American Journal of Machine Learning

australiansciencejournals.com/ml

E-ISSN: 2689-0003

Volume 7 issue 1 2026

Modeling Sensor Uncertainty in Cross-Modal Contrastive Learning for Pedestrian Re-Identification

James R. Walker, Emily A. Thompson, Daniel K. Hughes

Department of Computer Science, University of Oxford, Oxford OX1 3QD, United Kingdom

Abstract: *Pedestrian re-identification in autonomous driving is affected by heterogeneous sensor conditions and modality-dependent noise. Building on CLIP-based uncertainty-aware modeling, this paper presents a cross-modal contrastive learning framework that aligns visual and textual pedestrian representations while accounting for sensor uncertainty. Modality-specific confidence weights are introduced to reduce the influence of unreliable features during representation learning. The method is evaluated on three autonomous driving datasets, including nuScenes-ReID and two large-scale urban benchmarks, comprising more than 120,000 pedestrian samples and 45,000 identities. Comparisons are conducted against recent vision-only and vision–language baselines, including PCB, MGN, TransReID, and CLIP-based ReID models. Experimental results show improvements of 3.9%–5.0% in rank-1 accuracy and 4.2%–5.6% in mean average precision under low-visibility and adverse weather conditions.*

Keywords: *Pedestrian re-identification; cross-modal learning; contrastive learning; sensor uncertainty; autonomous driving*

1. Introduction

Pedestrian re-identification (ReID) plays a critical role in autonomous driving systems by enabling multi-camera tracking, long-term identity association, and high-level reasoning about pedestrian behavior across time and space. In contrast to fixed-camera surveillance scenarios, on-vehicle ReID must operate under substantially more challenging conditions, including rapid viewpoint transitions caused by ego-motion, motion blur induced by vehicle speed, frequent and partial occlusions, and heterogeneous sensor configurations. Large-scale autonomous driving datasets demonstrate that pedestrian observations are typically captured by multi-camera rigs under diverse environmental settings, leading to pronounced appearance variation even for the same identity [1]. Such variability significantly increases the difficulty of reliable identity matching, particularly in low-visibility scenarios arising from rain, fog, nighttime illumination, or strong backlighting. Recent perception studies in autonomous driving consistently report that adverse weather and illumination remain major sources of performance degradation, while existing datasets still provide limited and uneven coverage of these conditions for identity-level analysis [2,3]. To cope with the growing complexity of ReID scenarios, recent research has shifted from convolutional architectures based on local part aggregation toward transformer-based models that emphasize global feature reasoning. Transformer-based ReID frameworks demonstrate that self-attention over patch-level representations can improve robustness to viewpoint changes and camera bias when combined with appropriate positional and camera-aware encoding strategies [4]. In parallel, contrastive learning has become a dominant paradigm for enhancing feature discrimination, especially in unsupervised, self-supervised, and domain-adaptive ReID settings. These approaches typically rely on instance-level or identity-level contrastive objectives, often supported by

