Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.20260400406

Research on Cross-modal and Semantic Collaborative Unsupervised Domain Adaptation for All-weather Autonomous Driving Perception Enhancement

Zhuochao Du*

Portland institute, Nanjing University of Posts and Telecommnication, Nanjing, China

**Corresponding author: Zhuochao Du.*

Abstract

Autonomous driving perception systems often encounter severe “domain shift” during real-world deployment. Although deep learning models perform well under ideal weather, their accuracy drops significantly in extreme conditions like rain, snow, fog, or low-light nighttime, as well as during sensor modality transitions. Current Unsupervised Domain Adaptation (UDA) methods mostly focus on class-agnostic global alignment, often neglecting fine-grained semantic correlations and complex visual gaps. This paper systematically explores how to use UDA to enhance the robustness of autonomous driving in complex environments. We propose a multi-dimensional framework that synergizes visual style, feature extraction, and semantic correlation. This study combines theoretical analysis with an algorithmic review to organize several cutting-edge technical paths. First, we introduce Style Prompt Tuning guided by pre-trained Vision-Language Models (VLM). This achieves style transfer while preserving image geometry. Second, we utilize the Transformer-based DTN-DETR model, which cooperates with frequency and spatial domain optimization to improve feature extraction in low-light conditions. Finally, we analyze the Graph Embedding Interclass Relation-Aware Adaptive Network (GelraA-Net) based on Graph Convolutional Networks (GCN) and the Dual Semantic Correlation Alignment (DSCA) mechanism. These methods strengthen cross-scene semantic consistency through topological structures and contextual logic.

Keywords

autonomous driving, Unsupervised Domain Adaptation, all-weather perception, cross-modal knowledge transfer, Semantic Correlation Alignment

1. Introduction

With the rapid development of artificial intelligence and sensor technology, autonomous driving has become the core driver of global intelligent transportation. However, the robustness of perception systems, including semantic segmentation and object detection, still faces severe challenges in real-world environments. In practical deployment, deep learning models trained on specific scenes often suffer significant performance degradation when applied to new target domains due to variations in collection conditions, seasons, and regional styles [1]. Specifically, for all-weather operations, systems must remain stable under adverse

weather—such as rain, snow, and fog—and low-light nighttime conditions. For instance, nighttime scenes involve not only global low illumination but also sensor noise and sudden glare, making it difficult for traditional algorithms to capture weak object contours [2].

Furthermore, the limitations of single sensors become more prominent in extreme environments. While visible-light cameras perform well in clear daylight, they may fail in total darkness or under direct strong light. In such cases, cross-modal data like thermal imaging or LiDAR must be introduced as a supplement [3]. Therefore, researching Unsupervised Domain Adaptation (UDA) to achieve cross-environment and cross-modal knowledge transfer without manual labeling is of great scientific and engineering value for enhancing autonomous driving safety.

Current UDA research for autonomous driving has made significant progress. Mainstream directions include using Generative Adversarial Networks (GANs) or style transfer for image-level distribution alignment, employing self-training and pseudo-labeling for knowledge mining in target domains [4], and utilizing the long-range dependency modeling of Transformers to enhance feature extraction [5]. Recently, the introduction of Vision-Language Models (VLM) has provided new ways to bridge visual gaps [6].

However, existing UDA studies still have several limitations. First, most methods tend to perform class-agnostic global alignment. They neglect the inherent differences in feature distributions between categories, such as pedestrians and traffic signs, leading to poor fine-grained alignment [5]. Second, in cross-modal scenarios, current algorithms often ignore global information and inherent interclass relations. This leads to category confusion when dealing with complex and heterogeneous data sources [7]. Finally, in adverse weather or occlusion scenarios, existing research lacks deep exploration of semantic consistency and spatial geometric constraints, making it difficult to achieve true all-scene perception enhancement [8].

Based on these gaps, the core research question of this paper is: How can we build a multi-dimensional UDA system that synergizes visual style, cross-modal features, and fine-grained semantic correlations for all-weather autonomous driving? This question addresses the current lack of category-level alignment, multi-modal transfer, and global logical association. This paper introduces Style Prompt Tuning to resolve visual differences [6]. We also use Graph Embedding to capture topological consistency across scenes [7] and combine Dual Semantic Correlation Alignment (DSCA) with spherical geometric constraints to improve accuracy [1].

