_2, ..., \mathbf{p}_m, \mathbf{v}], \quad (5)$$

where $\mathbf{p}_i$ denotes learnable tokens, $\mathbf{a}_i \in \mathcal{A}$ represent selected attributes, and v encodes a textual description of the target domain. This construction enables prompts to flexibly adapt to different environmental conditions, thereby capturing fine-grained semantic variations and improving cross-domain generalization.

For instance, prompts such as "driving at night with low visibility and lighting effects" or "driving in snow with slippery and snow-covered roads" can be generated. These textual descriptions are then encoded into semantic representations via the text encoder:

$$\mathbf{t}_{\text{sem}} = \mathrm{E}_T(\mathbf{P}_T). \quad (6)$$

where CLIP is adopted as E_T in this paper. To avoid treating all attributes equally, we introduce learnable parameters $\{\alpha_1, \cdots, \alpha_i, \cdots, \alpha_M\}$ to fuse semantic embedding,

$$\mathbf{t}_{trg} = \sum_{j=1}^{M} w_j \, \mathrm{E}_T\left(\mathrm{G}(a_j)\right), \quad (7)$$

$$w_j = \frac{\exp(\alpha_j)}{\sum_{k=1}^{M}\exp(\alpha_k)}. \quad (8)$$

where $\mathrm{G}(a_j)$ is "driving under $< style >$ with a_j", and "$< style >$" specifies the overall target domain condition. Eq. 8 is used to make the summation of all w_j to 1.

This design offers two benefits. First, different attribute combinations provide diverse contextual descriptions, enabling the model to observe a richer distribution of cross-domain semantics during training. Second, the adaptive weighting mechanism allows the model to automatically focus on the most relevant attributes under different conditions (*e.g.*, emphasizing "lighting" in low-light scenarios or highlighting "rain/snow" in adverse weather), thereby enhancing the robustness and generalization capability of the learned representations.

3) Cross-modal Alignment (CMA): After the processing of above two units, visual and textual features gain stronger cross-domain adaptability and semantic expressiveness. However, without explicit constraints in the shared space, they may still drift apart, resulting in unstable correspondence between images and text. To address this issue, we design a cross-modal alignment mechanism that ensures consistency by optimizing a CLIP-style contrastive loss.

Concretely, the visual branch produces a global representation $\bar{\mathbf{f}}_{s\to t} = \mathrm{E}_{\ell 4}(x)$ from the last layer of the image encoder and the PIN module, which alleviates low-level statistical bias from the source domain. On the textual side, attribute-driven natural language prompts such as "driving at night with low visibility and lighting effects" are generated and encoded by the text encoder into a fused representation $\mathbf{t}_{\text{trg}}$. We then align the two modalities with a normalized cosine similarity loss:

$$\mathcal{L}_{\text{sty}} = 1 - \cos(\bar{\mathbf{f}}_{s\to t}, \mathbf{t}_{trg}). \quad (9)$$

This contrastive loss reduces the distance between semantically matched image-text pairs while enlarging the gap for mismatched combinations, thereby maintaining discriminability in the shared semantic space. Unlike attention-based or more complex interaction mechanisms, our design remains simple and efficient, while still leveraging attribute-driven prompts to provide explicit semantic guidance. As a result, visual features are directly anchored to the textual semantics that describe target-domain conditions, which stabilizes cross-modal consistency under domain shifts and significantly enhances the zero-shot generalization capability.

4) Semantic Consistency Regularization (SCR): Based on cross-modal alignment, the model is able to establish correspondences between images and text in a shared semantic space. However, the semantic boundaries between different categories can still become blurred due to domain shift, leading to unstable feature aggregation. To address this, we further introduce semantic consistency constraints, which aggregate intra-class features and disperse inter-class features to ensure discriminative power. Specifically, we first maintain a feature memory bank $\mathcal{M}$ during training, which stores the historical embedding mean and variance of each category. For the current input semantic feature $\mathbf{f}_c$, we calculate its consistency constraint with the corresponding category prototype $\bar{\mathbf{f}}_c$:

