SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2026, Vol. 69, Iss. 1, 182208:1–182208:20
https://doi.org/10.1007/s11432-025-5039-y

Structured prototype regularization for synthetic-to-real driving scene parsing

Jiahe FAN[1], Xiao MA[2], Sergey VITYAZEV[3], George GIAKOS[4], Shaolong SHU[1] & Rui FAN[1*]

[1]*College of Electronic and Information Engineering, Tongji University, Shanghai 201804, China*
[2]*Beijing Institute of Aerospace Control Devices, China Aerospace Science and Technology Corporation, Beijing 100039, China*
[3]*Faculty of Radio Engineering and Telecommunications, Ryazan State Radio Engineering University, Ryazan 390005, Russia*
[4]*Department of Electrical and Computer Engineering, Manhattan University, New York 10471, USA*

Received 1 July 2025/Revised 4 December 2025/Accepted 6 February 2026

Abstract Driving scene parsing is critical for autonomous vehicles to operate reliably in complex real-world traffic environments. To reduce the reliance on costly pixel-level annotations, synthetic datasets with automatically generated labels have become a popular alternative. However, models trained on synthetic data often perform poorly when applied to real-world scenes due to the synthetic-to-real domain gap. Despite the success of unsupervised domain adaptation in narrowing this gap, most existing methods mainly focus on global feature alignment while overlooking the semantic structure of the feature space. As a result, semantic relations among classes are insufficiently modeled, limiting the model's ability to generalize. To address these challenges, this study introduces a novel unsupervised domain adaptation framework that explicitly regularizes semantic feature structures to significantly enhance driving scene parsing performance in real-world scenarios. Specifically, the proposed method enforces inter-class separability and intra-class compactness by leveraging class-specific prototypes, thereby enhancing the discriminability and structural coherence of feature clusters. An entropy-based noise filtering strategy improves the reliability of pseudo labels, while a pixel-level attention mechanism further refines feature alignment. Extensive experiments on representative benchmarks demonstrate that the proposed method consistently outperforms recent state-of-the-art methods. These results underscore the importance of preserving semantic structure for robust synthetic-to-real adaptation in driving scene parsing tasks.

Keywords driving scene parsing, autonomous vehicles, unsupervised domain adaptation, synthetic-to-real adaptation, class prototypes, contrastive learning

Citation Fan J H, Ma X, Vityazev S, et al. Structured prototype regularization for synthetic-to-real driving scene parsing. Sci China Inf Sci, 2026, 69(1): 182208, https://doi.org/10.1007/s11432-025-5039-y

1 Introduction

1.1 Background

Driving scene parsing analyzes traffic environments to identify and categorize key elements, such as roads, vehicles, traffic lights, and sidewalks [1–4]. It is a key component of autonomous driving perception systems and provides a comprehensive machine-level understanding of traffic environments, supporting safe navigation and decision-making [5, 6]. Recent advances in fully supervised methods, primarily based on convolutional neural networks (CNNs), have achieved notable success by leveraging large-scale, manually annotated datasets [7–9]. Nevertheless, generating such pixel-level annotations remains prohibitively costly. For instance, labeling a single image in the Cityscapes dataset [10] requires approximately 1.5 h under normal conditions, and up to 3.3 h under adverse weather conditions [11].

An alternative to manual annotation is the use of synthetic datasets with automatic pixel-level labels for CNN training [10, 12]. However, models trained on synthetic data often suffer substantial performance drops when applied to real-world scenes due to the domain gap between the synthetic (source) and real (target) domains [13], as illustrated in Figure 1. This discrepancy arises from differences in image characteristics such as lighting, texture, and contrast, all of which hinder the generalization of CNNs to real-world driving environments. To address this issue, unsupervised domain adaptation (UDA) methods have been developed for driving scene parsing [14–20]. UDA methods transfer knowledge from labeled synthetic datasets to unlabeled real-world images by aligning the feature distributions across domains. By narrowing the gap between source and target representations, UDA provides a

* Corresponding author (email: rui.fan@ieee.org)

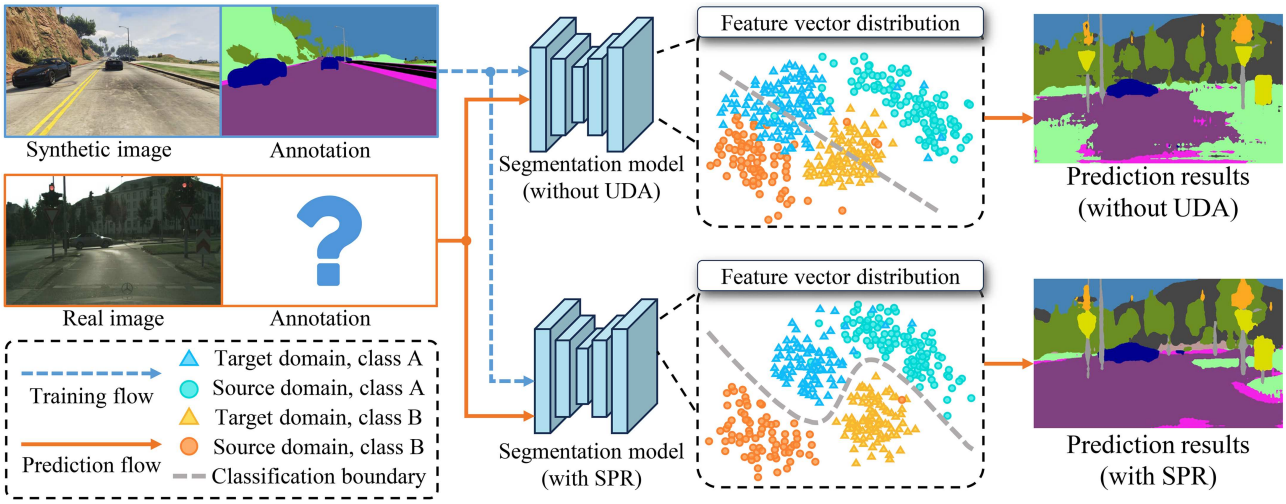

Figure 1 Image domain discrepancy and the effectiveness of the UDA method in driving scene parsing. The proposed SPR enhances feature alignment by explicitly modeling and preserving inter-class separability and intra-class compactness, leading to more structured and discriminative feature representations across domains.

practical and scalable solution for enhancing model performance in real-world autonomous driving applications. This study aims to bridge the domain gap between synthetic and real-world data, thereby improving the performance of driving scene parsing models in real-world environments.

1.2 Existing challenges and motivation

Despite significant progress in UDA for driving scene parsing, several key challenges remain. A primary issue is that most existing methods rely on global feature alignment strategies, such as adversarial learning, which often overlook the semantic structure of the feature space [13]. As a result, relationships among semantic classes, such as roads, vehicles, and pedestrians, are not explicitly preserved, which limits the effectiveness of feature alignment across domains.

Recent advances in prototype-based contrastive learning methods [21] have improved class-wise feature alignment by associating features with semantic prototypes. Nonetheless, these methods typically treat prototypes as isolated points, neglecting the structural relationships within the feature space [21]. Effective driving scene parsing requires not only accurate alignment of target features to source-domain prototypes, but also the preservation of a well-structured feature space characterized by high inter-class separability[1)] and strong intra-class compactness[2)] [22]. These two properties jointly contribute to more reliable decision boundaries and improved generalization across domains. However, most existing prototype-based contrastive learning methods do not explicitly enforce or model such structural constraints, thereby limiting their effectiveness in real-world domain adaptation tasks.

Moreover, existing strategies [23,24] often prioritize global feature alignment while overlooking the distribution of pixel-level features, which is essential for accurate driving scene parsing. This limitation becomes particularly pronounced under significant domain shifts, where pixel-wise predictions in the target domain are prone to uncertainty and noise [22]. In the absence of effective mechanisms to account for pixel-level discrepancies, alignment may be distorted by unreliable predictions, ultimately degrading the performance of scene parsing models. Addressing this issue requires more granular strategies capable of suppressing noisy pixel features while enhancing their consistency with the class-level semantic structure.

1.3 Novel contributions

To address the aforementioned limitations, this study presents a novel UDA framework based on structured prototype regularization (SPR), which explicitly enforces inter-class and intra-class consistency within the feature space. Given the significant discrepancy in feature distributions between the source and target domains, driving scene parsing on target-domain images often contains substantial noise. To mitigate the impact of this noise on feature

1) Inter-class separability ensures that features from different semantic categories, such as roads, vehicles, and pedestrians, are projected into clearly separated regions in the feature space, thereby minimizing class confusion.

2) Intra-class compactness requires that features from the same category remain tightly clustered, even across variations in appearance caused by lighting, occlusion, or domain shift.

alignment, the proposed method first computes class-specific prototypes from CNN logits, which serve as approximate centroids for each semantic category. These prototypes effectively capture the core characteristics of each class and suppress the influence of noisy features, thereby enhancing representation robustness and discriminability. To regularize the semantic structure of the feature space, the proposed method quantifies the relationships among prototypes by computing both inter-class and intra-class correlation matrices. These matrices guide the formulation of regularization terms that promote a semantically well-structured feature space, characterized by clear inter-class separability and strong intra-class compactness. Building on these refined prototypes, a contrastive loss is introduced to align feature distributions across domains. Since driving scene parsing fundamentally requires a dense, pixel-level understanding of the environment, aligning pixel-level feature distributions is critical for achieving fine-grained parsing performance. To this end, the correlation matrix between each pixel-level feature and each class prototype is further computed and utilized to guide the optimization process. Specifically, this study introduces an entropy-based measure derived from the correlation matrix to filter noisy pixel-level predictions in the target domain, thereby improving the prototype estimation accuracy. Furthermore, a pixel-level attention mechanism is incorporated into the contrastive loss to emphasize more reliable features during alignment, further enhancing the consistency of pixel-level representations across domains.

The novel contributions of this study can be summarized as follows.

- A novel contrastive learning framework based on structured prototype regularization, which explicitly regularizes inter-class separability and intra-class compactness within the feature distributions, thereby improving CNN generalization to real-world driving scenes.
- Structural modeling of inter-class and intra-class feature relations for domain adaptation, with imposed constraints that guide more effective cross-domain feature alignment.
- Enhanced pixel-level feature alignment with entropy-based noise filtering and pixel-level attention, both incorporated into the contrastive learning process to improve the driving scene parsing accuracy on real-world data.

The remainder of this article is structured as follows. Section 2 reviews related state-of-the-art methods. Section 3 presents the technical details of the proposed SPR method. Section 4 presents both quantitative and qualitative experimental results, comparing the proposed method with other state-of-the-art UDA methods. Section 5 provides theoretical insights into the role of structured prototype regularization in improving domain adaptation performance. Finally, Section 6 concludes the article and discusses potential directions for future research.

2 Related work

2.1 Unsupervised domain adaptation for driving scene parsing

Driving scene parsing aims to assign pixel-level semantic labels to road scene images, enabling a detailed understanding of the driving environment [25–31]. UDA methods alleviate the heavy dependency of CNN-based segmentation models on annotated data, thereby facilitating more cost-effective real-world deployment [32]. Currently, UDA methods for driving scene parsing primarily fall into three categories: (1) style transfer [33], (2) feature alignment [33, 34], and (3) self-training [35]. Inspired by recent advances in unpaired image-to-image translation [33], style transfer methods transform the style of synthetic (source) images to resemble real-world (target) images [33]. Feature alignment methods [23,36] reduce domain discrepancies by aligning feature distributions between the source and target domains. This is achieved either through statistical methods, such as minimizing the maximum mean discrepancy across different domains within domain-specific layers [37], or through adversarial learning, where a domain discriminator is used to encourage domain-invariant feature extraction [34]. Self-training methods [21,35,38] generate pseudo labels for unlabeled target-domain images and iteratively refine the model through retraining. Recent methods have incorporated more complex UDA frameworks. For instance, ADPL [39] proposes an adaptive dual-path learning strategy that alleviates visual inconsistency and enhances pseudo label quality by introducing two interactive single-domain adaptation paths, one for the source domain and one for the target domain. FREDOM [40] introduces a fairness-aware UDA framework for semantic segmentation by imposing a fairness objective on class distributions and leveraging a conditional structure network with self-attention to model structural dependencies.

Adversarial learning reduces the discrepancy in global feature distributions across domains by encouraging CNNs to extract domain-invariant representations. However, aligning global distributions alone does not guarantee effective class separability in the shared feature space [21]. To address this limitation, recent studies have incorporated class-specific information into the feature alignment process. For example, class labels are incorporated into the discriminator during adversarial learning to enable class-aware feature alignment and improve feature discriminability [24]. Nevertheless, most existing methods neglect to explicitly model structural correlations within the feature

space. Consequently, they struggle to ensure sufficient inter-class separability and intra-class compactness, thereby undermining the discriminative power of the learned representations in the target domain.

2.2 Contrastive learning for driving scene parsing

Contrastive learning [41] learns feature representations by encouraging similarity between positive (semantically similar) image pairs and dissimilarity between negative (semantically different) image pairs. Typically, positive pairs can be derived from different augmentations of the same image that preserve semantic content, or from samples of the same class across domains. In contrast, negative pairs are drawn from semantically different images, potentially spanning different domains. This learning paradigm guides the model to cluster semantically consistent features while pushing apart unrelated ones. Another prominent self-supervised method is self-training, which iteratively refines pseudo labels to improve performance in the absence of ground truth [42,43]. For instance, STC [44] utilizes contrastive objectives to develop association embeddings for video instance segmentation. More recently, contrastive learning has also been successfully adapted to driving scene parsing tasks [45]. In the field of UDA, contrastive learning has been employed to align image feature spaces across domains. CLST [13] leverages contrastive objectives to achieve more refined domain-adapted image features. EHTDI [46] explores target-domain decision boundaries and features via a mix-up strategy and multi-level contrastive learning. SePiCo [47] employs semantic-guided pixel contrast, incorporating both centroid-aware and distribution-aware pixel contrast, to learn class-discriminative and balanced pixel representations across domains. CACP [48] addresses domain adaptive semantic segmentation by leveraging both source-domain and target-domain prototypes to dynamically rectify pseudo labels, and applying contrastive learning on prototypes to enhance inter-class separability and reduce domain shifts.

