ISSN: 3079-6342

Reducing Annotation Cost in Vision Language Pedestrian Re Identification via Uncertainty Driven Sampling

Michael Anderson[1], Daniel Rodriguez[2], Yi Chen[3]*
Department of Computer Science, Stanford University, Stanford, CA 94305, USA

***Corresponding author:** yi.chen@stanford.edu

Abstract

Scaling pedestrian re-identification for autonomous driving is limited by the cost of identity labeling across large camera networks. Inspired by CLIP-based uncertainty modal modeling, this paper proposes an active learning approach that selects labeling candidates using uncertainty in the joint vision–language embedding space. The method combines (i) uncertainty sampling for ambiguous matches, (ii) diversity sampling based on embedding coverage, and (iii) batch acquisition with redundancy control. Experiments are conducted on a large-scale dataset with 400,000 images and 50,000 identities under incremental labeling budgets from 5% to 30%. Compared with random sampling, core-set selection, and margin-based acquisition using TransReID embeddings, the proposed strategy reaches 95% of full-supervision mAP using 18%–22% fewer labeled identities, while reducing annotation time by an estimated 20%–25% under standard labeling workflows.

Keywords

Active learning; annotation efficiency; pedestrian re-identification; uncertainty sampling; vision–language models

1. Introduction

Pedestrian re-identification (Re-ID) is widely deployed in large-scale camera networks and is playing an increasingly important role in autonomous driving systems. It supports cross-camera identity association, long-range tracking, and post-event analysis, and serves as a foundational component for higher-level perception and decision modules. Recent progress in Re-ID has been driven primarily by advances in feature representation learning and the availability of large labeled datasets. In particular, transformer-based architectures have significantly improved robustness to viewpoint variation, pose change, and spatial misalignment [1]. Despite these gains, large-scale deployment in real traffic environments remains constrained by the cost of identity annotation. Accurate labeling requires consistent identity association across cameras and over time, which is labor-intensive and difficult to scale across cities and vehicle fleets. As camera networks and data volumes continue to expand, annotation cost has become a major obstacle to frequent model updates and long-term system maintenance [2]. Vision–language models have recently attracted substantial attention as a potential means to reduce dependence on task-specific identity labels. CLIP-style pretraining learns transferable representations from large-scale image–text corpora and

ISSN: 3079-6342

has demonstrated strong generalization under distribution shift and domain variation [3,4]. Building upon this paradigm, uncertainty-aware modeling in vision–language embedding spaces has been shown to improve robustness and reliability for pedestrian Re-ID in autonomous-driving scenarios by explicitly accounting for representation disagreement rather than relying solely on similarity magnitude [5]. Several studies further adapt vision–language embeddings to Re-ID using prompt learning or text-free constraints, avoiding the need for manual identity descriptions while preserving semantic structure [6,7]. Survey analyses indicate that such embeddings retain useful cross-domain organization, but also reveal persistent limitations in fine-grained identity discrimination, occlusion handling, and score stability [8]. These findings suggest that vision–language representations provide a strong foundation for Re-ID, yet do not by themselves resolve the challenge of annotation efficiency. Active learning offers a complementary strategy by reducing labeling cost through selective annotation under a fixed budget. Early approaches rely primarily on uncertainty-based ranking, while more recent methods combine uncertainty with diversity criteria to avoid selecting redundant samples and to improve coverage of the data distribution [9,10]. However, batch acquisition strategies often perform poorly when uncertainty estimates are unreliable or when embedding redundancy is ignored, resulting in slow performance gains despite additional annotations [11,12]. These issues are particularly pronounced in autonomous driving data, where strong temporal correlation, repeated appearances of similar pedestrians, and long-tailed identity distributions are common. Applying active learning to pedestrian Re-ID introduces challenges beyond those encountered in standard classification tasks. Supervision is identity-based and depends on relational constraints between images or tracklets rather than independent class labels. Recent work explores relation-aware selection strategies, such as selecting informative pairs or support sets, to improve discriminative learning with fewer annotated identities [13]. Other studies investigate open-set and continual Re-ID settings, where new identities emerge over time and data distributions shift across locations and camera configurations [14]. In addition, occlusion, truncation, and partial observations reduce the reliability of naive uncertainty measures, since many visually ambiguous samples contribute limited information for improving identity discrimination [15]. Uncertainty estimation within joint embedding spaces provides a useful signal for addressing these challenges. In vision–language models, disagreement between visual and textual representations often reflects unreliable or ambiguous matches, offering a principled basis for sample prioritization. Recent evidence shows that calibrated uncertainty can improve active learning performance under distribution shift when compared with raw confidence scores