The final objectives are twofold. At the theoretical level, this study promotes the transition of UDA from "single distribution alignment" to a "style-feature-semantic" synergistic framework. We explore how logical correlations, such as context and class correlation, can correct perception biases caused by environmental interference. At the application level, this research directly supports all-weather autonomous driving tasks. By reducing reliance on expensive labeled data, it improves the feasibility of perception algorithms in nighttime, thick fog, and cross-modal environments, providing technical insights for safer and more robust intelligent transportation systems.

2. The Evolution of Unsupervised Domain Adaptation (UDA) Paradigms

2.1 From Distribution Alignment to Structural Alignment

Unsupervised Domain Adaptation (UDA) for autonomous driving has shifted from coarse global distribution alignment to fine-grained structural alignment. Early research relied primarily on Image-to-Image (I2I) translation, such as CycleGAN, using adversarial training for style transfer between domains. However, these methods have clear limitations in complex driving scenes. Generative models often require large datasets or complex cycle-consistency constraints. Moreover, they risk altering the original geometric content during style conversion. This can cause a mismatch between semantic labels and the transformed images, reducing segmentation accuracy [6]. Furthermore, traditional global alignment ignores feature distribution differences between categories. This class-agnostic approach often leads to "negative transfer," limiting the model's generalization in complex target domains [5].

To overcome these flaws, modern paradigms emphasize protecting and utilizing deep image structures. First, researchers have introduced semantic correlations. By modeling both image-level context and instance-level category correlations, models can use logical consistency—such as "vehicles are typically on roads"—

to assist alignment. This semantic logic significantly enhances detector robustness in adverse conditions like rain, fog, or occlusion [8]. Second, graph embedding has elevated alignment from the pixel level to the topological level. By using Graph Convolutional Networks (GCN) to model inter-class relationships across scenes, models capture global distribution patterns and ensure semantic consistency [7]. Finally, Frequency Domain Alignment (FDA) has evolved for day-to-night transitions involving drastic lighting changes. By extracting light-insensitive phase information or performing dual-domain interaction, models retain structural features more effectively than spatial-domain methods alone, improving nighttime accuracy [2].

2.2 Self-training and Pseudo-labeling Strategies

Self-training and pseudo-labeling have become foundational strategies for UDA in autonomous driving. Since target domains lack manual labels, generating pseudo-labels from model predictions is a core method for knowledge transfer. In cross-modal tasks, such as RGB-to-thermal migration, this strategy effectively transfers semantic knowledge from the visible light domain to thermal perception networks [3]. To ensure reliability, researchers use mathematical tools like Shannon Entropy constraints to optimize classification boundaries. By reducing prediction uncertainty and filtering high-quality samples, they mitigate gradient drift caused by incorrect pseudo-labels [1].

Meanwhile, combining consistency regularization with self-training further stabilizes models on unlabeled data. Modern mechanisms do not rely solely on original image predictions; they introduce multiple perturbations consistency learning. By applying perturbations at both input and feature levels and enforcing consistent predictions, models are forced to mine robust spatial context features in the target domain [4]. Additionally, the development of pre-trained Vision-Language Models (VLM) has brought breakthroughs to self-training. Through Style Prompt Tuning, models optimize style transfer networks guided by automatically generated textual prompts. Unlike early generative models, this prompt-based approach does not require large-scale end-to-end training. Instead, lightweight fine-tuning allows the model to adapt to target styles independently, greatly improving deployment efficiency and content preservation in all-weather environments [6].

3. Visual Bridging and Semantic Alignment in Extreme Weather

3.1 Style Prompt Tuning Guided by Vision-Language Models

Traditional Unsupervised Domain Adaptation (UDA) methods, such as Generative Adversarial Networks (GANs) or Diffusion Models, face significant limitations when handling visual shifts caused by adverse weather like rain or fog. Research by Suyeon Cha's team indicates that these methods often require expensive paired data. Furthermore, the pixel reconstruction process easily alters the geometric content and topological structure of the images. This leads to a mismatch between the original source-domain semantic labels (such as fine object contours) and the generated target-domain images, eventually causing a sharp decline in segmentation accuracy [6].

To address this challenge, the study proposes the Style Prompt Tuning (SPT) framework. Its core innovation lies in leveraging the semantic understanding of large-scale pre-trained Vision-Language Models (VLM). Specifically, SPT uses a VLM to automatically generate textual prompts representing target-domain styles. These prompts act as guidance signals to optimize a lightweight U-Net style network. Without changing the semantic structure of the source images, this network performs asymmetric style adjustments at the pixel level to align visual features with the target domain. Experimental results show that in transferring from Cityscapes to Foggy Cityscapes and Rain Cityscapes, this method significantly improves mIoU. It also offers superior training stability compared to traditional generative models, successfully bridging the visual gap between clear days and extreme weather [6].

