e-ISSN: 3079-6377 p-ISSN: 3079-6369

Learning Occlusion-Robust Pedestrian Representations via Uncertainty-Guided Feature Pruning

Lukas M. Schneider[1]*, Anna K. Vogel[2], Tobias R. Weber[3]

Department of Informatics, Technical University of Munich (TUM), 85748 Garching, Germany

***Corresponding author:** l.schneider@university.edu

Abstract

Occlusion and background interference remain major challenges for pedestrian re-identification in urban traffic environments. Inspired by uncertainty-aware CLIP-based frameworks, this paper introduces an uncertainty-guided feature selection mechanism that adjusts the contribution of local visual regions and semantic cues according to their estimated reliability. The proposed method is evaluated on two autonomous driving datasets with both real-world and synthetic occlusion patterns, covering occlusion ratios from 20% to 60%. Comparisons are conducted against attention-based and part-based ReID methods, including PCB, OSNet, and transformer-based attention models. The proposed approach achieves mAP improvements ranging from 4.5% to 6.2% under severe occlusion conditions, while maintaining comparable performance in fully visible settings.

Keywords

Pedestrian re-identification; occlusion handling; uncertainty-aware attention; autonomous driving; visual perception

1.Introduction

Pedestrian re-identification (ReID) in urban traffic aims to associate the same individual across non-overlapping cameras mounted on moving platforms. Unlike fixed surveillance systems, identity cues in autonomous driving scenarios vary rapidly due to ego-motion, scale change, viewpoint variation, and complex dynamic backgrounds. Among these factors, occlusion remains a dominant source of error because it removes discriminative body regions while introducing mismatched visual content across views. Vehicles, roadside infrastructure, and interactions among pedestrians frequently cause partial visibility, leading to abrupt and unstable changes in the observable area of a target across frames [1]. Such conditions significantly challenge identity consistency and reduce the reliability of appearance-based matching. A growing body of research has investigated occlusion handling in pedestrian ReID through visible-region estimation, feature refinement, and attention-based suppression of interference. Occlusion-aware masking approaches attempt to identify visible body parts and reduce the influence of occluded regions during feature extraction [2,3]. Feature refinement methods further model relations between reliable and unreliable components to correct

e-ISSN: 3079-6377 p-ISSN: 3079-6369

distorted global or part-level embedding [4]. Transformer-based architectures enhance global reasoning by applying structured attention to local tokens and body regions under partial visibility, showing improved robustness on occluded benchmarks [5]. Despite these advances, most existing methods rely on fixed partition strategies or implicit attention patterns, which limits adaptability when occlusion shapes, positions, and durations vary widely in real traffic scenes. Recent studies have begun to reinterpret occlusion as a reliability and uncertainty problem rather than a purely structural one. Instead of explicitly reconstructing missing regions, uncertainty-aware models aim to down-weight unreliable visual evidence during representation learning. Uncertainty-guided attention and context modeling have been shown to reduce the impact of background clutter and missing body parts by adaptively weighting local features [6,7]. Related ideas have also emerged in vision–language ReID, where uncertainty-aware cross-modal alignment accounts for ambiguity in visual observations and improves robustness under degraded sensing conditions [8]. These findings suggest that large vision–language models provide valuable semantic priors for ReID, but their effectiveness depends strongly on how noise and ambiguity are controlled during alignment rather than on model scale alone [9,10]. Despite steady progress, several challenges remain unresolved for occlusion-heavy pedestrian ReID in urban traffic environments. Occlusion patterns differ substantially in shape, spatial distribution, and temporal persistence, while many existing studies evaluate only a limited range of occlusion styles, restricting conclusions about generalization to real driving scenarios [11,12]. In addition, many attention-based and part-based methods implicitly treat selected regions as equally reliable, even though background fragments or occluding objects may receive high attention under motion blur or low-light conditions [13,14]. Transformer-based models further increase representation capacity but may amplify unreliable local tokens when the visible region becomes small, particularly under severe occlusion or strong background interference [15,16]. These limitations indicate that improved modeling of feature reliability is essential for robust identity association. Beyond accuracy, reliability is a critical requirement for autonomous driving systems. Overconfident but incorrect identity matches can propagate errors to downstream modules such as multi-object tracking, trajectory prediction, and behavior analysis. This motivates uncertainty-aware representation learning, where confidence information is explicitly encoded and used during matching. Probabilistic embeddings provide a principled alternative to deterministic feature vectors by modeling each instance as a distribution in feature space, allowing similarity computation to account for ambiguity. Prior work on probabilistic representations demonstrates improved stability and confidence estimation in large-scale retrieval tasks, with