$$\mathcal{L}_{\text{agg}} = \left\|\mathbf{f}_c - \bar{\mathbf{f}}_c\right\|^2, \quad (10)$$

where $\bar{\mathbf{f}}_c$ denotes the prototype feature of class c , which is obtained by aggregating the features of all samples belonging to class c. This loss achieves intra-class convergence by minimizing the Euclidean distance between features and class centers, thereby improving feature aggregation. At the same time, to prevent features of different classes from being too close in the shared space, we design an inter-class separation constraint:

$$\mathcal{L}_{\text{div}} = \sum_{k \neq c} \max(0, \tau - \cos(\mathbf{f}_c, \bar{\mathbf{f}}_k)), \tag{11}$$

τ is the similarity threshold and $\cos(\cdot, \cdot)$ is the cosine similarity. This loss requires the similarity between the current feature and the non-target category prototype to be less than τ, thereby maintaining a clear distinction boundary in the shared semantic space. Finally, we jointly optimize the cross-modal alignment loss and the consistency regularizer:

$$\mathcal{L} = \mathcal{L}_{\text{sty}} + \lambda_1 \mathcal{L}_{\text{agg}} + \lambda_2 \mathcal{L}_{\text{div}}, \tag{12}$$

where λ_1 and λ_2 are coefficients to balance the two losses and are set to 0.5 and 0.25. By jointly optimizing intra-class convergence and inter-class separation, the model can further enhance discriminability while maintaining cross-modal consistency. This ensures that features maintain robust semantic boundaries across different domains, significantly improving zero-shot generalization performance.

D. Domain Adaptation

After the style mining stage, we further perform adaptation on the target domain to enhance the model's robustness under distribution shifts. Different from the previous stage, the statistics (μ_t, σ_t) learned in the PIN module are fixed and no longer updated. The adaptation process relies solely on source-domain supervision, while the modulation with target-domain style parameters ensures that source samples are transformed to mimic the distribution of the target domain during training.

Concretely, given a source image x with pixel-level annotations y_s, we first extract their features through visual encoder and modulate them with randomly sampled target-domain statistics obtained in the feature augmentation stage. The resulting representations are then fed into the segmentation head D_{trg}, and the predictions are optimized with a standard cross-entropy loss against the source ground truth:

$$\mathcal{L}_{\text{seg}} = \mathrm{CE}(\hat{y}_s = \mathrm{D}_{trg}(\mathrm{E}_{img}^{PIN}(x)), y_s), \tag{13}$$

where E_{img}^{PIN} means the image encoder with PIN modulation, which is shown in the bottom-left of Fig. 2. During backpropagation, only the segmentation head and related learnable parameters are updated, while the PIN statistics remain fixed. By continuously perturbing source features with different target style parameters and training under supervision, the model gradually acquires robustness to target-domain.

Through this fine-tuning procedure, we can obtain the final adaptation model $\mathrm{F}(\cdot) = \mathrm{D}_{trg}(\mathrm{E}_{img}(\cdot))$, where E_{img} means the image encoder without PIN modulation, which effectively bridges the gap between the source and target domains, enabling robust semantic segmentation on unseen target images.

IV. Experiments

A. Datasets and Implementation details

1) Datasets: For segmentation, we use Cityscapes [24] (2975 training and 500 validation images) and GTA5 [25] (24,966 images, with 10% for validation) as source domains, and ACDC [26] as the target domain, which contains diverse weather conditions (night, snow, rain) with about 100 test images per subset. All datasets share 19 semantic categories. For classification, we adopt CUB [27] and CUB-Painting [28], both covering 200 bird species but from different domains, i.e., natural photographs and stylized paintings. For object detection, we employ the Diverse Weather Dataset (DWD) [29], which includes five weather conditions: Day Clear, Night Clear, Dusk Rainy, Night Rainy, and Day Foggy. The detector is trained on the Day Clear split (19,395 images) with 8,313 images for validation, while Night Clear, Dusk Rainy, and Day Foggy serve as target domains.