Inspired by these advances, this study presents a novel contrastive learning-based UDA framework that explicitly enforces the consistency of inter-class and intra-class relations within the feature space, thereby enhancing domain-invariant representation learning for driving scene parsing.

3 Methodology

3.1 Preliminaries

In this article, the source domain is associated with a labeled dataset $\mathcal{D}_s$, which contains N_s pairs of synthetic RGB images $\boldsymbol{X}_s \in \mathbb{R}^{H\times W\times 3}$ and corresponding pixel-level semantic annotations $\boldsymbol{Y}_s \in \mathbb{R}^{H\times W}$ with respect to C semantic categories. The target domain is associated with an unlabeled dataset $\mathcal{D}_t$ containing N_t real-world RGB images $\boldsymbol{X}_t \in \mathbb{R}^{H\times W\times 3}$ without corresponding semantic annotations. The semantic segmentation model $\mathcal{M}_s$, which consists of a feature extractor $\mathcal{F}$ and a pixel-level classifier $\mathcal{C}$, is initially trained on the labeled source-domain data in a supervised manner by minimizing the following pixel-wise cross-entropy loss:

$$\mathcal{L}_{ce} = -\mathcal{S}\left(\boldsymbol{H}_s \odot \log(\sigma(\boldsymbol{L}_s))\right), \tag{1}$$

where the operator $\mathcal{S}$ represents the summation over all elements of a given tensor or matrix, $\boldsymbol{H}_s \in \{0,1\}^{H\times W\times C}$ denotes the one-hot encoding of $\boldsymbol{Y}_s$, $\odot$ represents the Hadamard product, $\sigma(\cdot)$ denotes the softmax function, and $\boldsymbol{L}_s \in \mathbb{R}^{H\times W\times C}$ denotes the logits produced by $\mathcal{M}_s$.

Although the segmentation model trained on source-domain data can directly generate softmax-based label predictions $\mathcal{C}(\mathcal{F}(\boldsymbol{X}_t))$ for images in the target domain, the domain discrepancies between the source and target domains often result in performance degradation. Furthermore, the lack of pixel-wise annotations in the target dataset hinders the direct fine-tuning of the semantic segmentation model, further limiting its generalization performance. To address this issue, the overall workflow of the proposed SPR framework is illustrated in Figure 2.

3.2 Prototype-prototype interaction

3.2.1 *Class prototype estimation*

Following the training of the segmentation model on the source-domain dataset, the prototype vector $\boldsymbol{p}^c = (p_1^c, \ldots, p_D^c)^\top \in \mathbb{R}^D$ is defined as the centroid of the feature vectors assigned to the c-th class, where D denotes the number of feature channels. Specifically, each element of the prototype is computed as

$$p_d^c = \frac{\mathcal{S}\big(\boldsymbol{L}_d \odot \mathbb{1}[\boldsymbol{Y}=c]\big)}{\mathcal{S}\big(\mathbb{1}[\boldsymbol{Y}=c]\big)}, \tag{2}$$

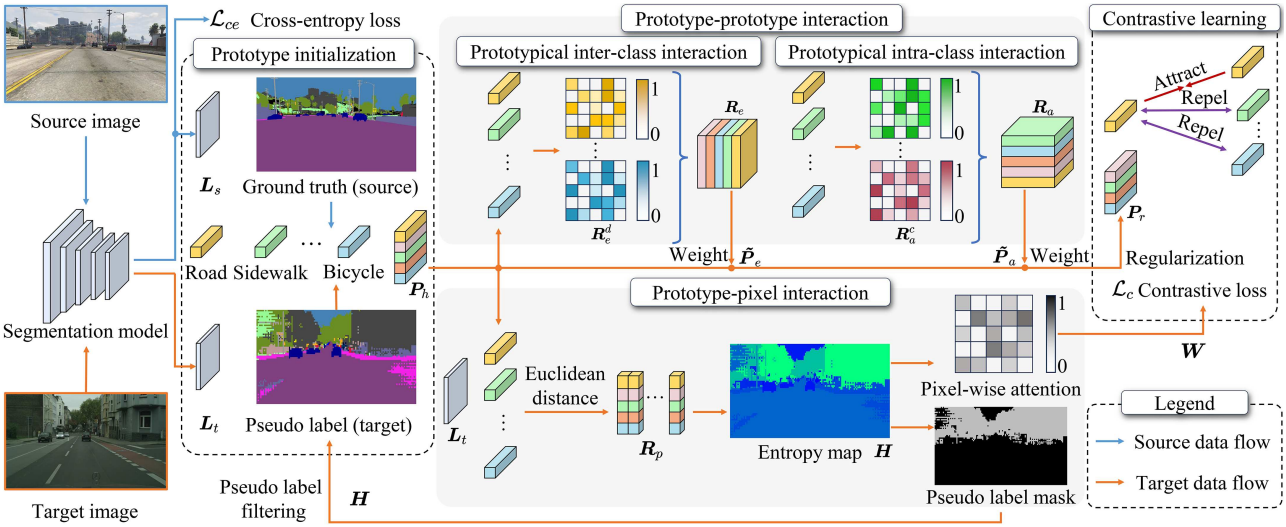

Figure 2 (Color online) Overview of the proposed SPR framework. The segmentation model is trained with cross-entropy loss $\mathcal{L}_{ce}$ on the source-domain dataset and contrastive loss $\mathcal{L}_c$ on datasets from both domains. Prototype-prototype interactions enforce inter-class separability and intra-class compactness by structurally regularizing class prototypes. Prototype-pixel interactions align pixel-wise features with these refined prototypes, incorporating entropy-based filtering and attention mechanisms to enhance semantic consistency in the target domain.

where $\boldsymbol{L}_d \in \mathbb{R}^{H\times W}$ denotes the d-th channel of the logits $\boldsymbol{L} \in \mathbb{R}^{H\times W\times D}$, and $\boldsymbol{Y} \in \mathbb{R}^{H\times W}$ represents the corresponding labels. The indicator function $\mathbb{1}[\cdot]$ selects pixels assigned to the c-th class based on the labels, where $\mathbb{1}[\boldsymbol{Y}=c] \in \{0,1\}^{H\times W}$. Finally, the prototype matrix $\boldsymbol{P} = \left(\boldsymbol{p}^1, \boldsymbol{p}^2, \ldots, \boldsymbol{p}^C\right) \in \mathbb{R}^{D\times C}$ is constructed by concatenating the class-specific prototypes of all C semantic classes, serving as an approximation of the semantic centroids in the feature space.

During the training process, the driving scene parsing model is trained using both the source and target domain data alternately. The source-domain prototype matrix $\boldsymbol{P}_s$ and target-domain prototype matrix $\boldsymbol{P}_t$ are computed based on the corresponding logits $\boldsymbol{L}_s$ and $\boldsymbol{L}_t$, and their respective ground-truth labels $\boldsymbol{Y}_s$ and pseudo labels $\boldsymbol{Y}_t$. The final prototype matrix $\boldsymbol{P}_h \in \mathbb{R}^{D\times C}$ is updated by linearly combining $\boldsymbol{P}_s$ and $\boldsymbol{P}_t$ with a weight γ, as shown in the following equation:

$$\boldsymbol{P}_h = \gamma \boldsymbol{P}_s + (1-\gamma)\boldsymbol{P}_t, \tag{3}$$

where γ is a hyperparameter that controls the contribution of the source and target domain prototypes to the final prototype matrix. In this study, γ is set to 0.5 to ensure a balanced integration of prototypes from both domains, allowing the model to retain source-domain representations that capture inter-class differences while adapting to the feature characteristics of the target domain for improved cross-domain stability.

3.2.2 *Prototypical intra-class and inter-class interactions*

To better characterize the feature distribution in the target domain and suppress the influence of noisy predictions, class prototypes are employed to represent the semantic structure of each category within the feature space. These prototypes serve as robust, noise-resistant summaries of class-specific features. By computing correlations between these prototypes, both inter-class separability and intra-class compactness are quantified, providing a structural perspective on feature distribution consistency.

As illustrated in Figure 3, the structural relationships between prototypes are first evaluated at the channel level. For the d-th feature channel, a prototype vector $\boldsymbol{p}^d \in \mathbb{R}^{1\times C}$ is extracted from the prototype matrix $\boldsymbol{P}_h$. The corresponding inter-class interaction matrix $\boldsymbol{R}_e^d \in \mathbb{R}^{C\times C}$ is then computed as

$$\boldsymbol{R}_e^d = \left(\boldsymbol{p}^d\right)^\top \boldsymbol{p}^d, \tag{4}$$

where each entry of $\boldsymbol{R}_e^d$ represents the correlation between the c-th and c'-th classes in that specific channel. Specifically, the off-diagonal elements ($c \neq c'$) indicate cross-class similarity; high values indicate significant correlation between distinct semantic categories, reflecting potential semantic ambiguity. Conversely, the diagonal entries represent class-specific self-correlation, which should be emphasized to ensure class compactness. Finally, the matrices $\{\boldsymbol{R}_e^d\}_{d=1}^D$ from all channels are stacked to form the comprehensive inter-class interaction tensor $\boldsymbol{R}_e \in \mathbb{R}^{D\times C\times C}$.

To explicitly enforce this desirable property, the interaction matrices are normalized within each channel to enhance the prominence of diagonal values (self-correlation) relative to off-diagonal (cross-correlation) values. This

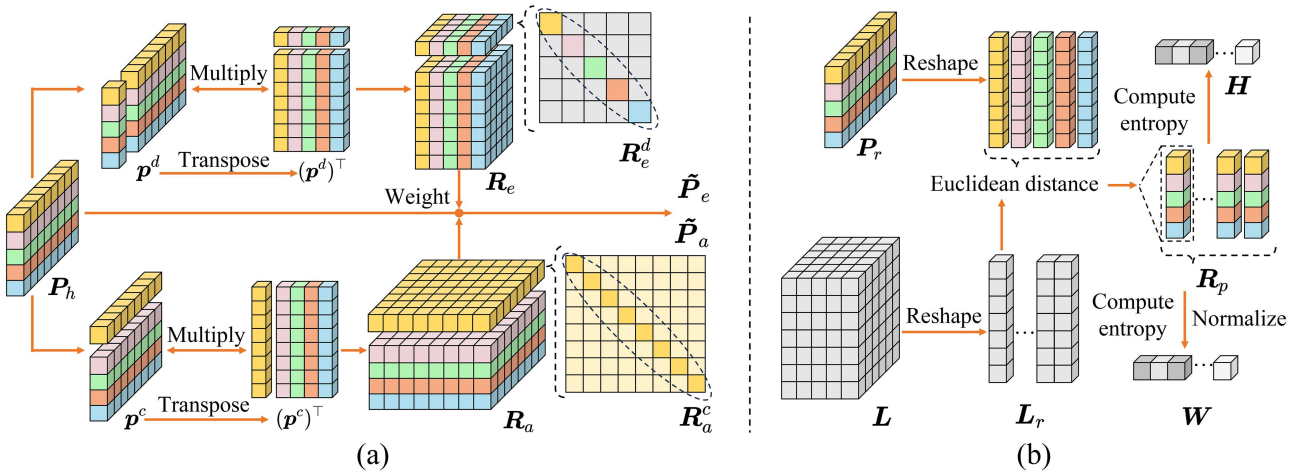

Figure 3 (Color online) Illustration of prototype-based structural modeling and pixel-wise uncertainty estimation. (a) Generation of the inter-class and intra-class weighted prototypes, $\tilde{\boldsymbol{P}}_e$ and $\tilde{\boldsymbol{P}}_a$, via prototype-prototype interaction; (b) estimation of the entropy map $\boldsymbol{H}$ and attention weight map $\boldsymbol{W}$ through prototype-pixel interaction.

normalization also ensures consistent scaling across channels, facilitating the formulation of regularization terms. For each channel, the normalized inter-class interaction $\tilde{\boldsymbol{R}}_e$ is defined as

$$\tilde{\boldsymbol{R}}_e^{(d,c,c')} = \frac{\boldsymbol{R}_e^{(d,c,c')}}{\boldsymbol{R}_e^{(d,c,c)} + \epsilon}, \tag{5}$$

where ϵ is a small constant introduced to prevent division by zero.

Complementary to inter-class relationships, the proposed framework further models intra-class structural consistency to characterize the internal feature distribution of each semantic category across channels. Specifically, the c-th column of the prototype matrix $\boldsymbol{P}_h$ is extracted to form a prototype vector $\boldsymbol{p}^c \in \mathbb{R}^{1\times D}$ representing the channel-wise feature response of the corresponding category. The intra-class interaction matrix is then defined as

$$\boldsymbol{R}_a^c = (\boldsymbol{p}^c)^\top \boldsymbol{p}^c, \tag{6}$$

where $\boldsymbol{R}_a^c$ quantifies the correlation between the d-th and d'-th feature channels for the c-th class. Specifically, the diagonal elements ($d = d'$) represent self-correlations of individual channels, reflecting the strength of class-specific features, while the off-diagonal elements ($d \neq d'$) capture the correlations between different channels. A higher prominence of diagonal entries indicates that the feature representation of the class is compact and internally consistent. Finally, the class-specific interaction matrices $\{\boldsymbol{R}_a^c\}_{c=1}^C$ are aggregated to construct a comprehensive intra-class interaction tensor $\boldsymbol{R}_a \in \mathbb{R}^{C\times D\times D}$. To explicitly encourage internal consistency, the matrices are normalized across channels, yielding the normalized intra-class correlation $\tilde{\boldsymbol{R}}_a$:

$$\tilde{\boldsymbol{R}}_a^{(c,d,d')} = \frac{\boldsymbol{R}_a^{(c,d,d')}}{\boldsymbol{R}_a^{(c,d,d)} + \epsilon}, \tag{7}$$

where ϵ is a small constant introduced to prevent division by zero. This normalization emphasizes the relative importance of off-diagonal elements compared to dominant self-correlations and ensures a consistent scale for regularization across channels.