e-ISSN: 3079-6377 p-ISSN: 3079-6369

recent advances reducing computational overhead to practical levels [17,18]. However, such representations remain underexplored in occlusion-focused pedestrian ReID for autonomous driving. Motivated by these observations, this work proposes an occlusion-aware pedestrian ReID framework based on uncertainty-guided feature selection. Instead of treating all local regions equally, the proposed method estimates region-level reliability and dynamically adjusts the contribution of local visual regions and semantic cues during embedding learning. By explicitly modeling uncertainty, the framework reduces the influence of occlusion-induced noise while preserving discriminative information from visible regions. Extensive experiments are conducted on two autonomous driving datasets containing both real-world and synthetic occlusion, covering occlusion ratios from 20% to 60%. The evaluation includes representative part-based, attention-based, and transformer-based baselines, as well as recent occlusion-focused ReID methods. Results demonstrate consistent improvements in mean average precision under severe occlusion while maintaining competitive performance in fully visible scenarios. These findings highlight uncertainty-guided feature selection as a practical and effective direction for improving pedestrian re-identification reliability in urban traffic environments.

2. Materials and Methods

2.1 Sample and Study Area Description

This work analyzes pedestrian images from two autonomous-driving datasets collected in urban traffic environments. The dataset includes 68,432 pedestrian samples corresponding to 21,906 identities. Images are captured by vehicle-mounted cameras operating at road intersections, arterial roads, and crowded pedestrian areas. Occlusion arises from vehicles, roadside facilities, and interactions among pedestrians. To support controlled analysis, synthetic occlusion is applied to part of the data by masking spatial regions, with occlusion ratios ranging from 20% to 60%. The samples cover a wide range of clothing styles, body proportions, and motion states. Images with incorrect labels or severe truncation are excluded before training and evaluation.

2.2 Experimental Design and Control Experiments

The experimental design compares an uncertainty-guided feature selection model with established pedestrian ReID methods. Part-based approaches such as PCB and compact convolutional networks such as OSNet are included as baseline models. Transformer-based attention architectures are also evaluated as references for region aggregation. All methods follow identical training and testing splits to ensure comparability. The experimental model

e-ISSN: 3079-6377 p-ISSN: 3079-6369

applies region-level reliability weighting during feature aggregation, whereas control models combine regional features without reliability modulation. Experiments are conducted on fully visible samples and on subsets with increasing occlusion levels. This setup isolates the effect of occlusion on identity discrimination across different modeling strategies.

2.3 Measurement Procedures and Quality Control

All images are resized to a fixed resolution and normalized using dataset-specific statistics. Feature extraction is performed on predefined spatial regions to retain local body information. For each region, a reliability score is computed based on feature stability and activation distribution. Mini-batches are constructed to contain balanced identity samples during training. To limit noise influence, samples affected by strong motion blur or illumination distortion receive reduced contribution during optimization instead of being removed. Training behavior is monitored through loss trends and embedding variance. Each experiment is repeated three times with different random seeds, and average values are reported.