2) Implementation Details: The training process of our LASC consists of the three steps: 1) Supervised training on source domain, where the encoder and PIN extract normalized features, attribute-driven prompts provide textual embeddings, and cross-modal alignment is optimized with semantic consistency; 2) Source features are augmented with learned statistics, and the segmentation head is fine-tuned under source-label supervision to reduce the domain gap; 3) The adapted model is directly applied to target-domain for zero-shot testing without any additional supervision.

We adopt the DeepLabv3+ architecture, where the feature extraction backbone is initialized from the image encoder E_{img} of the pre-trained CLIP-ResNet-50 model. To construct textual representations, we employ prompts of the form "driving under $< style >$ with [attributes]", where $< style >$ denotes the target domain condition (e.g., night, snow, rain, or game), and attributes are domain-related factors such as visibility, road condition, or lighting effects. The resulting domain-aware prompts are encoded by the CLIP text encoder to produce semantic embeddings that guide cross-modal alignment.

B. Results for Segmentation

This experiment focuses on examining domain extension in three scenarios: Rain, Snow and Night in ACDC. The source domain is Cityscapes or GTA5, we also perform evaluations from real to synthetic datasets (CS $\rightarrow$ GTA5) and from synthetic to real (GTA5 $\rightarrow$ CS).

1) Baselines: In this experiment, we mainly compare with the following methods:

- **Source-only** model consists of a ResNet-50 from CLIP as the backbone and employs DeepLabV3+ for segmentation. It is trained on the source domain using the same settings as described above.
- **CLIPstyler** [8] is a style conversion model, which realizes pixel level style conversion by providing semantic style information through pre-trained CLIP and U-Net.

TABLE I

DOMAIN ADAPTATION IN SEMANTIC SEGMENTATION. WE REPORT THE MIOU(%) ON TARGET DOMAINS (ACDC, GTA5, CITYSCAPES). THE BEST RESULTS ARE SHOWN IN BOLD.

Src	Trg	Method	Trg mIoU(%)	Src	Trg	Method	Trg mIoU(%)
CS	Night	source-only	18.31	GTA5	Night	source-only	12.22
		CLIPstyler	21.38±0.16			CLIPstyler	12.67±0.15
		PØDA	25.28±0.45			PØDA	15.52±0.37
		LASC	**27.04**±0.47			**LASC**	**16.81**±0.33
	Snow	source-only	39.82		Snow	source-only	32.32
		CLIPstyler	41.11±0.09			CLIPstyler	31.58±0.24
		PØDA	43.87±0.62			PØDA	34.04±0.22
		LASC	**45.88**±0.56			**LASC**	**35.89**±0.44
	Rain	source-only	38.20		Rain	source-only	33.32
		CLIPstyler	37.15±0.10			CLIPstyler	31.65±0.19
		PØDA	42.36±0.55			PØDA	35.18±0.35
		LASC	**44.67**±0.42			**LASC**	**37.06**±0.42
	GTA5	source-only	39.59		CS	source-only	36.38
		CLIPstyler	38.72±0.20			CLIPstyler	31.66±0.19
		PØDA	41.12±0.47			PØDA	41.56±0.33
		LASC	**44.53**±0.52			**LASC**	**41.82**±0.38

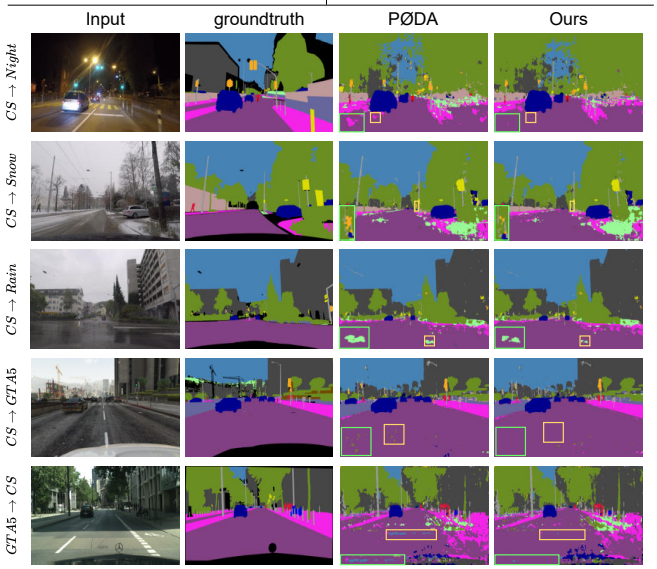

Fig. 3. Segmentation results on cross-domain settings. The first column shows the input images and the second column presents the corresponding ground-truth annotations. The third column illustrates the segmentation results of PØDA, while the fourth column reports the results of our method. Compared with PØDA, our approach yields more accurate segmentation under diverse domain shifts. The content in the green box is an enlarged display of the content in the yellow box.