ISSN: 3079-6342

[16]. Nevertheless, uncertainty alone may bias selection toward outliers or repeatedly sampled patterns. This limitation motivates the integration of uncertainty sampling with diversity-aware criteria that encourage broader coverage of the embedding space and reduce redundancy [17]. Such a combined strategy is particularly suitable for large-scale Re-ID, where both ambiguity control and representational coverage are essential.Based on these observations, this work presents an annotation-efficient pedestrian Re-ID framework that leverages uncertainty in a vision–language embedding space for active sample selection. The proposed method integrates uncertainty sampling to target ambiguous identity relations, diversity sampling to expand embedding coverage, and batch-level redundancy control to ensure efficient use of labeling budgets. Experiments are conducted on a large-scale dataset containing approximately 400,000 images and 50,000 identities, under incremental labeling budgets ranging from 5% to 30%. The approach is evaluated against random sampling, core-set selection, and margin-based acquisition strategies built on conventional Re-ID embeddings. The objective is to substantially reduce identity annotation effort while maintaining strong retrieval performance in conditions representative of autonomous driving, thereby enabling more scalable and sustainable deployment of Re-ID systems.

2.Materials and Methods

2.1 Samples and Study Scope

The experiments use pedestrian images collected from urban traffic scenes with multiple cameras. The dataset includes about 400,000 images from approximately 50,000 identities. Images are captured by vehicle-mounted cameras operating under normal driving conditions. Data cover daytime and nighttime scenes, different weather, and varying traffic density. Pedestrians differ in clothing style, body shape, walking speed, and level of occlusion. For annotation, images are grouped into short tracklets to reflect practical labeling procedures. Only a subset of identities is labeled at each acquisition stage. All images are resized to a fixed resolution and normalized using standard preprocessing steps.

2.2 Experimental Design and Baseline Comparison

An incremental labeling setting is adopted, with annotation budgets ranging from 5% to 30% of all identities. At each step, a selection method chooses a batch of unlabeled samples for annotation, followed by model retraining. The proposed selection strategy is compared with random sampling, core-set selection based on embedding distance, and margin-based acquisition using TransReID features. These baselines are widely used in active learning and Re-ID studies and provide clear reference points. All methods share the same backbone

ISSN: 3079-6342

network, training schedule, and data splits. This design ensures that performance differences are caused by the selection strategy rather than other factors.

2.3 Measurement Procedure and Quality Control

Feature extraction during sample selection uses a frozen vision–language encoder. Model evaluation uses a Re-ID network trained on the currently labeled set. All feature vectors are normalized before similarity and uncertainty computation. Images with severe blur or incomplete pedestrian regions are removed before selection to reduce noise. Each experiment is repeated with different random seeds to limit the effect of batch sampling variation. Validation data are kept separate from both training and selection pools. Performance results are averaged across runs to check stability.

2.4 Data Processing and Model Formulation

Let $f(x_i) \in R^d$ denote the normalized embedding of image x_i. Similarity between two samples x_i and x_j is computed using cosine similarity:

$$s_{ij} = \frac{f(x_i)^\top f(x_j)}{\|f(x_i)\| \|f(x_j)\|}.$$

Uncertainty for an unlabeled sample is estimated from the spread of its similarity scores with respect to the labeled set. Diversity is enforced by reducing the score of samples that are close to already selected candidates within the same batch. The final acquisition score is obtained by combining uncertainty and diversity terms with fixed weights. Samples with the highest scores are selected for annotation.

2.5 Evaluation Metrics and Analysis

Re-identification performance is measured using mean Average Precision and Rank-1 accuracy on a held-out test set. Annotation efficiency is evaluated by comparing model performance at the same labeling budget. For each method, the number of labeled identities needed to reach a target performance level is reported. Results are compared across repeated runs to verify consistency. This evaluation reflects realistic conditions where annotation cost, rather than dataset size, limits model improvement.

3. Results and Discussion

3.1 Label efficiency under incremental budgets

Under incremental labeling budgets from 5% to 30%, the proposed selection method reaches strong retrieval performance earlier than all baseline strategies. The advantage is most evident at low budgets below 15%, where random sampling often selects visually similar

ISSN: 3079-6342

samples and wastes annotation effort. Core-set selection improves coverage but frequently favors easy identities that contribute limited boundary refinement. By contrast, the proposed strategy selects samples that are both uncertain and spatially dispersed in the embedding space. This leads to faster improvement in mean Average Precision because the labeled set includes difficult identity boundaries and underrepresented appearance patterns [18,19]. The general workflow of such deep active learning systems is illustrated in Fig.1:Fig.1. General pipeline of deep active learning with iterative training, sample selection, annotation, and retraining.