3.2.3 *Regularization based on prototypical interaction*

To effectively preserve the semantic structure embedded in the prototypes during cross-domain feature alignment, the normalized correlation matrices, $\tilde{\boldsymbol{R}}_a$ and $\tilde{\boldsymbol{R}}_e$, are leveraged as adaptive weights to recalibrate the original prototype matrix $\boldsymbol{P}_h$. The inter-class weighted prototype $\tilde{\boldsymbol{P}}_e \in \mathbb{R}^{D\times C}$ is obtained by integrating cross-class information, where each feature channel is scaled by its corresponding inter-class normalized correlation:

$$\tilde{\boldsymbol{P}}_e^{(d,c)} = \sum_{c'=1}^{C} \tilde{\boldsymbol{R}}_e^{(d,c,c')} \cdot \boldsymbol{P}_h^{(d,c')}. \tag{8}$$

The intra-class weighted prototype $\tilde{\boldsymbol{P}}_a \in \mathbb{R}^{D\times C}$ is obtained by aggregating feature channels within each class, where each channel is weighted by its corresponding intra-class normalized correlation:

$$\tilde{\boldsymbol{P}}_a^{(d,c)} = \sum_{d'=1}^{D} \tilde{\boldsymbol{R}}_a^{(c,d,d')} \cdot \boldsymbol{P}_h^{(d',c)}. \tag{9}$$

These weighted prototypes, $\tilde{\boldsymbol{P}}_e$ and $\tilde{\boldsymbol{P}}_a$, represent the inter-class ambiguity and intra-class inconsistency captured by the prototype correlations, respectively. Intuitively, $\tilde{\boldsymbol{P}}_e$ emphasizes features shared across classes that may cause semantic confusion, while $\tilde{\boldsymbol{P}}_a$ emphasizes channel-level interactions that reduce compactness within each class. To incorporate structural regularization during cross-domain feature alignment, the regularized prototype matrix is derived by refining the original $\boldsymbol{P}_h$ with weighted correlation terms as follows:

$$\boldsymbol{P}_r = \boldsymbol{P}_h - \lambda_e \tilde{\boldsymbol{P}}_e - \lambda_a \tilde{\boldsymbol{P}}_a, \tag{10}$$

where the hyperparameters λ_e and λ_a control the strengths of the inter-class and intra-class regularization terms, respectively. Empirical evaluations show that setting both λ_e and λ_a to 0.1 achieves optimal performance. Under these settings, the structural corrections refine the prototype matrix $\boldsymbol{P}_h$, enhancing its semantic compactness within each class while preserving or improving its discriminability across classes. Serving as a robust anchor for contrastive learning, the refined $\boldsymbol{P}_r$ effectively integrates intrinsic structural properties into the alignment process, eliminating the need for additional supervision.

3.3 Prototype-pixel interaction

To enhance the domain adaptation process, a prototype-pixel interaction mechanism is introduced, which regularizes feature distributions by applying global structural constraints to local pixels, with the aim of reducing pixel-level semantic ambiguity. As shown in Figure 3(b), the logits $\boldsymbol{L}$ are reshaped into $\boldsymbol{L}_r \in \mathbb{R}^{N\times C}$, where $N = H \times W$ denotes the total number of spatial positions. The semantic similarity between pixel-wise logits and class prototypes is quantified using the squared Euclidean distance:

$$\boldsymbol{R}_p^{(n,c)} = \left\| \boldsymbol{L}_r^{(n,:)} - \boldsymbol{P}_r^{(:,c)} \right\|_2^2, \tag{11}$$

where $\boldsymbol{R}_p \in \mathbb{R}^{N\times C}$ captures the similarity between each pixel-level logits and different semantic prototype vectors, serving as an indicator of classification confidence.

3.3.1 *Adaptive pseudo label threshold*

To identify reliable pixels for effective alignment, the prediction confidence is characterized by the entropy of the probability distribution derived from the distance matrix $\boldsymbol{R}_p$. The distance matrix is first normalized using the following softmax function:

$$\boldsymbol{Q}^{(n,c)} = \frac{\exp\left(-\boldsymbol{R}_p^{(n,c)}\right)}{\sum_{c'=1}^{C} \exp\left(-\boldsymbol{R}_p^{(n,c')}\right)}. \tag{12}$$

The pixel-wise prediction uncertainty is then computed as the entropy map $\boldsymbol{H}$ as follows:

$$\boldsymbol{H}^{(n)} = -\sum_{c=1}^{C} \boldsymbol{Q}^{(n,c)} \log \boldsymbol{Q}^{(n,c)}, \tag{13}$$

where $\boldsymbol{H} \in \mathbb{R}^{N\times 1}$ quantifies the uncertainty of the pixel-level semantic predictions. Higher entropy values indicate greater uncertainty, reflecting more ambiguity in the pixel's predicted class, while lower entropy values suggest more confident and accurate predictions. To eliminate unreliable samples, only the top-α fraction of pixels with the lowest entropy is retained to construct a binary mask for the subsequent feature distribution alignment process. This ratio-based selection strategy eliminates the requirement for manual thresholding, thereby enhancing the adaptability and flexibility of the method across diverse target domains with varying uncertainty distributions.

3.3.2 *Pixel-wise attention for contrastive learning*

In addition to the selective filtering of unreliable samples, the entropy information is further leveraged as a soft attention mechanism to modulate the feature alignment process. By normalizing these entropy scores, the model assigns greater weight to pixels with higher ambiguity. This strategy prioritizes the refinement of challenging semantic regions, which typically necessitate more granular and precise domain alignment. Specifically, the pixel-wise weights $\boldsymbol{W} \in \mathbb{R}^{N\times 1}$ are derived by normalizing the entropy map $\boldsymbol{H}$ relative to its global maximum value. The normalized weight $\boldsymbol{W}$ serves as a relative measure of semantic uncertainty to guide the adaptation process. This mechanism assigns higher weights to more ambiguous features, directing contrastive supervision to the most challenging regions, which are crucial for bridging the domain gap.

3.3.3 *Contrastive adaptation*

To facilitate semantically structured domain alignment, a prototype-based contrastive loss is developed to incorporate both structural regularization and pixel-wise attention. Specifically, the pixel-wise similarity between features and category-specific prototypes is computed for each spatial location $\boldsymbol{x}$ as a vector $\boldsymbol{p}_t = (p_t^1, \cdots, p_t^C)^\top \in \mathbb{R}^C$, where each category-specific similarity score p_t^c is calculated as follows:

$$p_t^c = \frac{\exp\left(\boldsymbol{p}_r^c \cdot \boldsymbol{L}_t(\boldsymbol{x}) \cdot \boldsymbol{W}(\boldsymbol{x})/\tau\right)}{\sum_{k=1}^{C} \exp\left(\boldsymbol{p}_r^k \cdot \boldsymbol{L}_t(\boldsymbol{x}) \cdot \boldsymbol{W}(\boldsymbol{x})/\tau\right)}, \tag{14}$$

where $\boldsymbol{x}$ denotes the spatial coordinate (h, w) of a specific pixel and τ is the temperature coefficient that scales the similarity values. The vector $\boldsymbol{p}_t$ captures the pixel-wise similarity distribution across all C categories. Finally, the pixel-wise similarity vectors $\boldsymbol{p}_t$ across the entire image are combined into a similarity tensor $\boldsymbol{S}_t \in \mathbb{R}^{H\times W\times C}$.

To improve cross-domain feature alignment, higher weights are assigned to pixels near decision boundaries. This strategy amplifies the supervisory signal in ambiguous regions, ensuring that challenging spatial locations contribute more effectively to narrowing the domain gap. Due to the unavailability of ground-truth annotations in the target domain, pseudo labels $\boldsymbol{Y}_t$ derived from model predictions are employed. To ensure their quality, an adaptive thresholding scheme is applied to filter out low-confidence predictions, thereby mitigating the risk of noisy supervision. Leveraging the similarity scores and the filtered pseudo labels, the contrastive loss between the target-domain features and the class prototypes is defined as follows:

$$\mathcal{L}_t = -\mathcal{S}\left(\boldsymbol{H}_t \odot \log(\boldsymbol{S}_t)\right), \tag{15}$$

where $\boldsymbol{H}_t \in \{0,1\}^{H\times W\times C}$ denotes the one-hot encoding of pseudo labels $\boldsymbol{Y}_t$. Additionally, intra-domain consistency within the source domain is enforced by defining a source-to-source contrastive loss:

$$\mathcal{L}_s = -\mathcal{S}\left(\boldsymbol{H}_s \odot \log(\boldsymbol{S}_s)\right), \tag{16}$$

where $\boldsymbol{H}_s \in \{0,1\}^{H\times W\times C}$ denotes the one-hot encoding of ground-truth labels $\boldsymbol{Y}_s$. The tensor $\boldsymbol{S}_s$ is computed in a similar manner to $\boldsymbol{S}_t$, capturing the pixel-wise similarity between source-domain features and class prototypes. The overall contrastive adaptation objective integrates both directions:

$$\mathcal{L}_c = \mathcal{L}_s + \mathcal{L}_t. \tag{17}$$

This formulation facilitates the alignment of pixel-wise features with their corresponding class prototypes while allocating greater importance to semantically uncertain regions. By integrating contrastive learning with an attention mechanism and adaptive pseudo labeling, the proposed method improves inter-class separability and enhances intra-class compactness, facilitating more robust cross-domain alignment. To further improve adaptation performance, an additional self-training stage is incorporated. Following contrastive training, the model generates refined pseudo labels for target-domain images, which are subsequently employed to guide a self-supervised learning phase, thereby enhancing the model's performance in the target domain.

4 Experiments

4.1 Datasets and evaluation metrics

Following previous studies [22,32,39,40,47,48], the proposed method is evaluated on three widely used unsupervised domain adaptation benchmarks for driving scene parsing: (1) GTA5 → Cityscapes, (2) SYNTHIA → Cityscapes,

and (3) Cityscapes → ACDC. The GTA5 dataset [49] includes 24966 synthetic images with a resolution of 1914×1052 pixels and contains 19 semantic classes aligned with the Cityscapes dataset. The SYNTHIA dataset [12] contains 9400 synthetic images with a resolution of 1280×760 pixels. The Cityscapes dataset [10] contains 2975 real-world images for training, 500 real-world images for validation, and 1525 real-world images for testing, all with a resolution of 2048×1024 pixels. The Adverse Conditions Dataset with Correspondences (ACDC) [11] is specifically designed to capture adverse visual conditions in the real world, including fog, night, rain, and snow, and shares the same semantic classes as the Cityscapes dataset. In this study, mean intersection over union (mIoU) is used as the primary evaluation metric, providing a comprehensive quantification of overall scene parsing performance across semantic classes.

4.2 Implementation details

Consistent with previous studies [21, 24, 32, 50, 51], the DeepLab-v2 model [52] with a ResNet-101 [53] encoder is adopted as the baseline segmentation model. All methods involved in the comparison are initialized with weights pre-trained on the ImageNet dataset [54]. Atrous spatial pyramid pooling (ASPP) [52] is used after the final encoder layer with dilation rates of {6, 12, 18, 24}. An up-sampling layer is then employed to generate the segmentation outputs of the driving scene parsing model. To simulate real-world deployment scenarios, the source-domain dataset is partitioned into training, validation, and test sets in a ratio of 6:1:3, and the initial weights are selected based on the validation set. All methods involved in the comparison are implemented using PyTorch and executed on a heterogeneous system equipped with an Intel(R) Xeon(R) Gold 6154 CPU and a single NVIDIA RTX 3090 GPU with 24 GB memory. The proposed SPR model is optimized using the stochastic gradient descent optimizer with an initial learning rate of 2.5×10^{-4}, a momentum of 0.9, and a weight decay of 5.0×10^{-4}. The learning rate is adjusted according to a polynomial decay schedule with a power of 0.9.

4.3 Comparisons with state-of-the-art unsupervised domain adaptation methods

4.3.1 *Experimental results on GTA5 → Cityscapes adaptation task*

The quantitative and qualitative experimental results of the compared methods [13,22–24,32,36,38–40,46–48,50–52, 55–71] are presented in Table 1 and Figure 4, respectively. These results suggest that the proposed SPR method outperforms the baseline and various existing methods, particularly those based on adversarial learning [24, 36, 57, 67]. Notably, the proposed method acts primarily on the feature adaptation stage and is therefore independent of, yet complementary to, improvements obtained through self-training (ST) strategies. It can be effectively integrated with self-training methods to further enhance performance. Specifically, SPR achieves an mIoU of 53.1%, representing a substantial 16.0% improvement over the baseline. When integrated with self-training, SPR further boosts performance to 54.4%, supporting our hypothesis that explicitly modeling inter-class separability and intra-class compactness is essential for feature alignment. Furthermore, while several state-of-the-art (SoTA) methods optimize for long-tail categories, SPR demonstrates competitive performance on large-scale categories. In particular, SPR+ST outperforms recent methods such as CACP [48] and SePiCo [47] by achieving 91.5% IoU on building, 91.1% on vegetation, 58.9% on terrain, and 65.6% on bus. These results indicate that SPR extends traditional prototype-based methods by introducing structural regularization, yielding superior semantic consistency for major scene components.

To further evaluate SPR, qualitative segmentation results are presented in Figure 4. Notably, the driving scene parsing results exhibit significant improvement over the source-only baseline. Additionally, feature distributions in the output space are visualized using t-SNE [72], as illustrated in Figure 5. These visualizations demonstrate that SPR induces more compact intra-class clusters and clearer inter-class separation, thereby validating the effectiveness of the structural regularization strategy.