- **PØDA** [10] is an advanced model specially designed for zero shot domain adaptation. It uses the semantic information of the target domain to adjust the style vector of the source domain features, and successfully adapts to the target domain by fine-tuning.

All the above baselines use the same image encoder and use images of the same size as our method.

2) Comparison with baselines: According to the experimental results of CS → ACDC shown in the left part of Table I, this method shows significant performance advantages that LASC can outperform PØDA by average 2.4% on all four settings. In addition, to further demonstrate the performance of our LASC, we also show some qualitative results in Fig. 3, from which we can clearly observe that our method can get better result in some complex areas.

Similarly, the domain adaptation experiment from GTA5 to ACDC and CS also shows significant performance improvement according to the right part of Table I.

TABLE II

THE IMPORTANCE OF APL AND SCR FOR CS TO ACDC AND GTA5.

attributes	SCR	Night	Snow	Rain	GTA5
✗	✗	25.03±0.48	43.90±0.53	42.31±0.55	41.07±0.48
✓	✗	26.05±0.52	44.48±0.55	43.34±0.44	42.74±0.49
✓	✓	27.04±0.47	45.88±0.56	44.67±0.42	44.53±0.52
✗	✓	26.37±0.43	44.72±0.51	43.46±0.47	43.12±0.44

C. Importance of Attributes and SCR

To evaluate the effectiveness of attribute-driven prompts and class-wise consistency constraints, we conduct ablation studies by progressively adding these components. We begin with only target domain description, then incorporate attribute-driven prompt learning, and finally add the intra-class aggregation and inter-class separation losses.

As shown in Table II, domain-level descriptions alone capture global target domain characteristics but remain limited in modeling fine-grained semantics. Incorporating attribute prompts significantly improves mIoU by providing richer contextual cues such as visibility or road conditions, thereby alleviating semantic ambiguity. Adding semantic consistency regularization further enhances performance, showing that enforcing compactness within classes and separation across classes stabilizes cross-domain alignment and strengthens overall discriminability.

TABLE III

IMPACT OF SELECTED LAYERS FOR AUGMENTATION ON CS → NIGHT.

Layer1	Layer2	Layer3	Layer4	ACDC Night
✓	✗	✗	✗	27.04±0.47
✓	✓	✗	✗	23.96±0.44
✓	✗	✓	✗	23.28±0.51
✓	✗	✗	✓	21.86±0.39

D. Choice of Features to Augment

DeepLabV3+ utilizes both low-level features from Layer 1 and high-level features from Layer 4 as inputs to the segment decoder. In LASC, we augment only the Layer 1 features and propagate them through Layers 2–4 to obtain the Layer 4 representations. As shown in Table III, we further investigate whether augmenting additional layers is beneficial; the results indicate that augmenting only Layer 1 yields the best performance.

TABLE IV

IMPACT OF FROZEN LAYERS WHEN FINE-TUNING ON CS → NIGHT.

Layer1	Layer2	Layer3	Layer4	ACDC Night
✓	✗	✗	✗	20.03±0.54
✓	✓	✗	✗	25.41±0.46
✓	✓	✓	✗	25.74±0.41
✓	✓	✓	✓	27.04±0.47

E. Frozen Layer Selection for Backbone

The study of PØDA validates that employing a frozen backbone during the training phase enhances the model's ability to generalize to unseen domains. During fine-tuning, however, freezing the backbone is optional, with the exception of Layer1. For further clarification, Table IV reports the results when freezing the backbone at different layers.

The experimental results indicate that freezing the ResNet backbone remains the optimal strategy. Relying solely on single-domain training data is insufficient to enhance the

TABLE V

IMPACT OF SOURCE-ONLY PRE-TRAINING ON CS TO ACDC/GTA5.