3.2 Target Domain Feature Mining and Spatial Context Enhancement

In extreme weather applications for autonomous driving, obtaining and labeling data for foggy scenes is particularly difficult. Traditional UDA paradigms rely too heavily on "strong supervision" from the source domain to guide target-domain learning. When comparing clear city streets with blurry, foggy environments,

the vast difference in spatial context prevents the model from mining discriminative features within the target domain. This results in issues such as blurred segmentation boundaries and the loss of small objects.

Addressing this pain point, Chaoyu Rao's team proposed an enhanced target-domain learning scheme introducing the Multiple Perturbations Consistency (MPC) strategy. This strategy posits that even without labels, the target domain possesses high spatial correlation. By applying color jittering and Gaussian noise at the input level, alongside Dropout or feature randomization at the latent level, the model is forced to maintain consistent predictions for the same location under different perturbations. This "self-supervised" process compels the model to learn robust features immune to fog interference. Additionally, this method integrates spatial weight mapping to dynamically adjust the loss contribution between domains, thereby suppressing irrelevant "negative transfer" features. This strategy demonstrates strong adaptability in unlabeled foggy environments and effectively improves object recognition accuracy in complex backgrounds [4].

3.3 Fine-grained Semantic Correlation and Structural Alignment

Beyond visual style differences, extreme weather often causes the loss of local semantic information. In such cases, pixel alignment alone is insufficient; logical correlations between objects must be introduced to assist inference.

The Dual Semantic Correlation Alignment (DSCA) method, proposed by Yinsai Guo's team, provides a new perspective. This method advocates for utilizing two types of semantic information: context correlation and class correlation. For example, when detecting a vehicle in thick fog, the system can correct misclassifications by capturing the vehicle's context relative to the "road" and its distribution relative to other vehicles, even if the vehicle's features are blurred. By establishing this dual correlation alignment mechanism at the feature level, DSCA significantly reduces false alarm rates in foggy scenes [8].

Furthermore, for broader cross-scene transfer challenges, the GelraA-Net proposed by Teng Yang's team introduces Graph Convolutional Networks (GCN) to model topological relationships between categories. By mapping image features into a graph space, the study uses graph embedding to capture global distribution patterns. This ensures that the source and target domains remain consistent in complex semantic logic. This transition from "pixel alignment" to "logical consistency" provides a theoretical foundation for the generalization of all-weather autonomous driving across different regions and climates [7].

4. Cross-modal and Feature Synergy in Nighttime Low-light Environments

4.1 Feature Extraction and Category Alignment Optimization in Day-to-Night Adaptation

In nighttime driving, image signals suffer from low signal-to-noise ratios and blurred object edges. Research by Gong's team found that traditional UDA detection methods often fail to separate key objects from extremely dark backgrounds. To address this, they proposed the DTN-DETR architecture. This method leverages the global attention mechanism of Transformers and introduces a synergistic optimization strategy for both the frequency and spatial domains. At the algorithmic level, DTN-DETR uses image enhancement to simulate day-to-night differences and establishes a photometric consistency matching loss. Its core innovation involves filtering nighttime-specific high-frequency noise in the frequency domain while capturing geometric structures in the spatial domain. Furthermore, the team optimized specific prediction branches in the detection head for small objects like distant pedestrians and vehicles. Experiments on the BDD100K nighttime test set show that this method significantly improves detection accuracy in complex traffic flows while maintaining real-time speed [2].

However, resolving visibility at the visual level is not enough. Hanguang Xiao's team noted that many traditional UDA detectors perform class-agnostic global feature alignment, which ignores the inherent differences in feature distributions between categories at night. Consequently, they proposed the UDA-DETR framework. This framework introduces an Encoder Feature Alignment (EFA) module that uses domain queries for adversarial learning to capture global distributions. Meanwhile, it incorporates Category Discriminative Alignment (CDA) to ensure that category-specific semantic details are not lost during the alignment process. This evolution from global alignment to local instance alignment allows the detector to outperform mainstream

adversarial algorithms in challenging tasks like Foggy Cityscapes, providing a guarantee for precise nighttime localization [5].

4.2 Cross-modal Knowledge Transfer: Deep Alignment from RGB to Thermal Imaging

When visible-light cameras fail in total darkness or strong glare, thermal sensors become a critical supplement due to their light-independent imaging characteristics. However, thermal datasets are small and expensive to label. Research now focuses on using mature RGB domain knowledge to assist thermal perception.