2.4 Data Processing and Model Formulation

Let f_i^r denote the feature vector extracted from region r of image i. Each region is assigned a reliability weight $\alpha_i^r \in [0,1]$ derived from uncertainty estimation. The aggregated representation F_i is computed as

$$F_i = \sum_{r=1}^{R} \alpha_i^r f_i^r,$$

Where R denotes the number of spatial regions. Identity learning is supervised using a softmax-based classification loss,

$$L_{id} = -\sum_{i=1}^{N} \log \frac{\exp(w_{y_i}^T F_i)}{\sum_{k=1}^{K} \exp(w_k^T F_i)},$$

Where y_i represents the identity label and K is the total number of identities. This formulation reduces the influence of unreliable regions under partial occlusion.

2.5 Evaluation Metrics and Statistical Analysis

Performance is evaluated using rank-1 accuracy and mean average precision under a single-query protocol. Results are reported separately for each occlusion ratio to examine sensitivity to visibility loss. Statistical stability is assessed by reporting the mean and standard deviation across repeated runs. Performance trends are interpreted based on consistency across

e-ISSN: 3079-6377 p-ISSN: 3079-6369

occlusion levels rather than isolated values. Training, validation, and test sets are strictly separated throughout the experiments.

3. Results and Discussion

3.1 Performance trends across occlusion ratios

Across both autonomous-driving datasets, retrieval performance decreases as the occlusion ratio increases from 20% to 60%. The rate of degradation differs across method categories. Part-based pooling and standard attention aggregation show a rapid decline once large body regions are blocked, because the resulting embeddings rely on a limited set of visible patches mixed with background content [19, 20]. The uncertainty-guided feature selection model maintains higher mean average precision across all occlusion levels, with the largest margin observed in the 50%–60% range. In this regime, background interference frequently enters the feature map. Reliability weighting reduces the influence of regions with unstable activations, so the final representation depends mainly on cues that remain consistent across views [21]. A similar motivation has been reported in recent occlusion-aware fusion frameworks (Fig.1).

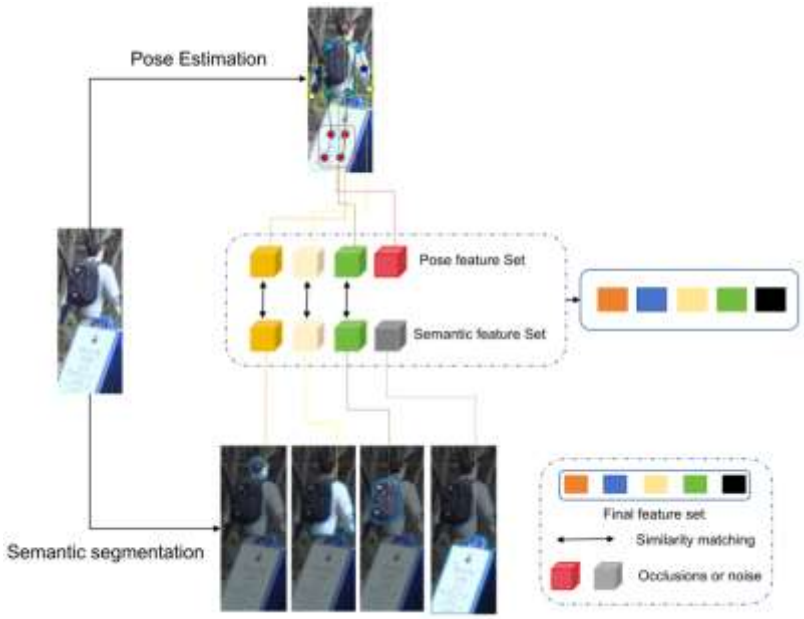

Figure 1 Occlusion-aware feature aggregation with region reliability weighting for pedestrian re-identification.

3.2 Comparison with PCB and OSNet

PCB shows competitive performance under low occlusion because stripe-based partitioning preserves local structure when most body parts remain visible. Its limitation becomes evident under irregular occlusion. Occluding objects often block the torso or legs unevenly, which causes fixed stripes to combine foreground fragments with occluders and reduces cross-view consistency. OSNet captures multi-scale information and handles scale variation, but it still

e-ISSN: 3079-6377 p-ISSN: 3079-6369

aggregates features from regions dominated by vehicles, poles, or nearby pedestrians when visibility is limited. Under 50%–60% occlusion, these effects increase the number of false matches between identities sharing similar background layouts. The uncertainty-guided selector reduces this error source by lowering the contribution of unreliable regions, which improves ranking stability while maintaining comparable accuracy in fully visible settings [22,23].