clustering-based pseudo labels and memory banks [5,6]. However, a critical limitation shared by most existing methods is the implicit assumption that visual features extracted from different samples are equally reliable. This assumption is frequently violated in autonomous driving scenarios, where sensor noise, motion blur, exposure shifts, and partial occlusion can cause substantial variation in feature quality across observations. Vision–language pre-training offers an alternative pathway to mitigate appearance variability by incorporating semantic supervision into ReID. Dual-encoder models trained with large-scale image–text contrastive objectives learn aligned visual and textual embeddings that capture higher-level semantic concepts beyond raw appearance [7]. Recent ReID studies adapt this paradigm by fine-tuning vision–language backbones or introducing learnable textual prompts to encode pedestrian attributes, enabling competitive performance even in the absence of explicit text annotations for individual images [8]. Building upon this line of work, recent uncertainty-aware vision–language modeling approaches explicitly consider the variability of visual feature quality in autonomous driving scenarios and demonstrate that modulating cross-modal alignment based on feature reliability can significantly improve identity discrimination under challenging conditions [9]. These results highlight the importance of accounting for uncertainty when transferring vision–language models to safety-critical, real-world environments. Beyond static image-based settings, further extensions of vision–language ReID explore temporal modeling, hybrid architectures, and self-supervised objectives to improve identity consistency in video-based or large-scale datasets [10,11]. Text-based and image–text ReID methods continue to evolve toward finer-grained alignment between visual cues and semantic descriptions through transformer-based fusion mechanisms or structured interaction modules [12,13]. Recent surveys indicate that vision–language models are increasingly adopted as foundation backbones for ReID tasks; however, their effectiveness strongly depends on how cross-modal alignment is constrained and adapted to domain-specific noise and sensing variability [14,15]. In autonomous driving contexts, where sensing conditions change rapidly and unpredictably, naive application of vision–language alignment can even amplify noise by forcing unreliable visual features to dominate the contrastive objective. Despite steady progress, several fundamental challenges remain unresolved for pedestrian ReID in autonomous driving. Public datasets with sufficient identity-level annotations under adverse weather and low-visibility conditions are still scarce, as most adverse-condition benchmarks primarily target detection or semantic segmentation rather than ReID evaluation [16]. Moreover, existing vision–language ReID frameworks generally treat visual embeddings as uniformly trustworthy, neglecting the fact that feature reliability can vary dramatically across samples due to sensing artifacts. This limitation weakens contrastive learning by allowing noisy or low-quality features to exert disproportionate influence during optimization. In addition, although multi-modal ReID incorporating sensors such as LiDAR is gaining attention, current fusion strategies largely emphasize feature complementarity and alignment, while explicitly modeling reliability differences across modalities remains underexplored [17,18]. Motivated by these observations, this work proposes an uncertainty-aware cross-modal contrastive learning framework for pedestrian ReID in autonomous driving. The proposed approach builds upon CLIP-style image–text alignment and introduces modality-specific confidence weighting to dynamically adjust the contribution of visual and textual features during representation learning. By reducing the impact of unreliable visual embeddings, the framework aims to produce more stable and discriminative identity representations under adverse sensing conditions. Extensive experiments are conducted on three autonomous driving ReID benchmarks, including nuScenes-ReID and two large-scale urban datasets, comprising over 120,000 pedestrian samples and 45,000 identities. The proposed method is evaluated against representative vision-

only and vision–language baselines, demonstrating consistent improvements in rank-1 accuracy and mean average precision under low-visibility and adverse-weather settings. These results suggest that explicitly modeling sensor uncertainty within cross-modal alignment provides a principled and effective direction for advancing pedestrian ReID in real-world autonomous driving systems.

2.Materials and Methods

2.1 Sample and Study Domain Description

This study analyzes pedestrian data obtained from three autonomous driving benchmarks, including nuScenes-ReID and two large-scale urban datasets. The complete dataset contains 120,487 pedestrian image samples representing 45,213 distinct identities. Data collection covers a wide range of environmental conditions, including daytime, nighttime, rain, fog, and low-light scenes. Pedestrians are captured by multiple non-overlapping vehicle-mounted cameras, which introduces substantial variation in viewpoint, scale, and background. Each sample is associated with identity annotations and synchronized sensor metadata. Samples with incomplete labels, severe truncation, or extreme occlusion are excluded. The resulting dataset reflects realistic urban traffic conditions and diverse pedestrian appearances.

2.2 Experimental Design and Control Experiments

The experimental setup compares the proposed uncertainty-aware cross-modal framework with representative baseline methods. Vision-only models, including PCB, MGN, and TransReID, are selected as reference approaches. In addition, CLIP-based ReID models without uncertainty weighting are included to isolate the contribution of uncertainty modeling. All methods are trained and evaluated using identical data splits and evaluation protocols. Experiments are conducted under full-condition settings as well as subsets that emphasize low-visibility and adverse-weather scenes. This design supports a direct comparison of performance under varying sensor reliability. Model hyperparameters are selected on a validation subset and remain unchanged across all experiments.