4.3.2 *Experimental results on SYNTHIA → Cityscapes adaptation task*

The quantitative and qualitative experimental results of the compared methods are presented in Table 2 and Figure 6, respectively. The domain discrepancy in the SYNTHIA → Cityscapes adaptation task is significantly more pronounced than in the GTA5 → Cityscapes adaptation task due to substantial differences in spatial layout. As shown in Table 2, SPR effectively addresses these alignment difficulties, delivering substantial improvements over the baseline model. While traditional methods [36, 67] based on adversarial learning often struggle with this heightened discrepancy and yield limited gains, the proposed SPR shows superior adaptability. Notably, when integrated with self-training strategies, SPR+ST achieves a remarkable performance boost, reaching an mIoU* of 59.4%. This confirms that the feature alignment strategy provides a solid foundation for subsequent refinement stages.

Table 1 Quantitative comparison of state-of-the-art unsupervised domain adaptation methods for GTA5 → Cityscapes adaptation task based on per-class IoU (%) and overall mIoU (%). Bold values indicate the best performance.

Method	Publication	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bike	mIoU	Gain
Source Only [52]	T-PAMI'17	53.8	15.6	69.3	28.1	18.8	27.6	34.9	18.2	82.5	27.8	71.6	59.4	35.3	44.1	25.9	37.5	0.1	28.9	24.9	37.1	+0.0
PatchAlign [55]	CVPR'19	92.3	51.9	82.1	29.2	25.1	24.5	33.8	33.0	82.4	32.8	82.2	58.6	27.2	84.3	33.4	46.3	2.2	29.5	32.3	46.5	+9.4
ADVENT [56]	CVPR'19	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5	+8.4
AdaptSeg [23]	CVPR'18	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4	+5.3
BDL [57]	CVPR'19	91.0	44.7	84.2	34.6	27.6	30.2	36.0	36.0	85.0	43.6	83.0	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5	+11.4
UIDA [58]	CVPR'20	90.6	37.1	82.6	30.1	19.1	29.5	32.4	20.6	85.7	40.5	79.7	58.7	31.1	86.3	31.5	48.3	0.0	30.2	35.8	46.3	+9.2
LTIR [59]	CVPR'20	92.9	55.0	85.3	34.2	31.1	34.9	40.7	34.0	85.2	40.1	87.1	61.0	31.1	82.5	32.3	42.9	0.3	36.4	46.1	50.2	+13.1
PIT [60]	CVPR'20	87.5	43.4	78.8	31.2	30.2	36.3	39.9	42.0	79.2	37.1	79.3	65.4	37.5	83.2	46.0	45.6	25.7	23.5	49.9	50.6	+13.5
LSE [61]	ECCV'20	90.2	40.0	83.5	31.9	26.4	32.6	38.7	37.5	81.0	34.2	84.6	61.6	33.4	82.5	32.8	45.9	6.7	29.1	30.6	47.5	+10.4
WeakSeg [62]	ECCV'20	91.6	47.4	84.0	30.4	28.3	31.4	37.4	35.4	83.9	38.3	83.9	61.2	28.2	83.7	28.8	41.3	8.8	24.7	46.4	48.2	+11.1
CrCDA [63]	ECCV'20	92.4	55.3	82.3	31.2	29.1	32.5	33.2	35.6	83.5	34.8	84.2	58.9	32.2	84.7	40.6	46.1	2.1	31.1	32.7	48.6	+11.5
FADA [24]	ECCV'20	92.5	47.5	85.1	37.6	32.8	33.4	33.8	18.4	85.3	37.7	83.5	63.2	39.7	87.5	32.9	47.8	1.6	34.9	39.5	49.2	+12.1
IAST [38]	ECCV'20	94.1	58.8	85.4	39.7	29.2	25.1	43.1	34.2	84.8	34.6	88.7	62.7	30.3	87.6	42.3	50.3	24.7	35.2	40.2	52.2	+15.1
ASA [64]	TIP'21	89.2	27.8	81.3	25.3	22.7	28.7	36.5	19.6	83.8	31.4	77.1	59.2	29.8	84.3	33.2	45.6	16.9	34.5	30.8	45.1	+8.0
CLAN [36]	T-PAMI'21	88.7	35.5	80.3	27.5	25.0	29.3	36.4	28.1	84.5	37.0	76.6	58.4	29.7	81.2	38.8	40.9	5.6	32.9	28.8	45.5	+8.4
DACS [65]	WACV'21	89.9	39.7	87.9	39.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1	+15.0
RPLL [66]	IJCV'21	90.4	31.2	85.1	36.9	25.6	37.5	48.8	48.5	85.3	34.8	81.1	64.4	36.8	86.3	34.9	52.2	1.7	29.0	44.6	50.3	+13.2
DAST [67]	AAAI'21	92.2	49.0	84.3	36.5	28.9	33.9	38.8	28.4	84.9	41.6	83.2	60.0	28.7	87.2	45.0	45.3	7.4	33.8	32.8	49.6	+12.5
ConTrans [68]	AAAI'21	95.3	65.1	84.6	33.2	23.7	32.8	32.7	36.9	86.0	41.0	85.6	56.1	25.9	86.3	34.5	39.1	11.5	28.3	43.0	49.6	+12.5
CIRN [69]	AAAI'21	91.5	48.7	85.2	33.1	26.0	32.3	33.8	34.6	85.1	43.6	86.9	62.2	28.5	84.6	37.9	47.6	0.0	35.0	36.0	49.1	+12.0
CLST [13]	IJCNN'21	92.8	53.5	86.1	39.1	28.1	28.9	43.6	39.4	84.6	35.7	88.1	63.9	38.3	86.0	41.6	50.6	0.1	30.4	51.7	51.6	+14.5
MetaCorrect [70]	CVPR'21	92.8	58.1	86.2	39.7	33.1	36.3	42.0	38.6	85.5	37.8	87.6	62.8	31.7	84.8	35.7	50.3	2.0	36.8	48.0	52.1	+15.0
ProDA [22]	CVPR'21	91.5	52.3	82.9	42.0	35.7	40.0	44.4	43.2	87.0	43.8	79.5	66.4	31.3	86.7	41.1	52.5	0.0	45.4	53.8	53.7	+16.6
UPLR [71]	ICCV'21	90.5	38.7	86.5	41.1	32.9	40.5	48.2	42.1	86.5	36.8	84.2	64.5	38.1	87.2	34.8	50.4	0.2	41.8	54.6	52.6	+15.5
EHTDI [46]	ACM MM'22	95.4	68.8	88.1	37.1	**41.4**	42.5	45.7	60.4	87.3	42.6	86.8	67.4	38.6	90.5	**66.7**	61.4	0.3	39.4	56.1	58.8	+21.7
EIC [51]	WACV'23	90.8	47.2	86.8	41.5	29.4	35.7	42.4	37.4	86.0	42.1	88.3	63.7	35.6	85.1	43.8	54.6	0.0	33.6	47.8	52.2	+15.1
ARAS [50]	TCSVT'23	91.9	45.2	81.8	21.9	25.6	35.5	41.5	33.4	85.1	34.8	73.8	62.5	31.6	85.9	33.8	42.5	7.3	33.8	42.8	47.9	+10.8
ADPL-Dual [39]	T-PAMI'23	93.4	60.6	87.5	45.3	32.6	37.3	43.3	55.5	87.2	44.8	88.0	64.5	34.2	88.3	52.6	61.8	49.8	41.8	59.4	59.4	+22.3
SePiCo [47]	T-PAMI'23	95.2	67.8	88.7	41.4	38.4	43.4	**55.5**	**63.2**	88.6	46.4	88.3	**73.1**	**49.0**	**91.4**	63.2	60.4	0.0	45.2	**60.0**	61.0	+23.9
FREDOM [40]	CVPR'23	90.9	54.1	87.8	44.1	32.6	**45.2**	51.4	57.1	88.6	42.6	89.5	68.8	40.0	89.7	58.4	62.6	55.3	47.7	58.1	61.3	+24.2
PBAL [32]	T-MM'24	88.3	38.2	85.3	32.5	29.0	37.5	43.7	40.5	86.7	33.5	84.8	65.9	28.6	85.5	33.3	31.3	26.8	20.4	48.6	49.5	+12.4
CACP [48]	T-MM'25	**95.5**	**70.4**	88.1	**52.3**	36.2	44.4	53.1	57.2	89.6	51.3	**90.0**	71.3	42.9	88.5	32.7	61.3	**65.0**	**53.1**	47.0	**62.6**	**+25.5**
SPR (Ours)	SCIS'26	90.7	43.2	91.0	40.3	29.6	32.9	41.5	38.0	91.0	50.9	89.0	63.1	38.2	81.7	43.3	60.2	10.8	31.2	37.5	53.1	+16.0
SPR+ST (Ours)	SCIS'26	90.4	40.7	**91.5**	47.1	30.8	32.8	41.0	33.2	**91.1**	**58.9**	87.7	61.3	35.0	79.1	51.5	**65.6**	0.2	43.7	52.6	54.4	+17.3

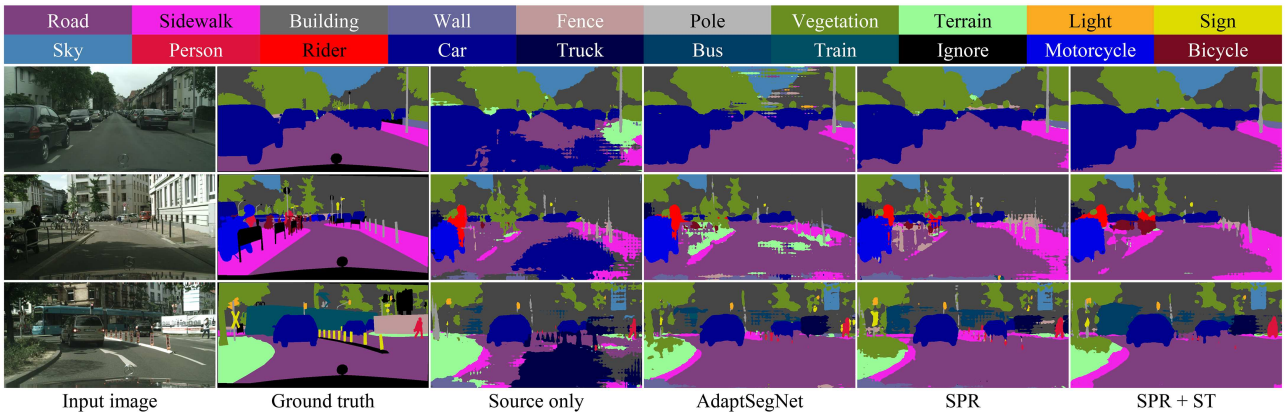

Figure 4 (Color online) Qualitative results obtained on the GTA5 → Cityscapes adaptation task. From left to right, the figure displays the input target-domain image and its corresponding ground truth, followed by the segmentation results predicted by four different methods: the source-only baseline model, AdaptSegNet [23], SPR, and SPR with the self-training strategy.

Compared to existing UDA methods, SPR offers enhanced consistency for major, safety-critical categories. While SoTA methods such as CACP [48] and SePiCo [47] optimize for rare categories to boost mIoU, they often suffer from severe degradation on essential navigable regions. For instance, these methods exhibit poor performance on critical categories like road and sidewalk, dropping to approximately 77% and 37% IoU, respectively. Conversely, SPR+ST maintains superior performance on these major classes, reaching 97.9% IoU on road and 52.6% IoU on sidewalk. These significant performance gaps, exceeding 20% and 15%, respectively, validate that structural regularization effectively maintains the global semantic layout required for safe autonomous driving. Finally, Figure 6 presents qualitative results for selected samples, visually confirming the segmentation quality of SPR.

Table 2 Quantitative comparison of state-of-the-art unsupervised domain adaptation methods for SYNTHIA → Cityscapes adaptation task based on per-class IoU (%) and overall mIoU (%). ∗ indicates challenging categories; mIoU* (%) is calculated over the remaining categories. Bold values indicate the best performance.