Method	Night	Snow	Rain	GTA5
LASC no src pretrain	23.17	37.06	40.73	40.65
LASC	**27.04**±0.47	**45.88**±0.56	**44.67**±0.42	**44.53**±0.52

backbone's discriminative ability on unseen domains and may even degrade the inherent adaptation capacity of the CLIP pretrained backbone. This finding is consistent with numerous zero-shot learning studies, which suggest that directly utilizing the weights of a pretrained network is often an effective approach for achieving cross-domain adaptation when data are limited.

F. Importance of source-only pre-training

To verify the role of source-only pre-training, we conducted an ablation study on the adaptation tasks from Cityscapes to ACDC (Night/Snow/Rain) as well as to GTA5. As shown in Table V, we observe that directly training LASC from scratch on the augmented features (LASC no src pretrain) leads to a significant drop in performance, whereas performing source-only pre-training before adaptation (LASC) substantially improves the overall mIoU. These results demonstrate that source-only pre-training provides a stable and effective initialization and is therefore a critical step in the LASC adaptation process.

TABLE VI

IMPACT OF DIFFERENT BACKBONES ON SEMANTIC SEGMENTATION.

Backbone	Method	Night	Snow	Rain	GTA5
Resnet-50	source-only	18.31	39.28	38.20	39.59
	PØDA	25.28±0.45	43.87±0.62	42.36±0.55	41.12±0.47
	LASC	27.04±0.47	45.88±0.56	44.67±0.42	44.53±0.52
Resnet-101	source-only	22.17	44.53	42.53	44.06
	PØDA	26.54±0.12	46.71±0.43	46.36±0.20	44.17±0.24
	LASC	**28.16**±0.24	**48.54**±0.41	**48.13**±0.28	**47.46**±0.32

G. Other backbone networks

In Table VI, we present the experimental resultes of LASC using different backbones (ResNet-50 and ResNet-101), along with the results of PØDA for comparison. The results show that changing the backbones does not compromise the domain extension capability of LASC. Compared with training on source-only data, LASC consistently achieves significant performance gains on the target domain while mainting superiority over PØDA. These findings further validate the robustness of LASC and highlight its potential for extension to more powerful backbone networks.

H. Effect of the number of optimization iterations

To determine the optimal number of optimization iterations, we conducted an evaluation on the validation set. As illustrated in Fig. 4, the results show that the performance steadily improves with more iterations and reaches its peak around 80–100 iterations. However, when the number of iterations continues to increase, the performance starts to decline, indicating an over-stylization effect that undermines the model's adaptation capability. Based on these validation results, we fix the number of iterations to 100 in the subsequent experiments, which achieves a favorable trade-off between accuracy and stability while avoiding overfitting.

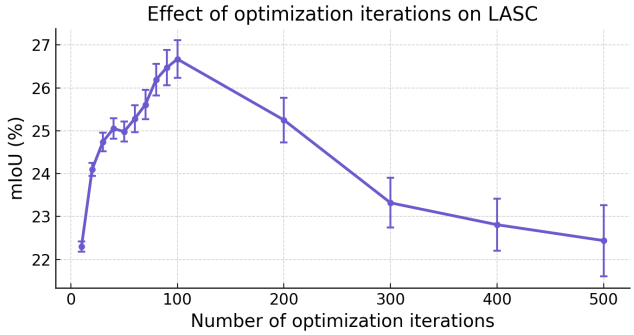

Fig. 4. Effect of the number of optimization iterations. The mIoU performance of LASC in the adaptation from Cityscapes to ACDC-Night with respect to the number of statistics optimization iterations. Each value represents the mean over five independent runs, and the error bars denote the standard deviation.

TABLE VII

PERFORMANCE OF LASC FOR CLASSIFICATION (ACC%)

Dataset	CLIP ZS	CLIP LP	CLIP LP(ZS init)	LADS	**LASC**
CUB → CUB-Painting	52.84	64.33±0.29	65.05±0.05	66.18±0.25	**66.59**±0.31

I. LASC for Other Tasks

In the semantic segmentation task, our model has already demonstrated outstanding performance. To further validate its superiority and adaptation capability, we conducted extensive experimental evaluations on additional tasks, *i.e.*, Classification and Object Detection.