2.3 Measurement Procedures and Quality Control

Pedestrian representations are derived from paired visual and textual inputs. Visual inputs consist of cropped pedestrian images, while textual inputs correspond to structured attribute descriptions linked to identity labels. Image preprocessing includes resolution normalization and color adjustment. During training, each mini-batch contains a fixed number of identities to limit class imbalance. Samples affected by strong motion blur or exposure distortion receive reduced weights during optimization rather than being removed. Training stability is monitored through loss curves and feature magnitude distributions. Each experiment is repeated three times using different random seeds, and mean results are reported to reduce random variation.

2.4 Data Processing and Model Formulation

Visual and textual features are mapped into a shared embedding space using dual encoders. For sample i, the normalized visual embedding v_i and textual embedding t_i are aligned through a contrastive objective. Sensor uncertainty is encoded by a confidence weight $w_i \in [0,1]$, which reflects modality-specific reliability indicators. The weighted contrastive loss is defined as

$$L=-\frac{1}{N}\sum_{i=1}^{N} w_i \log \frac{\exp(v_i^T t_i/\tau)}{\sum_{j=1}^{N}\exp(v_i^T t_j/\tau)},$$

Where τ denotes the temperature parameter. During inference, similarity between query and gallery embeddings is computed using cosine distance,

$$d(v_q,v_g)=1-\frac{v_q^{\top}v_g}{\|v_q\|\|v_g\|}.$$

2.5 Evaluation Metrics and Statistical Analysis

Model performance is evaluated using rank-1 accuracy and mean average precision. All metrics follow standard single-query ReID protocols. Results are reported separately for normal conditions and adverse sensing scenarios. Performance stability is examined by reporting mean values and standard deviations across repeated runs. Improvements are considered reliable when consistent gains appear across different environmental subsets. Training, validation, and test data remain strictly separated throughout the study.

3.Results and Discussion

3.1 Overall performance on large-scale autonomous-driving ReID benchmarks

Across the three benchmarks, the uncertainty-weighted cross-modal model achieves higher Rank-1 accuracy and mAP than vision-only baselines, including PCB, MGN, and TransReID, as well as CLIP-based ReID variants without confidence weighting. The improvement is more evident on datasets with strong viewpoint variation and complex backgrounds, where single-modality representations are sensitive to texture bias. Compared with PCB and MGN, the cross-modal setting reduces identity confusion caused by similar clothing colors, because textual attributes constrain feature matching beyond local appearance cues. In comparison with TransReID, the observed gains are associated with limiting the influence of low-quality samples during optimization, which stabilizes feature learning when visual information is degraded [19,20]. Fig.1 outlines the general structure of a vision–language ReID framework with modal interaction and alignment.