Method	Publication	Road	Sidewalk	Building	Wall*	Fence*	Pole*	Light	Sign	Vegetation	Sky	Person	Rider	Car	Bus	Motorcycle	Bike	mIoU	Gain	mIoU*	Gain
Source Only [52]	T-PAMI'17	55.6	23.8	74.6	9.2	0.2	24.4	6.1	12.1	74.8	79.0	55.3	19.1	39.6	23.3	13.7	25.0	33.5	0.0	38.6	+0.0
PatchAlign [55]	CVPR'19	82.4	38.0	78.6	–	–	–	9.9	10.5	78.2	80.5	53.5	19.6	67.0	29.5	21.6	31.3	–	–	46.2	+7.6
ADVENT [56]	CVPR'19	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2	+7.7	48.0	+9.4
AdaptSeg [23]	CVPR'18	84.3	42.7	77.5	–	–	–	4.7	7.0	77.9	82.5	54.3	21.0	72.3	32.2	18.9	32.3	–	–	46.7	+8.1
BDL [57]	CVPR'19	86.0	46.7	80.3	–	–	–	14.1	11.6	79.2	81.3	54.1	27.9	73.7	42.2	25.7	45.3	–	–	51.4	+12.8
UIDA [58]	CVPR'20	84.3	37.7	79.5	5.3	0.4	24.9	9.2	8.4	80.0	84.1	57.2	23.0	78.0	38.1	20.3	36.5	41.7	+8.2	48.9	+10.3
LTIR [59]	CVPR'20	92.6	53.2	79.2	–	–	–	1.6	7.5	78.6	84.4	52.6	20.0	82.1	34.8	14.6	39.4	–	–	49.3	+10.7
PIT [60]	CVPR'20	83.1	27.6	81.5	8.9	0.3	21.8	26.4	33.8	76.4	78.8	64.2	27.6	79.6	31.2	31.0	31.3	44.0	+10.5	51.8	+13.2
LSE [61]	ECCV'20	82.9	43.1	78.1	9.3	0.6	28.2	9.1	14.4	77.0	83.5	58.1	25.9	71.9	38.0	29.4	31.2	42.5	+9.0	49.4	+10.8
WeakSeg [62]	ECCV'20	92.0	53.5	80.9	11.4	0.4	21.8	3.8	6.0	81.6	84.4	60.8	24.4	80.5	39.0	26.0	41.7	44.3	+10.8	51.9	+13.3
CrCDA [63]	ECCV'20	86.2	44.9	79.5	8.3	0.7	27.8	9.4	11.8	78.6	86.5	57.2	26.1	76.8	39.9	21.5	32.1	42.9	+9.4	50.0	+11.4
FADA [24]	ECCV'20	84.5	40.1	83.1	4.8	0.0	34.3	20.1	27.2	84.8	84.0	53.5	22.6	85.4	43.7	26.8	27.8	45.2	+11.7	52.5	+13.9
IAST [38]	ECCV'20	81.9	41.5	83.3	17.7	4.6	32.3	30.9	28.8	83.4	85.0	65.5	30.8	86.5	38.2	33.1	52.7	49.8	+16.3	57.0	+18.4
ASA [64]	TIP'21	91.2	48.5	80.4	3.7	0.3	21.7	5.5	5.2	79.5	83.6	56.4	21.9	80.3	36.2	20.0	32.9	41.7	+8.2	49.3	+10.7
CLAN [36]	T-PAMI'21	82.7	37.2	81.5	–	–	–	17.1	13.1	81.2	83.3	55.5	22.1	76.6	30.1	23.5	30.7	–	–	48.8	+10.2
DACS [65]	WACV'21	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	**90.8**	67.6	38.3	82.9	38.9	28.5	47.6	48.3	+14.8	54.8	+16.2
RPLL [66]	IJCV'21	87.6	41.9	83.1	14.7	1.7	36.2	31.3	19.9	81.6	80.6	63.0	21.8	86.2	40.7	23.6	53.1	47.9	+14.4	54.9	+16.3
DAST [67]	AAAI'21	87.1	44.5	82.3	10.7	0.8	29.9	13.9	13.1	81.6	86.0	60.3	25.1	83.1	40.1	24.4	40.5	45.2	+11.7	52.5	+13.9
ConTrans [68]	AAAI'21	93.3	54.0	81.3	14.3	0.7	28.8	21.3	22.8	82.6	83.3	57.7	22.8	83.4	30.7	20.2	47.2	46.5	+13.0	53.9	+15.3
CIRN [69]	AAAI'21	85.8	40.4	80.4	4.7	1.8	30.8	16.4	18.6	80.7	80.4	55.2	26.3	83.9	43.8	18.6	34.3	43.9	+10.4	51.1	+12.5
CLST [13]	IJCNN'21	88.0	49.2	82.2	16.3	0.4	29.2	31.8	23.9	84.1	88.0	59.1	27.2	85.5	46.4	28.9	56.5	49.8	+16.3	57.8	+19.2
MetaCorrect [70]	CVPR'21	92.6	52.7	81.3	8.9	2.4	28.1	13.0	7.3	83.5	85.0	60.1	19.7	84.8	37.2	21.5	43.9	45.1	+11.6	52.5	+13.9
ProDA [22]	CVPR'21	87.1	44.0	83.2	26.9	0.0	42.0	45.8	34.2	86.7	81.3	68.4	22.1	**87.7**	50.0	31.4	38.6	51.9	+18.4	58.5	+19.9
UPLR [71]	ICCV'21	79.4	34.6	83.5	19.3	2.8	35.3	32.1	26.9	78.8	79.6	66.6	30.3	86.1	36.6	19.5	56.9	48.0	+14.5	54.6	+16.0
EHTDI [46]	ACM MM'22	93.0	**69.8**	84.0	36.6	**9.1**	39.7	42.2	43.8	**88.2**	88.1	68.3	29.0	85.5	54.1	37.1	56.3	57.8	+24.3	64.6	+26.0
ADPL-Dual [39]	T-PAMI'23	86.1	38.6	85.9	29.7	1.3	36.6	41.3	47.2	85.0	90.4	67.5	44.3	87.4	57.1	43.9	51.4	55.9	+22.4	63.6	+25.0
EIC [51]	WACV'23	77.5	32.3	82.6	25.5	1.9	34.6	33.6	32.4	81.7	85.1	63.8	31.8	82.3	35.2	31.9	54.6	49.2	+15.7	55.7	+17.1
ARAS [50]	TCSVT'23	85.6	39.2	79.9	15.5	0.3	32.2	19.3	23.9	79.1	81.7	61.1	19.3	82.9	25.7	10.6	51.9	44.3	+10.8	50.8	+12.2
SePiCo [47]	T-PAMI'23	77.0	35.3	85.1	23.9	3.4	38.0	51.0	**55.1**	85.6	80.5	73.5	**46.3**	87.6	69.7	50.9	**66.5**	58.1	+24.6	66.5	+27.9
FREDOM [40]	CVPR'23	86.0	46.3	**87.0**	33.3	5.3	**48.7**	**53.4**	46.8	87.1	89.1	71.2	38.1	87.1	54.6	51.3	59.9	59.1	+25.6	–	–
PBAL [32]	T-MM'24	85.3	42.5	81.7	10.2	0.1	36.8	23.6	31.8	85.1	87.6	64.1	27.7	85.2	31.4	23.0	35.6	47.0	+13.5	54.2	+15.6
CACP [48]	T-MM'25	77.8	37.5	85.9	**37.4**	0.0	44.4	52.7	50.9	**88.2**	88.0	**75.1**	38.2	87.4	**78.9**	**56.0**	55.7	**59.6**	**+26.1**	**67.1**	**+28.5**
SPR (Ours)	SCIS'26	96.2	48.0	86.0	17.6	1.1	29.5	17.0	14.7	81.4	81.0	60.3	25.9	85.1	38.9	27.5	51.4	47.6	+14.1	54.9	+16.3
SPR+ST (Ours)	SCIS'26	**97.9**	52.6	86.9	15.7	2.1	30.0	25.3	15.8	83.6	83.1	63.7	31.2	86.7	49.5	35.3	60.6	51.2	+17.7	59.4	+20.8

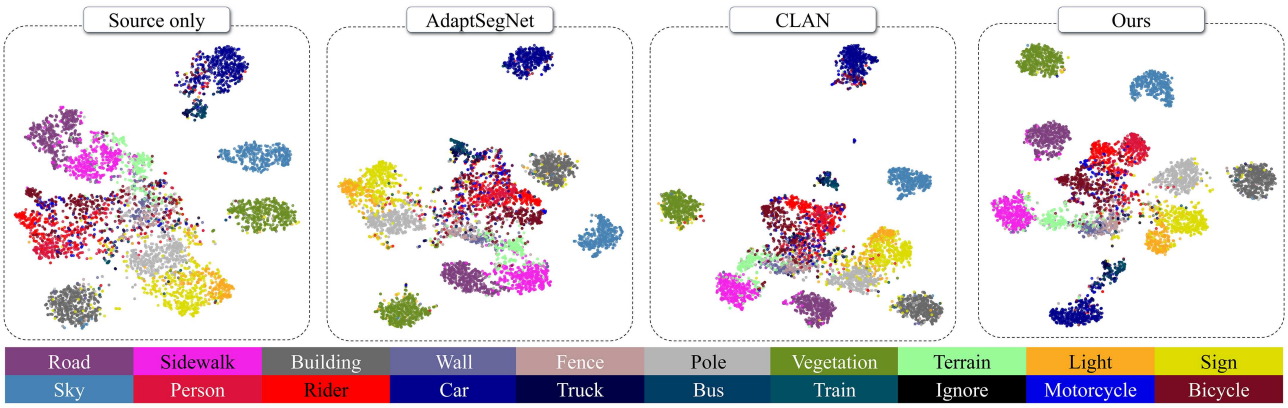

Figure 5 (Color online) Feature cluster visualization in the output space using t-SNE [72].

4.3.3 *Experimental results on Cityscapes → ACDC adaptation task*

To evaluate robustness against complex environmental variations in real-world scenarios, further experiments are conducted on the Cityscapes → ACDC adaptation task. Unlike the synthetic-to-real adaptation task, this setting involves severe visual degradation caused by adverse conditions such as fog, nighttime, rain, and snow. The quantitative and qualitative experimental results of the compared methods [23, 52, 56, 57, 65, 73–76] are presented in Table 3 and Figure 7, respectively. The proposed SPR shows superior stability under these domain shifts. Specifically, the method achieves mIoU of 49.5%, outperforming the source-only baseline by a substantial 20.7% and surpassing SoTA methods including VBLC [76] and Refign [73]. A key advantage of SPR lies in its ability to preserve semantic structure when visual cues are compromised. For instance, on large-scale background categories such as the sky and wall, which are heavily affected by lighting and weather changes, SPR maintains remarkable consistency. Notably, the method achieves 84.7% IoU on the sky and 36.1% on the wall, significantly outperforming

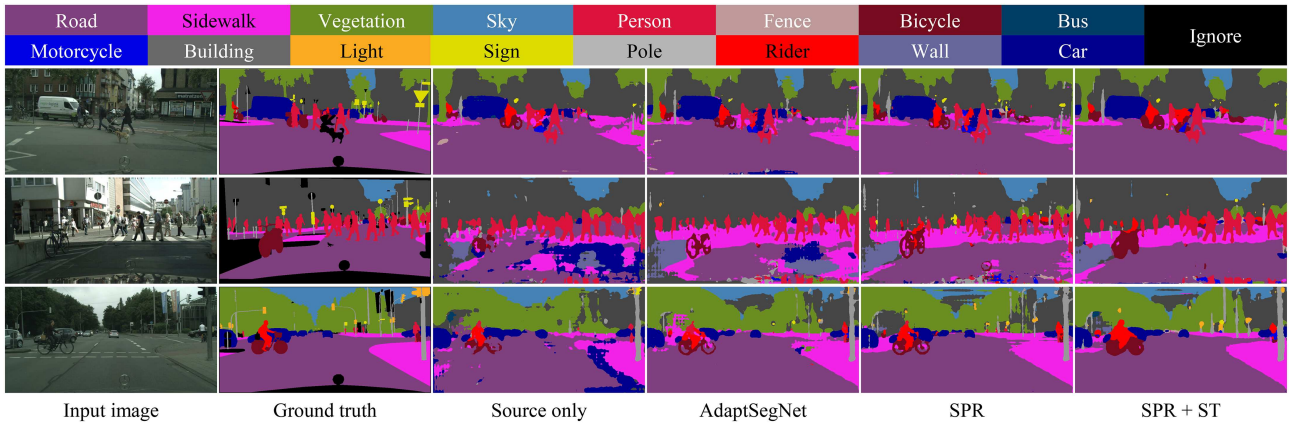

Figure 6 (Color online) Qualitative results on the SYNTHIA → Cityscapes adaptation task. From left to right, the figure displays the input target-domain image and its corresponding ground truth, followed by the segmentation results predicted by four different methods: the source-only baseline model, AdaptSegNet [23], SPR, and SPR with the self-training strategy.

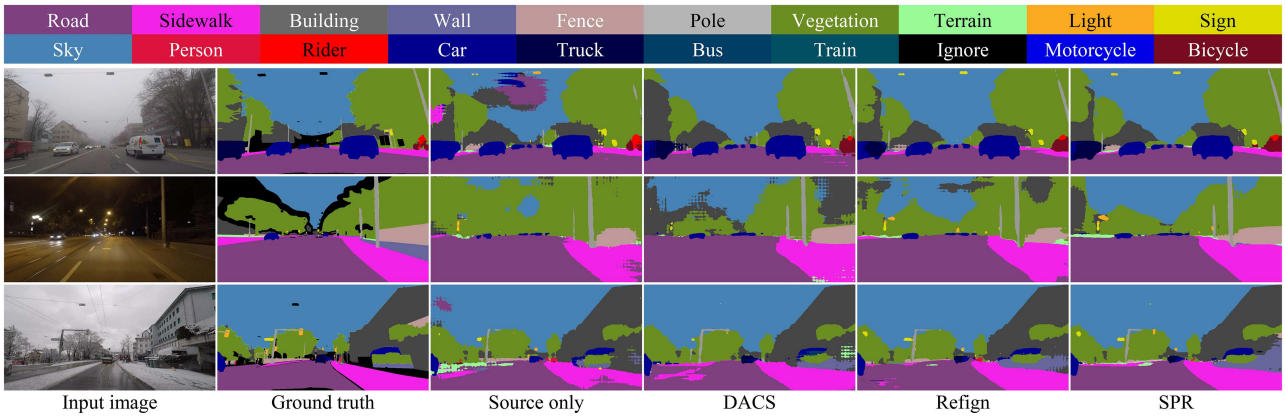

Figure 7 (Color online) Qualitative results obtained on the Cityscapes → ACDC adaptation task. From left to right, the figure displays the input target-domain image and its corresponding ground truth, followed by the segmentation results predicted by four different methods: the source-only baseline model, DACS [65], Refign [73], and SPR.

Table 3 Quantitative comparison of state-of-the-art unsupervised domain adaptation methods for Cityscapes → ACDC adaptation task based on per-class IoU (%) and overall mIoU (%). Bold values indicate the best performance.