1) Results for Classification: This section presents the experimental validation of LASC on the image classification task. We primarily conducted test on CUB → CUB-Painting.

In the task of image classification, LASC mainly compares with the following baseline methods:

- **CLIP ZS** [5] uses the category name as a hint of the pre-training CLIP model.
- **CLIP LP** [5] is a variant of the CLIP model, and fine tune the image embedding through a linear probe.
- **CLIP LP (ZS-init)** [5] is a CLIP model that uses text embedding to initialize linear probes.
- **LADS** [18] uses domain specific semantic information to guide image embedding to adjust to the target domain. This conditioning is achieved by training a standard two-tier MLP. In the fine-tuning phase, LADS introduces a linear probe to consider the deviation.

All the above methods use the same image encoder, VIT-L/14 [30], and the image size is adjusted to 224 × 224. The experimental results are shown in the Table VII, where the baseline results are cited from their original papers.

As shown in Table VII, our model significantly outperforms all compared methods except LADS. This demonstrates that our model not only excels in semantic segmentation but also exhibits strong performance in classification.

2) Results for Object Detection: This section presents the experimental validation of LASC on the object detection task. The selected dataset shares similar characteristics with the one used for semantic segmentation, containing images captured under various real-world weather conditions. In our experiments, the testing was conducted under multiple scenarios, using daytime clear weather as the source domain:Day Clear→Night Clear,Day Clear→Dusk Rainy,Day

TABLE VIII
PERFORMANCE OF LASC FOR OBJECT DETECTION (MAP%)

Method	Night Clear	Dusk Rainy	Day Foggy
S-DGOD	36.6	28.2	33.5
CLIP The Gap	36.9	32.3	38.5
OA-DG	38.0	33.9	38.3
PØDA	43.4	40.2	44.4
LASC	**44.2**±0.37	**41.7**±0.35	**45.6**±0.41

Clear→Day Foggy. In this task, LASC mainly compares with the following baseline methods besides **PØDA**:

- **CLIP The Gap** [31] is a method specifically designed for SDG, which leverages textual guidance to enhance the diversity of source data.
- **OA-DG** [32] is another approach to SDG, which directly extracts visual features by perceiving objects in image, aiming to capture domain-invariant information.
- **S-DGOD** [33] is an SGD method that achieves domain alignment for object detection by grouping proposals based on visual similarity, aiming to coarsely align instance features across domains.

As shown in Table VIII, when using Day Clear as the source domain, LASC demonstrates clear advantages in cross-domain object detection. On the target domains Night Clear, Dusk Rainy, and Day Foggy, LASC achieves mAP scores of 44.2%, 41.7%, and 45.6%, respectively, all surpassing the strongest baseline PØDA (43.4%, 40.2%, and 44.4%). Compared with other methods, LASC consistently maintains superior performance under significant appearance shifts and environmental changes, highlighting its effectiveness and robustness in capturing domain-invariant features and enhancing cross-domain adaptation.

V. CONCLUSION

We presented LASC, a novel framework for cross-domain semantic segmentation that integrates feature modulation with learnable statistics, attribute-driven prompt learning, and cross-modal alignment with semantic consistency regularization. By further fine-tuning with source supervision, our method achieves strong adaptation to unseen domains without requiring additional target labels. Extensive experiments confirm its effectiveness under diverse conditions, and future work will extend this approach to broader multi-modal scenarios and tasks such as video understanding and cross-modal retrieval.

REFERENCES

[1] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE TNN*, vol. 22, no. 2, pp. 199–210, 2010.

[2] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *ECCV Workshops*, 2016.

[3] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *ICML*, 2015, pp. 1180–1189.

[4] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with mixstyle," in *ICLR*, 2021, pp. 1–14.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[6] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," in *CVPR*, 2022, pp. 7266–7276.

[7] Z. Li, Y. Song, M.-M. Cheng, X. Li, and J. Yang, "Advancing textual prompt learning with anchored attributes," in *ICCV*, 2025.