The MS-UDA framework proposed by Yeong-Hyeon Kim's team offers a systematic solution. Instead of attempting to bridge the massive spectral gap directly, this framework builds a continuous multi-spectral knowledge transfer path. It first performs RGB-to-RGB style adaptation to handle day-night differences, followed by cross-spectral RGB-to-Thermal alignment to address modality differences. Finally, it uses Thermal-to-Thermal adaptive fine-tuning for domain differences. This step-by-step strategy effectively transfers segmentation knowledge from large-scale labeled RGB datasets to thermal networks. Experiments prove that MS-UDA significantly improves the segmentation robustness for pedestrians, vehicles, and roads without requiring any target-domain labels [3].

4.3 Spherical Pre-alignment for 3D Point Clouds in Sparse Nighttime Environments

In all-weather autonomous driving, 3D geometric features from LiDAR serve as a vital defense against visual failure at night. However, cross-scene point cloud data also face severe domain shift. Distant areas are particularly affected by sparse point distributions and regional style variations.

Qingwang Wang's team proposed the PS-UDA framework for cross-scene multispectral point cloud classification. Their core contribution is the introduction of L2-norm spherical constraints and a Laplace matrix pre-alignment mechanism. They argue that features from source and target domains, which are distant in Euclidean space, are much easier to align if mapped to the same spherical surface using L2 norms. Based on this spherical mapping, the model uses a Laplace matrix to capture sample similarities and applies Shannon entropy constraints for self-training. This approach effectively protects the semantic consistency of distant objects. This mathematical pre-alignment at the 3D geometric level provides crucial spatial support for the "visual-geometric" synergistic UDA scheme proposed in this paper, ensuring reliable 3D perception in extreme nighttime environments [1].

5. Semantic Enhancement: Category Relation Modeling and Consistency Learning

5.1 Graph Embedding for Interclass Relation-Aware Modeling

In cross-scene Unsupervised Domain Adaptation (UDA) tasks, pixel-level or feature-level alignment often overlooks two critical factors: the macro contribution of global image information and the inherent interclass relations between categories. Teng Yang's team noted that without modeling logical relationships—such as the co-occurrence frequency of "buildings" and "roads"—models are prone to category confusion during alignment.

To address this, the team proposed the GelraA-Net (Graph Embedding Interclass Relation-Aware Adaptive Network). The core innovation is the use of Graph Convolutional Networks (GCN) to build category relationship graphs. During feature extraction, the system captures isolated sample features while establishing spatial and logical relationships through a Graph Embedding Module. Specifically, GelraA-Net employs manifold embedded distribution alignment to map high-dimensional features into a low-dimensional manifold space with topological consistency. This allows the model to capture global distribution patterns accurately. Experimental results show that GelraA-Net effectively compensates for local details lost in global alignment when processing complex multisource remote sensing data. Its classification accuracy outperforms conventional adversarial alignment methods on multiple public datasets [7]. This graph-based paradigm provides robust topological constraints for understanding complex urban landscapes.

5.2 Dual Semantic Correlation and 3D Geometric Constraints

In extreme environments, such as large-scale occlusion or sensor failure, systems must possess "reasoning abilities" to infer missing information using known semantics. The Dual Semantic Correlation Alignment (DSCA) method, proposed by Yinsai Guo's team, provides theoretical support for this capability.

The core logic of DSCA lies in mining both context correlation and class correlation. In autonomous driving object detection, context correlation uses environmental information to constrain the object search space. Meanwhile, class correlation utilizes co-occurrence probabilities to assist classification. By introducing these dual correlations into the UDA framework, DSCA guides the model to find semantically consistent feature anchors in the target domain. This maintains high detection accuracy even in adverse weather or severe occlusion [8].

For the more complex dimension of 3D perception, Qingwang Wang's team introduced a spherical geometric constraint mechanism within the PS-UDA framework. Because 3D point cloud density and scanning angles vary greatly across scenes, traditional Euclidean distance alignment often fails. PS-UDA introduces an L2-norm constraint to project features from both source and target domains onto the same unit sphere surface. In this spherical space, discrete feature distributions become easier to align. Combined with adjacency relationships established by a Laplace matrix, this method achieves efficient self-training optimization through Shannon entropy constraints [1].