Method	Publication	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bike	mIoU	Gain
Source Only [52]	T-PAMI'17	79.0	21.8	53.0	13.3	11.2	22.5	20.2	22.1	43.5	10.4	18.0	37.4	33.8	64.1	6.4	0.0	52.3	30.4	7.4	28.8	0.0
AdaptSeg [23]	CVPR'18	69.4	34.0	52.8	13.5	18.0	4.3	14.9	9.7	64.0	23.1	38.2	38.6	20.1	59.3	35.6	30.6	53.9	19.8	33.9	33.4	+4.6
ADVENT [56]	CVPR'19	72.9	14.3	40.5	16.6	21.2	9.3	17.4	21.2	63.8	23.8	18.3	32.6	19.5	69.5	36.2	34.5	46.2	26.9	36.1	32.7	+3.9
BDL [57]	CVPR'19	56.0	32.5	68.1	20.1	17.4	15.8	30.2	28.7	59.9	25.3	37.7	28.7	25.5	70.2	39.6	40.5	52.7	29.2	38.4	37.7	+8.9
DANNet [74]	CVPR'21	**82.9**	53.1	75.3	32.1	28.2	26.5	39.4	40.3	70.0	39.7	83.5	42.8	28.9	68.0	32.0	31.6	47.0	21.5	36.7	46.3	+17.5
DACS [65]	WACV'21	58.5	34.7	76.4	20.9	22.6	31.7	32.7	46.8	58.7	39.0	36.3	43.7	20.5	72.3	39.6	34.8	51.1	24.6	38.2	41.2	+12.4
FDA [75]	ICLR'23	73.2	34.7	59.0	24.8	**29.5**	28.6	43.3	44.9	70.1	28.2	54.7	47.0	28.5	74.6	44.8	**52.3**	63.3	28.3	39.5	45.7	+16.9
VBLC [76]	AAAI'23	49.6	39.3	79.4	35.8	**29.5**	**42.6**	**57.2**	**57.5**	69.1	**42.7**	39.8	**54.5**	**29.3**	**77.8**	43.0	36.2	32.7	**38.7**	**53.4**	47.8	+19.0
Refign [73]	WACV'23	49.5	**56.7**	**79.8**	31.2	25.7	34.1	48.0	48.7	**76.2**	42.5	38.5	48.3	24.7	75.3	**46.5**	43.9	**64.3**	34.1	43.6	48.0	+19.2
SPR (Ours)	SCIS'26	77.8	36.4	78.9	**36.1**	29.1	40.1	46.9	51.2	74.9	38.5	**84.7**	49.3	22.3	73.3	40.4	42.6	57.0	26.5	34.5	**49.5**	**+20.7**

Refign, which achieves only 38.5% and 31.2%, respectively. These results confirm that SPR mitigates the impact of realistic texture corruption, ensuring reliable scene parsing performance in the most challenging real-to-real adaptation tasks.

4.4 Ablation study

To evaluate the effectiveness of individual components, ablation experiments are conducted on the GTA5 → Cityscapes adaptation task, as detailed in Table 4. The source-only baseline, based on DeepLab-v2 with a ResNet-101 backbone, achieves mIoU of 37.1% on the Cityscapes validation set. Specifically, the structural regularization

Table 4 Ablation study of each component on GTA5 → Cityscapes adaptation task. "P-Inter" and "P-Intra" refer to prototypical inter-class and intra-class interactions, respectively; "P-Pixel" denotes prototype-pixel interaction; "L" represents pixel-wise attention for contrastive loss; and "Ada-ST" indicates adaptive threshold self-training. Bold values indicate the best performance.

Source only	P-Intra	P-Inter	P-Pixel	L	Ada-ST	mIoU (%)
✓						37.1
✓	✓					51.5
✓		✓				49.7
✓	✓	✓				52.2
✓			✓	✓		51.6
✓	✓	✓	✓			52.7
✓	✓	✓	✓	✓		53.1
✓					✓	45.5
✓	✓	✓	✓	✓	✓	**54.4**

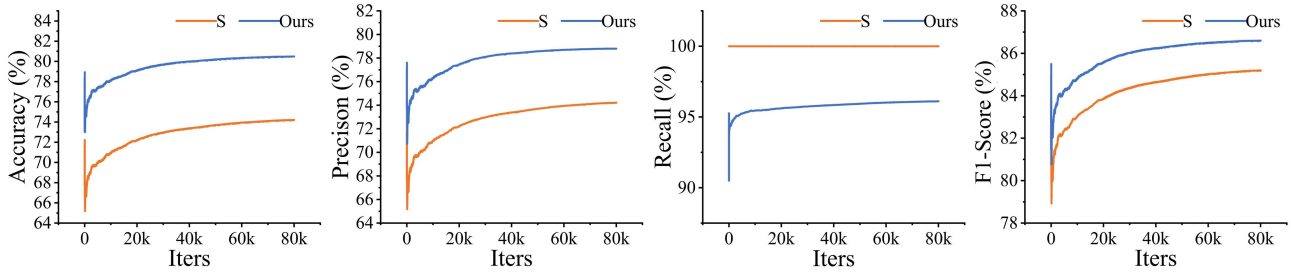

Figure 8 (Color online) Performance comparison of pseudo label filtering strategies. "S" represents the softmax-based pseudo label filtering method with fixed probability thresholds.

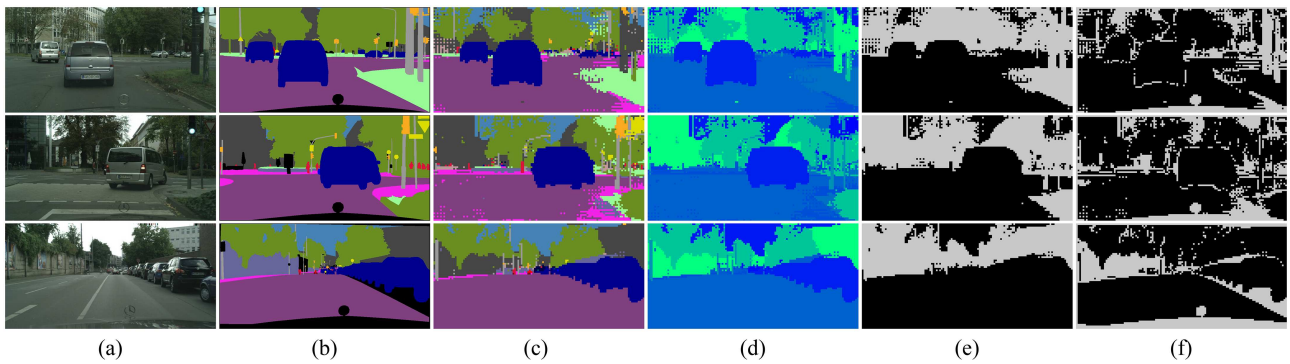

Figure 9 (Color online) Visualization of pseudo label filtering. (a) Input target-domain images; (b) ground truth; (c) segmentation results of SPR; (d) entropy maps of SPR segmentation results; (e) mask for pseudo label filtering; (f) ground truth mask for pseudo label filtering.

Table 5 Sensitivity analysis of the hyperparameter α for partitioning target-domain predictions into reliable and unreliable subsets for pseudo label filtering. Bold values indicate the best performance.

GTA5 → Cityscapes									
α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
mIoU (%)	51.8	52.2	52.1	51.9	51.0	52.4	52.5	**54.4**	52.7

mechanisms serve as the primary drivers of performance improvement. Independent applications of prototypical intra-class interaction, prototypical inter-class interaction, and prototype-pixel interaction each yield substantial gains of over 12%, demonstrating that enforcing semantic consistency at both the prototype and pixel levels is critical for bridging the domain gap. The integration of these complementary modules further refines feature alignment, establishing a robust foundation and achieving mIoU of 53.1%. Notably, the framework exhibits strong potential for further optimization via self-training. While applying self-training to the source-only baseline yields 45.5% mIoU, coupling it with the proposed SPR further boosts performance to 54.4%. This improvement confirms that the structurally regularized feature distribution facilitates effective self-training on the target domain.

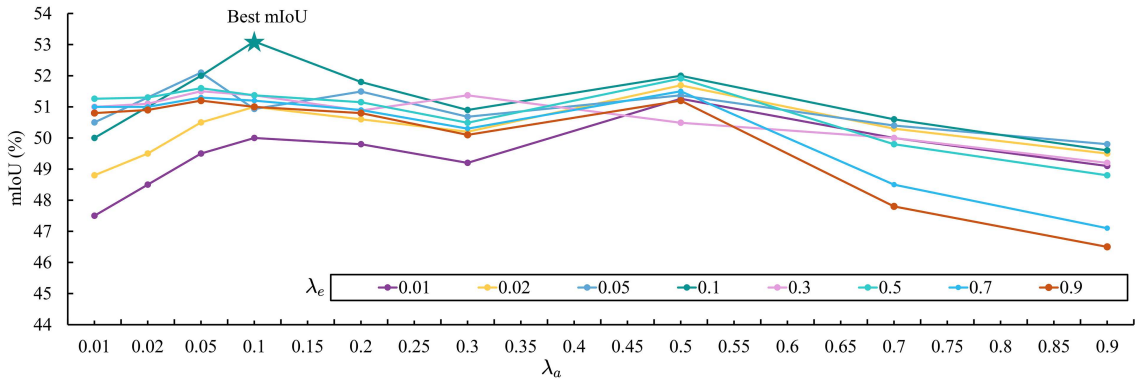

Figure 10 (Color online) Analysis for the hyperparameters λ_e and λ_a of the proposed structured prototype regularization.

4.5 Parameter sensitivity analysis

The performance of the proposed SPR is influenced by two critical sets of hyperparameters: (1) the pseudo label generation threshold α, which determines the selection of reliable pixels during the filtering process, and (2) the regularization weights λ_a and λ_e, which modulate the intensity of intra-class and inter-class structural prototype regularization terms. A detailed analysis of their respective effects is presented as follows.

Pseudo label threshold: To validate the effectiveness of prototype-pixel interaction for pseudo label generation, a comprehensive evaluation is conducted. In contrast to conventional methods that rely on softmax logits [21], the proposed SPR leverages similarity between features and prototypes to filter high-confidence pixels. This selection process is controlled by the threshold parameter α, which determines the proportion of reliable pixels retained. As evidenced by the sensitivity analysis in Table 5, an optimal mIoU of 54.4% is achieved when α is set to 0.8, representing the most effective balance between label quantity and quality. Furthermore, the quality of generated pseudo labels is evaluated against ground truth labels using standard classification metrics. As illustrated in Figure 8, the comparison with the logit-based baseline [21] reveals a distinct performance trade-off. While the reference method exhibits higher recall, this suggests the inclusion of noisier labels. In contrast, the proposed SPR significantly enhances precision and accuracy, ultimately surpassing the baseline in F1-score. This result indicates that a cleaner and more accurate label set contributes more effectively to adaptation than a larger yet noisier one, further suggesting the superior capability of SPR in filtering ambiguous regions. To qualitatively illustrate the effect of pseudo label filtering, Figure 9 shows that the proposed strategy preserves reliable regions while suppressing uncertain predictions, producing high-confidence pseudo labels closely aligned with the ground truth masks.

Structural regularization weights: The impact of intra-class and inter-class regularization weights, λ_a and λ_e, is further analyzed on the GTA5 $\rightarrow$ Cityscapes adaptation task, as illustrated in Figure 10. Optimal performance is achieved with moderate weights ($\lambda_a = \lambda_e = 0.1$). This configuration provides balanced structural constraints that enhance both inter-class separability and intra-class compactness, while preserving the discriminative capacity of the prototypes. Consequently, distinct class boundaries are maintained while allowing sufficient flexibility for target-specific variations, thereby ensuring reliable feature alignment. In scenarios with minimal regularization ($\lambda_a = \lambda_e = 0.01$), structural guidance becomes negligible. Without sufficient regularization, the prototypes act merely as source-domain class centroids, making the adaptation process equivalent to a standard prototype-based approach. This functional similarity results in performance comparable to that of the baseline [21], thereby underscoring the critical role of the proposed SPR. Conversely, excessive weights ($\lambda_a = \lambda_e = 0.9$) impose overly rigid constraints on the prototypes. Such over-regularization restricts representational capacity and degrades adaptation performance, highlighting the necessity of balanced regularization. These results indicate that moderate structural constraints optimize adaptation by balancing inter-class separability and intra-class compactness effectively. Furthermore, the stable performance observed across a broad intermediate range (0.05–0.3) demonstrates the framework's robustness to hyperparameter variations.

4.6 Computational costs and runtime performance

Computational efficiency is crucial for driving scene parsing models, particularly in real-world autonomous driving scenarios. To provide a comprehensive evaluation, computational costs and runtime performance are analyzed for both training and inference phases. Experiments are conducted on the GTA5 [49] $\rightarrow$ Cityscapes [10] adaptation task. The training process is restricted to 20 epochs, comprising 2975 images. The batch size is set to 1, with input

Table 6 Comparison of training efficiency and adaptation performance for different unsupervised domain adaptation methods. Bold values indicate the best performance.

Method	Epoch duration (s) ↓	Each iter (s) ↓	Memory (MB) ↓	mIoU (%) ↑
AdaptSeg [23]	1561	0.525	11251	42.4
UIDA [58]	1560	0.525	10277	46.3
FADA [24]	2182	0.734	**5218**	49.2
ASA [64]	1570	0.520	12855	45.1
DACS [65]	**1147**	**0.380**	11483	52.1
MetaCorrect [70]	15358	5.163	19679	52.1
DAST [67]	2181	0.733	11685	49.6
ProDA [22]	2582	0.868	11469	53.7
ProCA [21]	1254	0.422	21640	48.8
SPR (Ours)	1332	0.448	17825	**54.4**

Table 7 Comparison of inference efficiency and segmentation performance for different segmentation backbones. Bold values indicate the best performance.