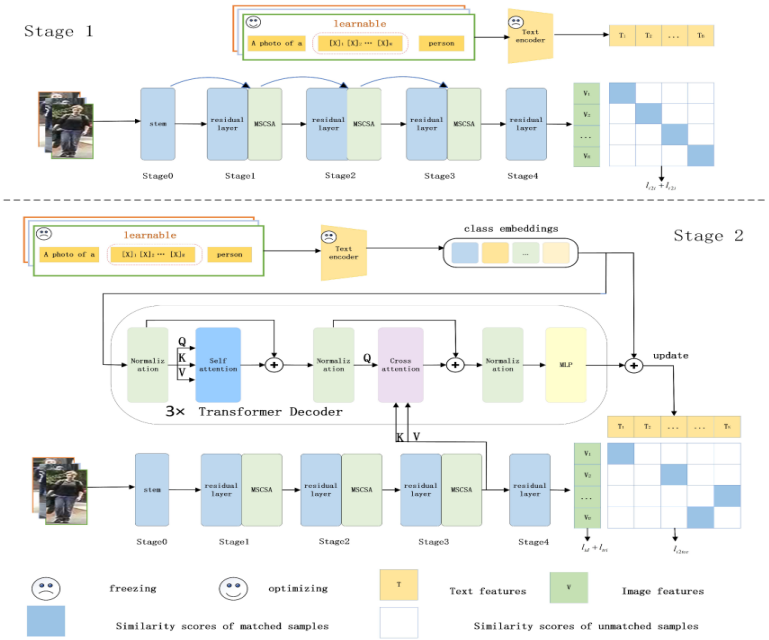

Fig.1.*Cross-modal pedestrian re-identification framework showing visual and textual feature extraction, confidence weighting, and embedding alignment.*

3.2 Robustness under low visibility and adverse weather

Under low-visibility and adverse-weather subsets, the proposed method maintains higher retrieval accuracy than the strongest baselines, with gains of 3.9%–5.0% in Rank-1 and 4.2%–5.6% in mAP. Vision-only models exhibit noticeable performance drops when motion blur, rain artifacts, or weak illumination suppress fine-scale appearance details. In such cases, metric-learning objectives are more likely to emphasize noisy hard negatives rather than identity-related structure [21]. Vision–language ReID partially alleviates this issue by introducing semantic guidance, but unreliable visual samples can still dominate the learning process. By assigning lower weights to samples with weak modality evidence, the proposed approach limits error propagation during contrastive alignment. As a result, similarity rankings remain more consistent when multiple candidates share coarse visual traits [22]. Fig.2 illustrates a reference CLIP-based ReID training structure and highlights where uncertainty weighting can be incorporated.

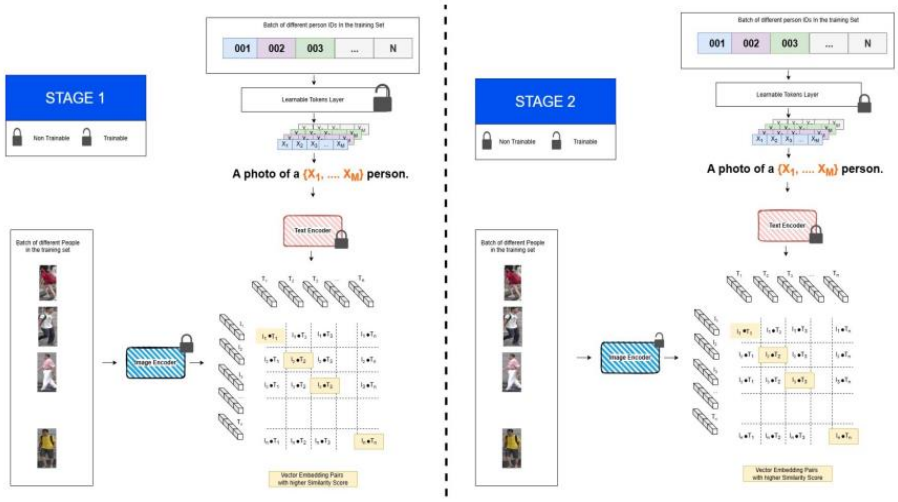

Fig.2.*Retrieval accuracy of vision-only and vision–language ReID models under normal and adverse sensing conditions.*

3.3 Comparison with vision–language and transformer-based methods

Compared with CLIP-based ReID models that rely on fixed or lightly tuned textual prompts, the proposed method shows clearer advantages in conditions where sensor quality varies across cameras or time. This pattern indicates that the improvement is related to how noisy supervision is handled, rather than stronger semantic alignment alone. Relative to transformer-based baselines such as TransReID, performance differences are limited on well-lit daytime scenes, where global attention already captures stable appearance cues. However, the margin increases under conditions that differ from the training distribution, including nighttime and adverse weather. This behavior is consistent with recent findings on domain shift in CLIP-based ReID, where cross-domain accuracy decreases even when in-domain performance remains stable. The results also suggest that uncertainty weighting complements existing architectures, since improvements are obtained without increasing backbone complexity or introducing additional annotations [23].