3.3 Comparison with transformer attention baselines and related occlusion models

Transformer-based attention improves global context modeling but remains sensitive to saliency shifts introduced by occluders. Under heavy occlusion, high-contrast objects often receive strong attention, and the corresponding tokens increase similarity scores between different identities captured in similar traffic scenes. The uncertainty-guided mechanism limits this effect by applying reliability weighting before region aggregation. As a result, unstable regions cannot dominate the final descriptor even when attention responses peak [24,25]. This behavior differs from attention-only strategies that rely on sharper focus without explicit reliability control. Similar attention-enhanced ViT designs have been reported for occluded ReID (Fig.2), where improved token discrimination is observed, but performance remains affected when unreliable regions dominate the visible area.

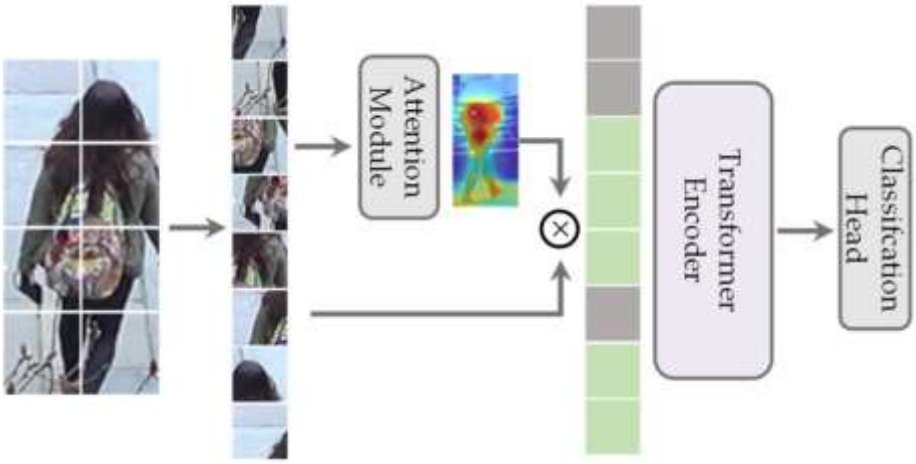

Figure 2 Vision transformer with attention-based region modeling for pedestrian re-identification under occlusion.

3.4 Error patterns, ablation interpretation, and remaining limits

Qualitative inspection reveals two frequent error patterns in baseline models under high occlusion. The first involves background-driven matches, where large occluders appear in similar spatial locations across camera views and lead to high similarity between different identities. The second involves part misalignment, where visible limbs fall into different stripes or tokens across views and weaken identity consistency. With uncertainty-guided

e-ISSN: 3079-6377 p-ISSN: 3079-6369

selection, both error types occur less frequently because regions with unstable evidence receive lower weights, and similarity scores become less sensitive to occluder texture. Ablation results follow the same trend. Removing reliability weighting mainly affects performance in the 50%–60% occlusion subsets, while differences under full visibility remain limited. Remaining errors are concentrated in cases of near-complete occlusion or strong truncation, where discriminative cues are scarce and single-frame inference offers limited recovery. This observation indicates that region selection alone cannot fully resolve extreme occlusion and that sequence-level reasoning is required in such cases [26].