Family	Backbone	Params (M) ↓	FLOPs (G) ↓	FPS ↑	Latency (ms) ↓	Memory (MB) ↓	mIoU (%) ↑
ResNet [53]	ResNet-18	**11.5**	101.6	82.0	12.2	**144**	35.6
	ResNet-34	21.6	184.7	45.5	22.0	260	38.3
	ResNet-50	24.9	211.9	39.4	25.3	298	46.4
	ResNet-101	43.9	367.9	23.0	43.4	518	54.4
	ResNet-152	59.5	496.4	17.0	58.9	698	**55.0**
VGG [77]	VGG11	19.6	**77.8**	**120.9**	**8.2**	321	37.3
	VGG13	23.3	112.5	83.9	11.9	354	39.7
	VGG16	29.5	190.7	60.7	16.4	428	46.5
	VGG19	34.8	234.2	51.0	19.5	436	48.6

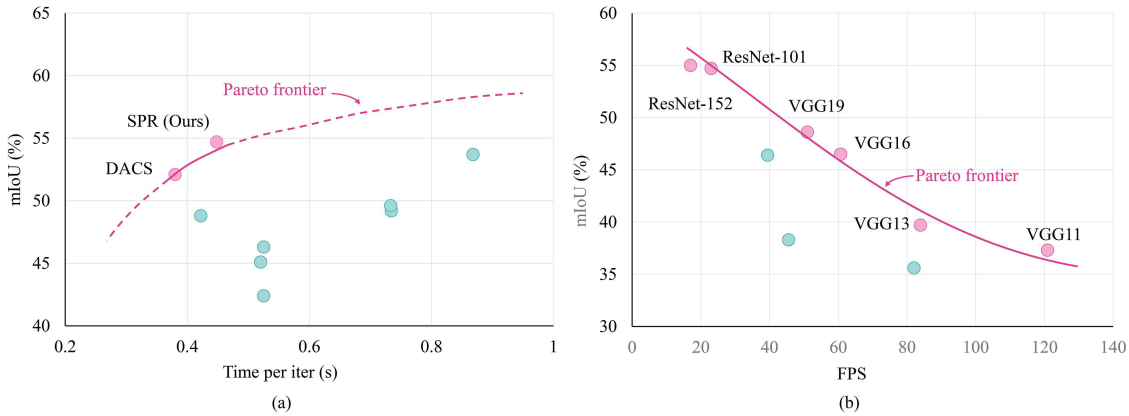

Figure 11 (Color online) Analysis of the trade-off between performance and runtime. (a) Performance and training efficiency across UDA methods; (b) performance and inference efficiency across segmentation backbones. The Pareto frontier highlights non-dominated methods. The proposed method lies on the frontier and exhibits a favorable trade-off.

images resized to 512×1024 pixels. Table 6 details the epoch duration, iteration time, memory usage, and mIoU for representative UDA methods [21–24, 58, 64, 65, 67, 70]. Specifically, compared with ProDA [22], SPR reduces the training time by nearly 50% while achieving a 0.7% improvement in mIoU. Although it is slightly slower than DACS [65], SPR yields a substantial performance gain of 2.3%. These results indicate that the proposed SPR method introduces only moderate training complexity relative to other SoTA methods implemented on the same driving scene parsing model. Consequently, SPR achieves a superior trade-off between computational efficiency and segmentation accuracy, which is advantageous for practical deployment in autonomous driving systems.

During the inference phase, the computational complexity is primarily determined by the backbone architecture, as the proposed SPR method does not introduce any additional parameters or modules into the inference pipeline. Table 7 summarizes the FPS, latency, memory usage, and mIoU across representative backbones [53, 77]. These metrics serve as key indicators of deployment feasibility. The results demonstrate that SPR maintains a favorable

balance between accuracy and efficiency. For example, with a ResNet-101 backbone, SPR achieves 54.4% mIoU, 23 FPS, and 43.4 ms latency. These metrics confirm the capability of the proposed SPR method for real-time application. Consequently, this demonstrates that SPR significantly enhances segmentation accuracy without compromising inference efficiency, thereby effectively balancing performance improvements and computational overhead. Furthermore, adopting lightweight backbones such as ResNet-18 and VGG11 significantly increases inference speed with only a moderate drop in accuracy, offering flexible solutions for resource-constrained driving scenarios.

Based on the quantitative results reported in Tables 6 and 7, Figure 11 further visualizes the trade-off between segmentation performance and computational efficiency during both training and inference. As shown, the proposed SPR lies on the Pareto frontier, indicating that it achieves a competitive balance between accuracy and runtime without being dominated by alternative methods.

5 Discussion

5.1 Theoretical analysis

Domain shift arises from the discrepancy between the joint probability distributions over images and labels in the source and target domains, i.e., $P^s(\mathcal{X}^s, \mathcal{Y}^s) \neq P^t(\mathcal{X}^t, \mathcal{Y}^t)$, resulting in degraded performance when a source-domain scene parsing model is directly applied to the target domain. Specifically, UDA aims to learn a hypothesis h that minimizes the target-domain prediction error $\epsilon^t(h)$, even in the absence of target-domain annotations. Conventionally, this is achieved by aligning source- and target-domain features or model outputs so that the probability distributions, $P^s(h(\mathcal{X}^s) \mid \mathcal{X}^s)$ and $P^t(h(\mathcal{X}^t) \mid \mathcal{X}^t)$, become similar across domains. Conventional UDA methods mainly reduce the marginal gap, i.e., $P^s(\mathcal{X}^s) \approx P^t(\mathcal{X}^t)$, but do not ensure the alignment of conditional probability distributions $P^s(\mathcal{Y}^s \mid \mathcal{X}^s)$ and $P^t(\mathcal{Y}^t \mid \mathcal{X}^t)$, which often results in high intra-class variability and low inter-class separability. The proposed SPR method mitigates this by employing structured regularization to explicitly preserve the structure of class-specific features in the target domain, which effectively promotes inter-class separability and intra-class compactness. Let class-specific prototypes be denoted by $\boldsymbol{p}_c^s$ and $\boldsymbol{p}_c^t$, and the method enforces:

$$Corr(\{\boldsymbol{p}_c^s\}_{c=1}^C) \approx Corr(\{\boldsymbol{p}_c^t\}_{c=1}^C), \tag{18}$$

where $Corr(\cdot)$ denotes the calculation of the correlation matrix. This preserves inter-class separability and intra-class compactness, guiding $P^t(\mathcal{Y}^t \mid \mathcal{X}^t)$ to approximate $P^s(\mathcal{Y}^s \mid \mathcal{X}^s)$ in a class-aware manner. Leveraging the theoretical framework for domain adaptation by Ben-David et al. [78], the expected prediction error on the target domain for a hypothesis h is bounded as follows:

$$\epsilon^t(h) \leqslant \epsilon^s(h) + \frac{1}{2} d(P^s, P^t) + \lambda, \tag{19}$$

where $\epsilon^s(h)$ denotes the expected prediction error of h on the source domain, which is accessible during training, and $\epsilon^t(h)$ denotes the expected prediction error on the target domain, which remains unknown due to the lack of target-domain annotations. The term $d(P^s, P^t)$ quantifies the divergence between the source- and target-domain marginal feature distributions, while λ quantifies the error arising from misaligned class-conditional distributions $P^s(\mathcal{Y}^s \mid \mathcal{X}^s)$ and $P^t(\mathcal{Y}^t \mid \mathcal{X}^t)$. Conventional UDA methods mainly focus on reducing $d(P^s, P^t)$ but have limited control over λ, which tends to remain substantial when intra-class compactness and inter-class separability are compromised. By enhancing intra-class compactness and enforcing consistent inter-class correlations across domains, SPR effectively reduces λ while maintaining a low $d(P^s, P^t)$, thereby tightening the theoretical bound on $\epsilon^t(h)$ and enabling more reliable adaptation.

5.2 Limitations and prospective extensions

5.2.1 *Computational complexity and efficiency trade-offs*

Experiments demonstrate that the proposed method maintains computational efficiency. In terms of training cost, its performance is comparable to existing UDA methods. Notably, the domain adaptation strategy incurs no additional computational overhead during inference. This efficiency enables the final model to satisfy deployment requirements in real time. However, the prototype interaction mechanism, which computes correlations between matrices, results in a slight increase in training time compared with conventional methods. Consequently, strategies to simplify these computations are explored, and additional experiments are conducted to evaluate their impact on both performance and efficiency. Specifically, a decoupling strategy is adopted to mitigate the computational

Table 8 Comparison of training efficiency and adaptation performance for the original and simplified SPR framework.

Method	Epoch duration (s) ↓	Each iter (s) ↓	Memory (MB) ↓	mIoU (%) ↑
SPR (Simplified)	1025	0.344	13536	47.9
SPR (Original)	1332	0.448	17825	54.4

Table 9 Quantitative comparison with state-of-the-art source-free unsupervised domain adaptation methods on the GTA5 → Cityscapes adaptation task based on per-class IoU (%) and overall mIoU (%). Bold values indicate the best performance.

Method	Publication	Source-free	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bike	mIoU
SFDA [79]	CVPR'21	✓	84.2	39.2	82.7	27.5	22.1	25.9	31.1	21.9	82.4	30.5	85.3	58.7	22.1	80.0	33.1	31.5	3.6	27.8	30.6	43.2
SFUDA [80]	ACM MM'21	✓	**95.2**	40.6	85.2	30.6	26.1	35.8	34.7	32.8	85.3	41.7	79.5	61.0	28.2	**86.5**	41.2	45.3	**15.6**	33.1	40.0	49.4
SimT [81]	CVPR'22	✓	88.5	45.4	85.2	38.4	28.0	34.9	40.8	37.0	84.9	42.4	80.4	60.6	26.2	85.6	38.0	48.0	0.0	35.1	35.9	49.2
CSFDA [82]	CVPR'23	✓	90.4	42.2	83.2	34.0	29.3	34.5	36.1	38.4	84.0	43.0	75.6	60.2	28.4	85.2	33.1	46.4	3.5	28.2	44.8	48.3
ATP [83]	T-PAMI'24	✓	93.2	**55.8**	86.5	45.2	27.3	36.6	42.8	37.9	86.0	43.1	**87.9**	**63.5**	15.3	85.5	41.2	55.7	0.0	38.1	**57.4**	52.6
DBC [84]	T-MM'24	✓	89.3	38.7	85.7	34.1	28.5	**39.1**	**43.5**	**44.4**	86.2	36.0	84.5	60.2	25.2	84.0	38.2	49.5	6.5	32.2	45.9	50.1
IAPC [85]	T-IV'24	✓	90.9	36.5	84.4	36.1	**31.3**	32.9	39.9	38.7	84.3	38.6	87.5	58.6	28.8	84.3	33.8	49.5	0.0	34.1	47.6	49.4
SPR (Ours)	SCIS'26	✓	86.0	25.6	81.9	39.5	23.2	19.7	29.6	17.3	83.9	36.0	78.9	57.6	31.2	75.6	31.9	37.2	0.2	26.2	42.6	43.4
SPR (Ours)	SCIS'26	✗	90.4	40.7	**91.5**	**47.1**	30.8	32.8	41.0	33.2	**91.1**	**58.9**	87.7	61.3	**35.0**	79.1	**51.5**	**65.6**	0.1	**43.7**	52.6	**54.4**

overhead associated with calculating and storing multiple correlation matrices. The correlation matrices for classes and channels, $\boldsymbol{R}_e^l \in \mathbb{R}^{C\times C}$ and $\boldsymbol{R}_a^l \in \mathbb{R}^{D\times D}$, are derived directly from the global class-specific prototypes $\boldsymbol{P} \in \mathbb{R}^{D\times C}$. This modification reduces the space complexity from $\mathcal{O}(D \cdot C^2 + C \cdot D^2)$ to $\mathcal{O}(C^2 + D^2)$, eliminating redundant computations in high dimensions.

Comparative results are presented in Table 8. While the simplified version effectively reduces memory usage, a performance gap in mIoU is observed when compared to the original design. This performance gap can be attributed to two structural distinctions. (1) The simplified method relies on a single global correlation matrix shared across all classes, whereas distinct semantic classes exhibit unique feature dependencies. In contrast, the proposed SPR preserves interaction matrices for individual classes, enabling the model to capture unique structural correlations for each category. (2) The results indicate that the higher dimensionality of the proposed SPR is not redundant; rather, it plays a critical role in preserving fine-grained semantic structure and ensuring superior performance. Current observations confirm that the interaction in high dimensions is essential for optimal performance. Nevertheless, there is still potential for further optimization of the proposed method. Consequently, future research will explore more efficient approximation techniques to reduce memory usage during training, with the objective of maintaining robust correlation modeling for class-specific features.

5.2.2 *Dependency on source-domain annotations and comparison with source-free UDA methods*

The primary objective of this study is to address challenges in domain adaptation from synthetic to real environments, where labeled source-domain data are available during training. To further assess the impact of this dependency, a comprehensive comparison with representative source-free unsupervised domain adaptation (SFUDA) methods is conducted on the GTA5 → Cityscapes adaptation task. As presented in Table 9, the proposed SPR leverages a labeled dataset from the source domain to achieve superior performance compared to representative SFUDA methods [79–85]. Additionally, to simulate a source-free scenario, supervision from source-domain data is disabled within the framework. Although SPR is not specifically developed for the SFUDA setting, it achieves performance comparable to several representative SFUDA methods. However, a noticeable performance gap persists when compared to leading SFUDA methods. This indicates that the naive removal of supervision from the source-domain data is insufficient to preserve structural consistency. Consequently, extending SPR to settings without source-domain data would require the integration of self-supervised objectives to compensate for the absence of explicit cross-domain supervision. Accordingly, investigating efficient self-training strategies to bridge this gap remains a priority for future research. The SPR is presented as a framework for synthetic-to-real domain adaptation in driving scene parsing, offering a favorable balance between performance and efficiency.