[8] G. Kwon and J. C. Ye, "Clipstyler: Image style transfer with a single text condition," in *CVPR*, 2022, pp. 18 062–18 071.

[9] H. Niu, L. Xie, J. Lin, and S. Zhang, "Exploring semantic consistency and style diversity for domain generalized semantic segmentation," in *AAAI*, vol. 39, no. 6, 2025, pp. 6245–6253.

[10] M. Fahes, T.-H. Vu, A. Bursuc, P. Pérez, and R. De Charette, "Poda: Prompt-driven zero-shot domain adaptation," in *ICCV*, 2023, pp. 18 623–18 633.

[11] O. Sener, H. O. Song, A. Saxena, and S. Savarese, "Learning transferrable representations for unsupervised domain adaptation," in *NeurIPS*, 2016, pp. 2118–2126.

[12] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial training of neural networks," *JMLR*, vol. 17, no. 59, pp. 1–35, 2016.

[13] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *CVPR*, 2019, pp. 4893–4902.

[14] V. Kumar, H. Patil, R. Lal, and A. Chakraborty, "Improving domain adaptation through class aware frequency transformation," *IJCV*, vol. 131, no. 11, pp. 2888–2907, 2023.

[15] R. Gal, O. Patashnik, H. Maron, A. H. Bermano, G. Chechik, and D. Cohen-Or, "Stylegan-nada: Clip-guided domain adaptation of image generators," *ACM ToG*, vol. 41, no. 4, pp. 1–13, 2022.

[16] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *CVPR*, 2022, pp. 16 816–16 825.

[17] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv:2104.08691*, 2021.

[18] L. Dunlap, C. Mohri, D. Guillory, H. Zhang, T. Darrell, J. E. Gonzalez, A. Raghunathan, and A. Rohrbach, "Using language to extend to unseen domains," in *ICLR*, 2023, pp. 1–14.

[19] M. Singha, H. Pal, A. Jha, and B. Banerjee, "Ad-clip: Adapting domains in prompt space using clip," in *ICCV*, 2023, pp. 4355–4364.

[20] T. Kalluri, B. P. Majumder, and M. Chandraker, "Tell, don't show!: Language guidance eases transfer across domains in images and videos," *arXiv preprint arXiv:2403.05535*, 2024.

[21] C. Jiang, J. Zhao, J. Deng, Z. Li, and H. Zhang, "Imbuing, enrichment and calibration: Leveraging language for unseen domain extension," *IJCV*, vol. 133, p. 4064–4090, 2025.

[22] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "Styleclip: Text-driven manipulation of stylegan imagery," in *ICCV*, 2021, pp. 2085–2094.

[23] S. Petryk, L. Dunlap, K. Nasseri, J. Gonzalez, T. Darrell, and A. Rohrbach, "On guiding visual attention with language specification," in *CVPR*, 2022, pp. 18 092–18 102.

[24] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016, pp. 3213–3223.

[25] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *ECCV*, 2016, pp. 102–118.

[26] C. Sakaridis, D. Dai, and L. Van Gool, "Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *ICCV*, 2021, pp. 10 765–10 775.

[27] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," CalTech, Tech. Rep., 2011.

[28] S. Wang, X. Chen, Y. Wang, M. Long, and J. Wang, "Progressive adversarial networks for fine-grained domain adaptation," in *CVPR*, 2020, pp. 9213–9222.

[29] A. Wu and C. Deng, "Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation," in *CVPR*, 2022, pp. 847–856.

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021, pp. 1–14.

[31] V. Vidit, M. Engilberge, and M. Salzmann, "Clip the gap: A single domain generalization approach for object detection," in *CVPR*, 2023, pp. 3219–3229.

[32] W. Lee, D. Hong, H. Lim, and H. Myung, "Object-aware domain generalization for object detection," in *AAAI*, 2024, pp. 2947–2955.

[33] F. Rezaeianaran, R. Shetty, R. Aljundi, D. O. Reino, S. Zhang, and B. Schiele, "Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection," in *ICCV*, 2021, pp. 9204–9213.