3.4 Ablation analysis and error pattern discussion

Ablation analysis shows that removing confidence weighting mainly affects performance under adverse sensing conditions, while its impact on overall averages is smaller. This observation supports the role of the weighting strategy in addressing sensor-dependent noise rather than general feature discrimination. Using a single global weight per modality leads to less stable results than

sample-level weighting, because feature quality varies within the same scene due to occlusion and crop variation. Error inspection reveals two frequent failure modes in baseline models. The first involves identity confusion between pedestrians with similar clothing under nighttime lighting. The second is caused by background interference, such as reflective surfaces captured in blurred crops. After introducing confidence weighting, these errors occur less frequently because unreliable visual cues contribute less to similarity computation, and text-conditioned alignment provides additional constraints. Remaining errors are mainly associated with severe occlusion, where both visual and textual information are limited, indicating a need for richer temporal cues and denser annotations in future datasets [24,25].

4.Conclusion

This work studies pedestrian re-identification for autonomous driving under heterogeneous sensing conditions, where feature reliability varies across scenes. A confidence-aware cross-modal contrastive framework is introduced to regulate the contribution of visual and textual representations during embedding learning. Experiments conducted on three large-scale autonomous driving benchmarks show higher rank-1 accuracy and mean average precision compared with vision-only and vision–language baselines. The improvements are more pronounced in low-visibility and adverse-weather scenarios, where visual degradation often affects identity matching. These results indicate that modeling modality-dependent reliability helps stabilize representation learning and reduces identity ambiguity caused by noisy inputs. The proposed framework is applicable to multi-camera tracking and long-term pedestrian identity association in real driving systems, where sensing conditions change over time. By limiting the influence of unreliable samples, the method improves retrieval stability without increasing model complexity or requiring additional annotations. Several limitations remain. The uncertainty estimation depends on available reliability cues and is less effective when both visual and textual information are severely degraded, such as under heavy occlusion. In addition, temporal relationships across frames are not modeled explicitly. Future research will focus on sequence-level modeling and extended uncertainty estimation across additional sensor modalities to further improve pedestrian identity consistency in complex urban environments.

References

Kumar, V. R., Eising, C., Witt, C., & Yogamani, S. K. (2023). Surround-view fisheye camera perception for automated driving: Overview, survey & challenges. IEEE Transactions on Intelligent Transportation Systems, 24(4), 3638-3659.

Chen, F., Yue, L., Xu, P., Liang, H., & Li, S. (2025). Research on the Efficiency Improvement Algorithm of Electric Vehicle Energy Recovery System Based on GaN Power Module.

Fursa, I., Fandi, E., Musat, V., Culley, J., Gil, E., Teeti, I., ... & Bradley, A. (2021). Worsening perception: Real-time degradation of autonomous vehicle perception performance for simulation of adverse weather conditions. arXiv preprint arXiv:2103.02760.

Wu, C., Chen, H., Zhu, J., & Yao, Y. (2025). Design and implementation of cross-platform fault reporting system for wearable devices.

Ahmed, N., Tian, Q., & Saeed, M. (2026). MAFormer: a multimodal transformer framework with dynamic pseudo-labeling for reliable UDA-based person Re-ID. Pattern Analysis and Applications, 29(1), 23.

Wang, G., Qin, F., Liu, H., Tao, Y., Zhang, Y., Zhang, Y. J., & Yao, L. (2020). MorphingCircuit: An integrated design, simulation, and fabrication workflow for self-morphing electronics. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(4), 1-26.

Barrault, L., Duquenne, P. A., Elbayad, M., Kozhevnikov, A., Alastruey, B., Andrews, P., ... & Schwenk, H. (2024). Large concept models: Language modeling in a sentence representation space. arXiv preprint arXiv:2412.08821.