6 Conclusion and future work

In this article, a novel UDA framework was proposed for driving scene parsing in synthetic-to-real scenarios, which leveraged contrastive learning to explicitly model the structural relations within the semantic feature space. The proposed method constructed and regularized class prototypes and their correlation matrices, effectively promoting

a well-structured semantic feature space and thereby facilitating more precise domain alignment. Specifically, the proposed approach introduced three key improvements: (1) explicit quantification and regularization of inter-class and intra-class relations within the feature space, (2) integration of an entropy-based noise filtering mechanism to enhance the quality of pseudo labels in the target domain, and (3) refinement of pixel-level feature alignment through a pixel-level attention mechanism embedded in the contrastive learning process. Extensive experiments on multiple driving scene parsing benchmarks demonstrated the effectiveness of the proposed method compared to state-of-the-art UDA techniques. Future work will involve extending the proposed framework to the source-free setting, where source data are inaccessible for prototype initialization. To address this constraint, strategies for adapting structural regularization and alignment mechanisms will be explored, potentially by leveraging target-domain statistics or self-supervised learning.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62473288, 62388101, 62333017, 62233013), Fundamental Research Funds for the Central Universities, and Xiaomi Young Talents Program.

References

1 Huang J, Li J, Jia N, et al. RoadFormer+: delivering RGB-X scene parsing through scale-aware information decoupling and advanced heterogeneous feature fusion. IEEE Trans Intell Veh, 2025, 10: 3156–3165
2 Xue B, Yan X, Wu J, et al. Visual-marker-based localization for flat-variation scene. IEEE Trans Instrum Meas, 2024, 73: 1–16
3 Guan H, Song C, Zhang Z. LiDAR-camera cooperative semantic segmentation. Mach Intell Res, 2025, 22: 956–968
4 Wu X J, Liu C L. Hierarchical open-vocabulary part-object segmentation with knowledge-guided SAM. Mach Intell Res, 2026, 23: 214–226
5 Chen S, Jian Z, Huang Y, et al. Autonomous driving: cognitive construction and situation understanding. Sci China Inf Sci, 2019, 62: 81101
6 Ruan Y, Wang D, Yuan Y, et al. SKPNet: snake KAN perceive bridge cracks through semantic segmentation. Intell Robot, 2025, 5: 105–18
7 Fan R, Wang H, Cai P, et al. Learning collision-free space detection from stereo images: homography matrix brings better data augmentation. IEEE ASME Trans Mechatron, 2022, 27: 225–233
8 Guo S, Long Z, Wu Z, et al. LIX: implicitly infusing spatial geometric prior knowledge into visual semantic segmentation for autonomous driving. IEEE Trans Image Process, 2025, 34: 7250–7263
9 Tang G, Wu Z, Li J, et al. TiCoSS: tightening the coupling between semantic segmentation and stereo matching within a joint learning framework. IEEE Trans Automat Sci Eng, 2025, 22: 18646–18658
10 Cordts M, Omran M, Ramos S, et al. The Cityscapes dataset for semantic urban scene understanding. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 3213–3223
11 Sakaridis C, Dai D, Van Gool L. ACDC: the adverse conditions dataset with correspondences for semantic driving scene understanding. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 10765–10775
12 Ros G, Sellart L, Materzynska J, et al. The SYNTHIA dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 3234–3243
13 Marsden R A, Bartler A, Döbler M, et al. Contrastive learning and self-training for unsupervised domain adaptation in semantic segmentation. In: Proceedings of International Joint Conference on Neural Networks (IJCNN), 2022. 1–8
14 Sun H, Li M. Enhancing unsupervised domain adaptation by exploiting the conceptual consistency of multiple self-supervised tasks. Sci China Inf Sci, 2023, 66: 142101
15 Zhai Y M, Ren C X, Luo Y W, et al. Maximizing conditional independence for unsupervised domain adaptation. Sci China Inf Sci, 2024, 67: 152108
16 Ma S, Pang Y W, Pan J, et al. Preserving details in semantics-aware context for scene parsing. Sci China Inf Sci, 2020, 63: 120106
17 Li D Y, Liu M, Zhao F, et al. Challenges and countermeasures of interaction in autonomous vehicles. Sci China Inf Sci, 2019, 62: 50201
18 Zheng A, Fei Z, Ding Y, et al. RULER: source-free domain adaptive person re-identification via uncertain label refinery. Mach Intell Res, 2025, 22: 900–916
19 Li D, Zhang Z, Zhang Y, et al. AdaGPAR: generalizable pedestrian attribute recognition via test-time adaptation. Mach Intell Res, 2025, 22: 783–796
20 Lin R, Xiong A, He Q, et al. From uni- to multi-modal: progressive multi-source domain adaptation with gradient disentangling for audio-visual deception detection. Mach Intell Res, 2026, 23: 484–500
21 Jiang Z, Li Y, Yang C, et al. Prototypical contrast adaptation for domain adaptive semantic segmentation. In: Proceedings of European Conference on Computer Vision (ECCV), 2022. 36–54
22 Zhang P, Zhang B, Zhang T, et al. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 12414–12424
23 Tsai Y H, Hung W C, Schulter S, et al. Learning to adapt structured output space for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 7472–7481
24 Wang H, Shen T, Zhang W, et al. Classes matter: a fine-grained adversarial approach to cross-domain semantic segmentation. In: Proceedings of European Conference on Computer Vision (ECCV), 2020. 642–659
25 Fan R, Wang H, Cai P, et al. SNE-RoadSeg: incorporating surface normal information into semantic segmentation for accurate freespace detection. In: Proceedings of European Conference on Computer Vision (ECCV), 2020. 340–356
26 Li J, Zhang Y, Yun P, et al. RoadFormer: Duplex transformer for RGB-normal semantic road scene parsing. IEEE Trans Intell Veh, 2024, 9: 5163–5172
27 Wu Z, Feng Y, Liu C W, et al. S^3M-Net: joint learning of semantic segmentation and stereo matching for autonomous driving. IEEE Trans Intell Veh, 2024, 9: 3940–3951
28 Liu C, Li S, Chen H, et al. Semi-supervised joint adaptation transfer network with conditional adversarial learning for rotary machine fault diagnosis. Intell Robot, 2023, 3: 131–44
29 Wang G, Li Z, Weng G, et al. An overview of industrial image segmentation using deep learning models. Intell Robot, 2025, 5: 143–80
30 Chen Y, Ge P, Wang G, et al. An overview of intelligent image segmentation using active contour models. Intell Robot, 2023, 3: 23–55
31 Wu D, Liang Z, Chen G. Deep learning for LiDAR-only and LiDAR-fusion 3D perception: a survey. Intell Robot, 2022, 2: 105–129
32 Ren Q, Mao Q, Lu S. Prototypical bidirectional adaptation and learning for cross-domain semantic segmentation. IEEE Trans Multimedia, 2024, 26: 501–513
33 Hoffman J, Tzeng E, Park T, et al. CyCADA: cycle-consistent adversarial domain adaptation. In: Proceedings of International Conference on Machine Learning (ICML), 2018. 1989–1998
34 Hoffman J, Wang D, Yu F, et al. FCNs in the wild: pixel-level adversarial and constraint-based adaptation. 2016. ArXiv:1612.02649
35 Zou Y, Yu Z, Kumar B, et al. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of European Conference on Computer Vision (ECCV), 2018. 297–313

36 Luo Y, Liu P, Zheng L, et al. Category-level adversarial adaptation for semantic segmentation using purified features. IEEE Trans Pattern Anal Mach Intell, 2022, 44: 3940–3956
37 Long M, Cao Y, Wang J, et al. Learning transferable features with deep adaptation networks. In: Proceedings of International Conference on Machine Learning (ICML), 2015. 97–105
38 Mei K, Zhu C, Zou J, et al. Instance adaptive self-training for unsupervised domain adaptation. In: Proceedings of European Conference on Computer Vision (ECCV), 2020. 415–430
39 Cheng Y, Wei F, Bao J, et al. ADPL: adaptive dual path learning for domain adaptation of semantic segmentation. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 9339–9356
40 Truong T D, Le N, Raj B, et al. FREDOM: fairness domain adaptation approach to semantic scene understanding. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 19988–19997
41 Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006. 1735–1742
42 He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 9726–9735
43 Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations. In: Proceedings of International Conference on Machine Learning (ICML), 2020. 1597–1607
44 Jiang Z, Gu Z, Peng J, et al. STC: spatio-temporal contrastive learning for video instance segmentation. In: Proceedings of European Conference on Computer Vision (ECCV), 2023. 539–556
45 Alonso I, Sabater A, Ferstl D, et al. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 8219–8228
46 Li J, Wang Z, Gao Y, et al. Exploring high-quality target domain information for unsupervised domain adaptive semantic segmentation. In: Proceedings of ACM International Conference on Multimedia (ACM MM), 2022. 5237–5245
47 Xie B, Li S, Li M, et al. SePiCo: semantic-guided pixel contrast for domain adaptive semantic segmentation. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 9004–9021
48 Xue Y, Tian X, Zhang F, et al. CACP: Covariance-aware cross-domain prototypes for domain adaptive semantic segmentation. IEEE Trans Multimedia, 2025, 27: 5023–5034
49 Richter S R, Vineet V, Roth S, et al. Playing for data: ground truth from computer games. In: Proceedings of European Conference on Computer Vision (ECCV), 2016. 102–118
50 Cao Y, Zhang H, Lu X, et al. Adaptive refining-aggregation-separation framework for unsupervised domain adaptation semantic segmentation. IEEE Trans Circuits Syst Video Technol, 2023, 33: 3822–3832
51 Chung I, Yoo J, Kwak N. Exploiting inter-pixel correlations in unsupervised domain adaptation for semantic segmentation. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2023. 12–21
52 Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 834–848
53 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 770–778
54 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009. 248–255
55 Tsai Y H, Sohn K, Schulter S, et al. Domain adaptation for structured output via discriminative patch representations. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 2019. 1456–1465
56 Vu T H, Jain H, Bucher M, et al. ADVENT: adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 2517–2526
57 Li Y, Yuan L, Vasconcelos N. Bidirectional learning for domain adaptation of semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 6929–6938
58 Pan F, Shin I, Rameau F, et al. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 3763–3772
59 Kim M, Byun H. Learning texture invariant representation for domain adaptation of semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 12972–12981
60 Lv F, Liang T, Chen X, et al. Cross-domain semantic segmentation via domain-invariant interactive relation transfer. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 4333–4342
61 Subhani M N, Ali M. Learning from scale-invariant examples for domain adaptation in semantic segmentation. In: Proceedings of European Conference on Computer Vision (ECCV), 2020. 290–306
62 Paul S, Tsai Y H, Schulter S, et al. Domain adaptive semantic segmentation using weak labels. In: Proceedings of European Conference on Computer Vision (ECCV), 2020. 571–587
63 Huang J, Lu S, Guan D, et al. Contextual-relation consistent domain adaptation for semantic segmentation. In: Proceedings of European Conference on Computer Vision (ECCV), 2020. 705–722
64 Zhou W, Wang Y, Chu J, et al. Affinity space adaptation for semantic segmentation across domains. IEEE Trans Image Process, 2021, 30: 2549–2561
65 Tranheden W, Olsson V, Pinto J, et al. DACS: domain adaptation via cross-domain mixed sampling. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2021. 1378–1388
66 Zheng Z, Yang Y. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. Int J Comput Vis, 2021, 129: 1106–1120
67 Yu F, Zhang M, Dong H, et al. DAST: unsupervised domain adaptation in semantic segmentation based on discriminator attention and self-training. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI), 2021. 10754–10762
68 Lee S, Hyun J, Seong H, et al. Unsupervised domain adaptation for semantic segmentation by content transfer. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI), 2021. 8306–8315
69 Gao L, Zhang L, Zhang Q. Addressing domain gap via content invariant representation for semantic segmentation. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI), 2021. 7528–7536
70 Guo X, Yang C, Li B, et al. MetaCorrection: domain-aware meta loss correction for unsupervised domain adaptation in semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 3927–3936
71 Wang Y, Peng J, Zhang Z. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 9092–9101
72 Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res, 2008, 9: 2579–2605
73 Brüggemann D, Sakaridis C, Truong P, et al. Refign: align and refine for adaptation of semantic segmentation to adverse conditions. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2023. 3173–3183
74 Wu X, Wu Z, Ju L, et al. A one-stage domain adaptation network with image alignment for unsupervised nighttime semantic segmentation. IEEE Trans Pattern Anal Mach Intell, 2023, 45: 58–72
75 Yi L, Xu G, Xu P, et al. When source-free domain adaptation meets learning with noisy labels. 2023. ArXiv:2301.13381
76 Li M, Xie B, Li S, et al. VBLC: visibility boosting and logit-constraint learning for domain adaptive semantic segmentation under adverse conditions. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI), 2023. 8605–8613
77 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556

78 Ben-David S, Blitzer J, Crammer K, et al. A theory of learning from different domains. Mach Learn, 2010, 79: 151–175
79 Liu Y, Zhang W, Wang J. Source-free domain adaptation for semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 1215-1224
80 Ye M, Zhang J, Ouyang J, et al. Source data-free unsupervised domain adaptation for semantic segmentation. In: Proceedings of ACM International Conference on Multimedia (ACM MM), 2021. 2233–2242
81 Guo X, Liu J, Liu T, et al. SimT: handling open-set noise for domain adaptive semantic segmentation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 7032–7041
82 Karim N, Mithun N C, Rajvanshi A, et al. C-SFDA: a curriculum learning aided self-training framework for efficient source free domain adaptation. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023. 24120–24131
83 Wang Y, Liang J, Zhang Z. A Curriculum-style self-training approach for source-free semantic segmentation. IEEE Trans Pattern Anal Mach Intell, 2024, 46: 9890–9907
84 Tian Y, Li J, Fu H, et al. Self-mining the confident prototypes for source-free unsupervised domain adaptation in image segmentation. IEEE Trans Multimedia, 2024, 26: 7709–7720
85 Cao Y, Zhang H, Lu X, et al. Towards source-free domain adaptive semantic segmentation via importance-aware and prototype-contrast learning. IEEE Trans Intell Veh, 2025, 10: 2736–2747