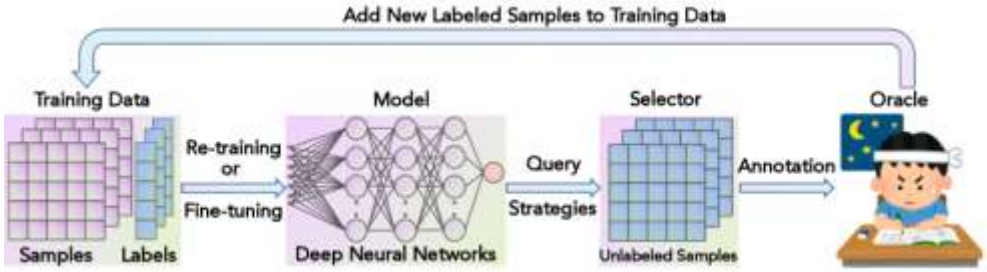

Figure 1 Iterative active learning process with repeated cycles of training, sample selection, annotation, and model update.

3.2 Comparison with uncertainty-only and diversity-only selection

Selection based only on uncertainty tends to focus on samples with poor image quality, such as blur or partial crops. These samples are difficult, but they do not always improve identity separation. As a result, performance gains slow after the first few annotation rounds. Diversity-only methods reduce redundancy, but they often select many easy samples that already lie far from decision boundaries. The combined strategy avoids both issues. Uncertainty highlights samples that affect identity discrimination, while diversity spreads selection across cameras and appearance modes. This balance allows the method to reach near-saturated performance using fewer labeled identities. The result suggests that sample usefulness depends not only on difficulty, but also on how new information complements the existing labeled set [20,21].

3.3 Effects on long-tail identities and occlusion-heavy subsets

The largest improvements appear for long-tail identities and scenes with frequent occlusion. Under limited budgets, rare identities are often missed by random selection and remain poorly represented in the feature space. The uncertainty-driven step increases the chance that these identities are labeled early. Redundancy control further prevents repeated selection of similar views from the same camera or time segment. In occlusion-heavy subsets, the method favors samples that clarify ambiguous body regions rather than repeatedly

ISSN: 3079-6342

selecting severely degraded views. This behavior reduces wasted labels and improves cluster formation for partially observed pedestrians. These results support the view that uncertainty in a vision–language embedding space can reveal samples with high corrective value [22].

3.4 Practical implications and relation to recent vision–language studies

From a practical perspective, the results support an annotation strategy that emphasizes label impact rather than label quantity. Early annotation rounds benefit from focusing on uncertain and diverse samples, while later rounds naturally shift toward broader coverage. The observed learning curves are consistent with recent studies that use vision–language models as guides for data selection and representation shaping [23,24]. These studies show that structured teacher signals can improve learning efficiency under limited supervision. Fig.2 presents an example of such a teacher-guided active learning loop:Fig.2. Active learning framework guided by a vision–language teacher for sample selection and representation refinement.