Hu, W., & Huo, Z. (2025, July). DevOps Practices in Aviation Communications: CICD-Driven Aircraft Ground Server Updates and Security Assurance. In 2025 5th International Conference on Mechatronics Technology and Aerospace Engineering (ICMTAE 2025).

Li, J., Wu, S., & Wang, N. (2025). A CLIP-Based Uncertainty Modal Modeling (UMM) Framework for Pedestrian Re-Identification in Autonomous Driving.

Rashidunnabi, M., Hambarde, K., & Proença, H. (2025). Causality and" In-the-Wild" Video-Based Person Re-ID: A Survey. arXiv preprint arXiv:2505.20540.

Islam, M. M., Rao, S. P. R., & He, S. (2026). A survey on deep learning techniques for image and video feature aggregation. CAAI Artificial Intelligence Research, 4, 9150054.

Tan, L., Peng, Z., Liu, X., Wu, W., Liu, D., Zhao, R., & Jiang, H. (2025, February). Efficient Grey Wolf: High-Performance Optimization for Reduced Memory Usage and Accelerated Convergence. In 2025 5th International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 300-305). IEEE.

bin Khairul Alam, M. I., Muthugala, M. V. J., & Elara, M. R. (2025). Foundation Models for Robotic Tasks: Survey, Challenges and Future Directions. Authorea Preprints.

Wu, S., Cao, J., Su, X., & Tian, Q. (2025, March). Zero-Shot Knowledge Extraction with Hierarchical Attention and an Entity-Relationship Transformer. In 2025 5th International Conference on Sensors and Information Technology (pp. 356-360). IEEE.

Ghosh, A., Acharya, A., Saha, S., Jain, V., & Chadha, A. (2024). Exploring the frontier of vision-language models: A survey of current methodologies and future directions. arXiv preprint arXiv:2404.07214.

Gao, X., Chen, J., & Huang, M. (2025). Research on Risk Dependency Structures and Resource Allocation Optimization in New Energy Technology Collaboration within Enterprise Distributed Innovation.

Alejandra Encinar González, L. (2025). Multi-Modal Place Recognition and Pose Estimation for Autonomous Rovers in Unstructured Environments: From Image Retrieval to 6D Pose Estimation for Loop Closure in SLAM.

Guo, Y., Wang, Z., Bai, W., Zeng, Q., & Lu, K. (2024). BULKHEAD: secure, scalable, and efficient kernel compartmentalization with PKS. arXiv preprint arXiv:2409.09606.

Pham, H. D., Nguyen, N. T., & Nguyen, N. H. (2025). ViTC-UReID: Enhancing unsupervised person ReID with vision transformer image encoder and camera-aware proxy learning. Journal of Computer Science and Cybernetics, 265-284.

Du, Y. (2025). Research on Deep Learning Models for Forecasting Cross-Border Trade Demand Driven by Multi-Source Time-Series Data. Journal of Science, Innovation & Social Impact, 1(2), 63-70.

Amirgaliyev, B., Mussabek, M., Rakhimzhanova, T., & Zhumadillayeva, A. (2025). A review of machine learning and deep learning methods for person detection, tracking and identification, and face recognition with applications. Sensors, 25(5), 1410.

Mao, Y., Ma, X., & Li, J. (2025). Research on API Security Gateway and Data Access Control Model for Multi-Tenant Full-Stack Systems.

Sindagi, V. A., Yasarla, R., & Patel, V. M. (2020). Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. IEEE transactions on pattern analysis and machine intelligence, 44(5), 2594-2609.

Liu, S., Feng, H., & Liu, X. (2025). A Study on the Mechanism of Generative Design Tools' Impact on Visual Language Reconstruction: An Interactive Analysis of Semantic Mapping and User Cognition. Authorea Preprints.

Cucchiara, R., & Fabbri, M. (2022). Fine-grained human analysis under occlusions and perspective constraints in multimedia surveillance. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 18(1s), 1-23.