4. Conclusion

This work addresses pedestrian re-identification in urban traffic scenarios where occlusion and background interference are common. A feature selection strategy guided by uncertainty is introduced to adjust the contribution of local visual regions based on their reliability, instead of assuming equal importance across all regions. Experiments on autonomous-driving datasets with controlled occlusion levels show higher mean average precision under moderate and severe occlusion, while performance under full visibility remains comparable to existing methods. These results show that reducing the influence of unreliable regions during feature aggregation helps limit occlusion-related noise and background bias when visible body parts differ across views. The proposed method is suitable for vehicle-based multi-camera tracking systems operating in crowded or visually complex environments, where partial visibility is frequent. By relying on region-level reliability rather than explicit occlusion masks, the approach remains applicable across varied occlusion patterns without additional supervision. Several limitations should be noted. The current design does not model long-term temporal continuity, which restricts its effectiveness when pedestrians are almost fully occluded. In addition, the evaluation focuses on single-frame representations and does not fully use information from consecutive frames. Future studies will explore sequence-level modeling and temporal uncertainty estimation to further improve identity consistency in real-world autonomous driving systems.

References

[1] De Borba, T., Vaculín, O., Marzbani, H., & Jazar, R. N. (2025). Increasing safety of vulnerable road users in scenarios with occlusion: A collaborative approach for smart infrastructures and automated vehicles. IEEE Access.

e-ISSN: 3079-6377 p-ISSN: 3079-6369

[2] 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.

[3] Nguyen, V. D., Mantini, P., & Shah, S. K. (2025). Occlusion-aware appearance and shape learning for occluded cloth-changing person re-identification. Pattern Analysis and Applications, 28(2), 1-17.

[4] 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.

[5] Pereira, G. A., & Hussain, M. (2024). A review of transformer-based models for computer vision tasks: Capturing global context and spatial relationships. arXiv preprint arXiv:2408.15178.

[6] 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.

[7] Cha, D., Kakuba, S., Bitwire, G. A., & Han, D. S. (2025). EDATRAF: An Enhanced Depth-Aware Transformer for Monocular 3D Object Detection Using Feature Fusion and Cross-Query Attention. IEEE Access.

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

[9] Ramos, L. T., & Casas, E. (2025). Applications, trends, and perspectives of large language models in education: A literature review. Authorea Preprints.

[10] Vilakati, S. (2025). Prompt engineering for accurate statistical reasoning with large language models in medical research. Frontiers in Artificial Intelligence, 8, 1658316.

[11] 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.

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

[13] Sikdar, A., Liu, Y., Kedarisetty, S., Zhao, Y., Ahmed, A., & Behera, A. (2025). Interweaving insights: High-order feature interaction for fine-grained visual recognition. International Journal of Computer Vision, 133(4), 1755-1779.

[14] 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.

[15] El-Saleh, A. A. (2025). A Comprehensive Review of Face Detection Techniques for Occluded Faces: Methods, Datasets, and Open Challenges. Comput Model Eng Sci, 143(3).

[16] 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.

e-ISSN: 3079-6377 p-ISSN: 3079-6369

[17] Pendleton, C., Harrington, E., Fairbrother, G., Arkwright, J., Fenwick, N., & Katrix, R. (2025). Probabilistic lexical manifold construction in large language models via hierarchical vector field interpolation. arXiv preprint arXiv:2502.10013.

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

[19] Cosma, A., Catruna, A., & Radoi, E. (2023). Exploring self-supervised vision transformers for gait recognition in the wild. Sensors, 23(5), 2680.

[20] 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.

[21] Truman, A., & Kutas, M. (2024). Flexible Conceptual Representations. Cognitive Science, 48(6), e13475.

[22] 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).

[23] Mohammed, T. K., & Motupalli, R. (2026). CORA-Net: Calibration-oriented, Robust, and Ordinal-aware Network with Dual-path Cross-scale Attention and Uncertainty-guided Co-teaching for Knee OA Grading. International Journal of Intelligent Engineering & Systems, 19(2).

[24] Chowdhury, M. H., Chowdhury, M. E., Khan, M. S., Ullah, M. A., Mahmud, S., Khandakar, A., ... & Hasan, A. (2023). Self-attention MHDNet: A novel deep learning model for the detection of R-peaks in the electrocardiogram signals corrupted with Magnetohydrodynamic effect. Bioengineering, 10(5), 542.

[25] 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.

[26] Trivigno, G. (2025). Towards robust visual geo-localization: Cross-domain, sequential, and fine-grained approaches for place recognition.