6. Conclusion

This paper addresses the core "domain shift" problem in all-weather autonomous driving perception systems. We explored and constructed a multi-dimensional Unsupervised Domain Adaptation (UDA) framework that synergizes cross-modal features and semantics. Our findings indicate that single-level distribution alignment is no longer sufficient for the complex challenges of extreme weather and low-light environments. Instead, a coordinated approach across three dimensions—style, feature, and semantics—is necessary. Methodologically, this study organized the use of Style Prompt Tuning guided by Vision-Language Models (VLM) to eliminate visual interference from rain and fog. We also combined the global attention of Transformers with frequency-domain optimization to extract key object contours in dark nighttime conditions (Cha et al., 2025; Gong et al., 2025). Furthermore, by introducing the GelraA-Net, this research verified the importance of building category relationship graphs. This ensures topological logic stability during transitions between different geographic regions and sensor modalities [7].

Reflecting on the research process, this paper confirms that protecting image geometry and semantic logic within the UDA framework is key to robust perception. Experimental analysis shows that while traditional generative alignment bridges visual gaps, a lack of modeling for object co-occurrence and spatial layout leads to semantic label misalignment. Additionally, cross-modal knowledge transfer—such as from RGB to thermal imaging or 3D point clouds—requires more than just spectral feature matching. It also demands mathematical tools like L2-norm spherical constraints to overcome inherent differences in data distribution [1]. Although our multi-dimensional framework demonstrates significant robustness, the model still experiences latency in perceiving small objects during extreme compound conditions, such as heavy rain at night. Furthermore, semantic blurring persists when foreground objects share high physical similarity with the background [5].

The core value of this research lies in its theoretical breakthroughs and engineering potential, directly addressing the demand for all-weather safety. To resolve the lack of precision in class-agnostic global alignment, this paper implements instance-level matching through Category Discriminative Alignment (CDA) and Dual Semantic Correlation Alignment (DSCA). This fills the knowledge gap regarding local semantic alignment in complex scenes [8]. By using graph embedding to model interclass topological relations, we elevated the alignment paradigm from the pixel level to the logical level, expanding existing structural alignment theories. Moreover, this framework requires no expensive manual labeling. Its applicability spans heterogeneous scenarios, including urban streets and remote sensing, providing a systematic solution for generalizing all-weather autonomous driving algorithms.

Based on our findings, future research should incorporate temporal consistency. Using motion features from continuous video streams can enhance perception coherence in dynamic driving scenes, compensating for the limitations of single-frame alignment [5]. Additionally, exploring real-time dynamic prompt generation

will be crucial for improving system adaptability to unpredictable environmental changes [6]. Finally, a key research direction for higher-level perception enhancement involves integrating 3D spherical constraints more closely with 2D semantic logic in an end-to-end framework. This will be essential to resolve boundary blurring for distant, sparse objects [1].

References

[1] Wang, Q., Wang, M., Huang, J., Liu, T., Shen, T., & Gu, Y. (2024). Unsupervised domain adaptation for cross-scene multispectral point cloud classification. IEEE Transactions on Geoscience and Remote Sensing, 62, 1-15.

[2] Gong, T., Lu, X., Sang, Y., Li, S., & Yu, B. (2025). DTN-DETR: Day-night domain adaptive Transformer for nighttime object detection. Computer Engineering, 1–16. https://doi.org/10.19678/j.issn.1000-3428.0252921

[3] Kim, Y. H., Shin, U., Park, J., & Kweon, I. S. (2021). MS-UDA: Multi-spectral unsupervised domain adaptation for thermal image semantic segmentation. IEEE Robotics and Automation Letters, 6(4), 6497-6504.

[4] Rao, C., Fang, X., Zhang, Y., Fan, W., & Zhou, D. (2025). Cross-domain autonomous driving visual segmentation based on enhanced target data learning. ICT Express, 11(1), 53-58.

[5] Xiao, H., Zhou, T., Xiong, S., Li, J., Li, Z., Liu, X., & Deng, T. (2025). Unsupervised domain-adaptive object detection: An efficient method based on UDA-DETR. Neurocomputing, 631, 129711.

[6] Cha, S., Choi, G., Kwak, M., & Choi, J. (2025). Style prompt tuning for bridging visual gaps in autonomous driving. Engineering Applications of Artificial Intelligence, 161, 112105.

[7] Yang, T., Xiao, S., Qu, J., Dong, W., Du, Q., & Li, Y. (2024). Graph embedding interclass relation-aware adaptive network for cross-scene classification of multisource remote sensing data. IEEE Transactions on Image Processing.

[8] Guo, Y., Yu, H., Xie, S., Ma, L., Cao, X., & Luo, X. (2024). Dsca: A dual semantic correlation alignment method for domain adaptation object detection. Pattern Recognition, 150, 110329.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access

This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.