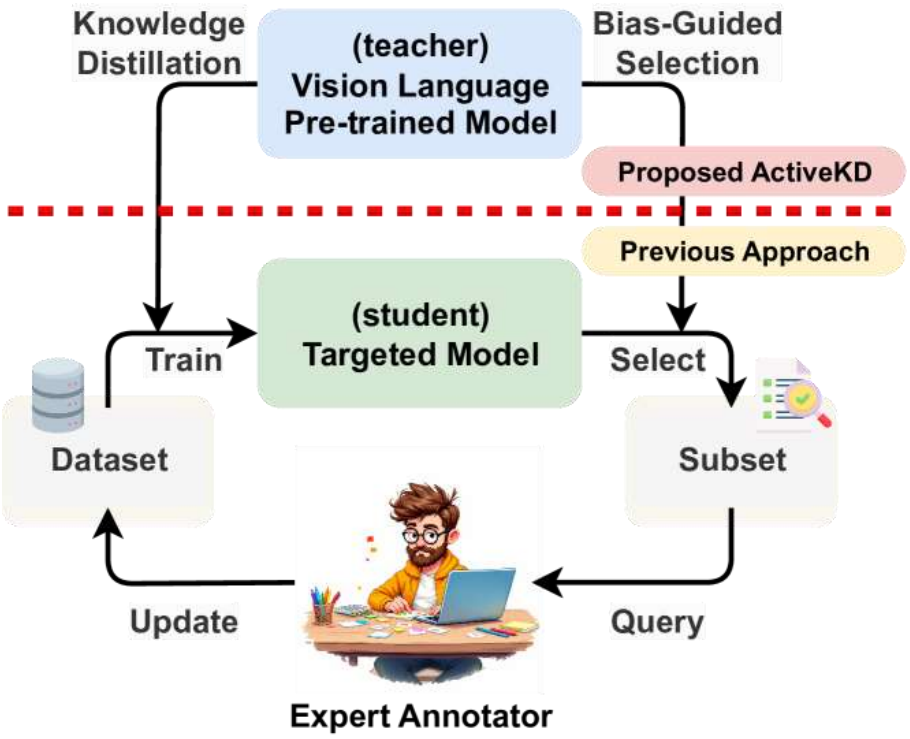

Figure 2 Active learning scheme guided by vision–language embeddings for uncertainty-based sample selection.

4. Conclusion

This study examines annotation-efficient pedestrian re-identification in large-scale autonomous driving settings. An uncertainty-driven sample selection method is used within a vision–language embedding space. Experimental results show that the method reduces labeling effort while maintaining strong retrieval accuracy, particularly under limited annotation budgets. The improvement comes from selecting samples that are both uncertain

ISSN: 3079-6342

and diverse, which helps refine identity boundaries and improves coverage of rare appearance patterns. This approach addresses a common weakness of existing active learning methods, which often select redundant or weakly informative samples. The findings suggest that uncertainty estimated in joint embeddings provides a practical signal for guiding identity annotation. The method is suitable for large camera networks, where data volume grows quickly and labeling cost limits frequent retraining. Possible applications include incremental model updates for autonomous vehicles and long-term multi-camera systems. The study is limited by the use of offline selection and fixed acquisition rules. Future work will investigate adaptive selection strategies and online integration to improve performance in changing environments.

References

[1] 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.

[2] Karim, M. M., Khan, S., Van, D. H., Liu, X., Wang, C., & Qu, Q. (2025). Transforming data annotation with ai agents: A review of architectures, reasoning, applications, and impact. Future Internet, 17(8), 353.

[3] 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.

[4] Madsen, P., & Spiik, W. (2025). Evaluating Subjective Understanding in Multimodal Models: A Case Study on Fashion Style Representation using Embeddings.

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

[6] Asperti, A., Fiorilla, S., Nardi, S., & Orsini, L. (2025). A review of recent techniques for person re-identification. arXiv preprint arXiv:2509.22690.

[7] 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.

[8] Raja, R., Vats, A., Thawakar, O., & Ashraf, T. (2025). Object Tracking: A Comprehensive Survey From Classical Approaches to Large Vision-Language and Foundation Models. Available at SSRN 5541079.

[9] 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.

ISSN: 3079-6342

[10] Mohammadi, S., & Ascenso, J. (2025). Uncertainty-driven Sampling for Efficient Pairwise Comparison Subjective Assessment. IEEE Transactions on Multimedia.

[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] Kim, D. D., Chandra, R. S., Yang, L., Wu, J., Feng, X., Atalay, M., ... & Bai, H. X. (2024). Active learning in brain tumor segmentation with uncertainty sampling and annotation redundancy restriction. Journal of Imaging Informatics in Medicine, 37(5), 2099-2107.

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

[14] Cunico, F., & Cristani, M. (2024, September). Multi-Camera Industrial Open-Set Person Re-Identification and Tracking. In European Conference on Computer Vision (pp. 121-135). Cham: Springer Nature Switzerland.

[15] 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.

[16] Querol, L. S., Nagahara, H., & Hayashi, H. (2024, September). CALICO: Confident Active Learning with Integrated Calibration. In International Conference on Artificial Neural Networks (pp. 116-130). Cham: Springer Nature Switzerland.

[17] 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.

[18] Raza, A., Hanif, F., & Mohammed, H. A. (2025). Analyzing the enhancement of CNN-YOLO and transformer based architectures for real-time animal detection in complex ecological environments. Scientific Reports, 15(1), 39142.

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

[20] Shwartz Ziv, R., & LeCun, Y. (2024). To compress or not to compress—self-supervised learning and information theory: A review. Entropy, 26(3), 252.

[21] 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.

[22] Venkataramanan, A., Bodesheim, P., & Denzler, J. (2025). Probabilistic Embeddings for Frozen Vision-Language Models: Uncertainty Quantification with Gaussian Process Latent Variable Models. arXiv preprint arXiv:2505.05163.

ISSN: 3079-6342

[23] 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).

[24] Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., ... & Chandra, V. (2024). An introduction to vision-language modeling. arXiv preprint arXiv:2405.17247.