International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 15, Issue 02, February 2026)

Evaluation of Real-Time Person Detection and Re-Identification Systems Using YOLO, Byte Track, OSNet_x1.0, and ResNet50

Kushal Ahuja[1], Katukam Ravi[2], Arjun Nagulapally[3]

[1]Trainee Data Scientist, [2]Principal Data Scientist, [3]Chief Technical Officer, AIONOS, Hyderabad, India

***Abstract*— This paper presents a systematic evaluation of two real-time person detection, tracking, and re-identification pipelines: YOLO + ByteTrack + OSNet_x1.0 and YOLO + ByteTrack + ResNet50. The proposed system leverages YOLO for accurate person detection, ByteTrack for robust multi-object tracking, and either OSNet_x1.0 or ResNet50 as the re-identification backbone, with identity association achieved using cosine similarity and the Hungarian assignment algorithm. Experiments conducted using a Python-based implementation on an NVIDIA RTX 3050 GPU demonstrate that both pipelines maintain approximately 8 FPS in real-time deployments, with OSNet_x1.0 achieving up to 80% Top- 1 ReID accuracy. Comparative analyses reveal that OSNet_x1.0 produces superior embedding separability, greater identity stability, and fewer identity switches compared to ResNet50, as reflected in tracking metrics such as MOTA and IDF1. The primary contributions include a rigorous quantitative comparison of these approaches, a detailed exploration of computational and accuracy trade-offs, and an analysis of practical limitations in challenging real-world scenarios.**

I. INTRODUCTION

Real-time person detection, tracking, and re-identification play a central role in a wide range of computer vision applications, including intelligent surveillance, access control, and human–computer interaction. Achieving robust performance in these tasks requires systems that can maintain accurate identities over time, especially in environments characterized by dynamic backgrounds, occlusions, and frequent appearance changes.

The motivation for this study stems from the need to systematically assess how different re-identification backbones—specifically OSNet_x1.0 and ResNet50—influence embedding quality, identity consistency, and computational efficiency within a unified real-time pipeline. This research provides a detailed quantitative comparison between these architectures, evaluates their identity assignment effectiveness under real-time constraints, and investigates practical trade-offs between accuracy and efficiency.

The remainder of the paper is structured as follows: Section 2 reviews related work; Section 3 describes the integrated system architecture; Section 4 presents the experimental methodology; Section 5 reports and analyzes the results; Section 6 discusses failure cases and limitations; and Section 7 concludes the paper.

II. RELATED WORK

Considerable research attention has been devoted to advancing real-time person detection, multi-object tracking, and person re-identification, leading to a diverse ecosystem of algorithms that balance speed, accuracy, and robustness. Among detection approaches, the YOLO family stands out for its efficient object localization and classification, supporting robust deployment under time constraints.

In tracking, methods such as ByteTrack excel in maintaining identity associations across challenging sequences, particularly by pairing effectively with real-time detectors. In the domain of person re-identification, architectures like OSNet_x1.0 and ResNet50 provide strong baselines, differing in embedding representation quality and suitability for low- latency applications.

This study extends previous evaluations by systematically comparing YOLO-based pipelines coupled with ByteTrack while contrasting distinct re-identification backbones, thereby contextualizing embedding discriminability and identity preservation under real- world computational constraints.

Furthermore, advances in embedding learning and assignment strategies have significantly improved person re-identification systems. Cosine similarity has emerged as a preferred metric for comparing ReID embeddings due to its effectiveness in quantifying directional similarity across varying appearance conditions. The Hungarian assignment algorithm, when employed alongside such embeddings, enables optimal identity correspondence by minimizing association cost matrices over consecutive frames. Consequently, these approaches have shaped recent benchmarks and heightened evaluation standards emphasizing identity consistency and temporal alignment.

III. System Architecture

At the core of the evaluated system pipeline lies a modular integration of proven components, each fulfilling a specialized role to achieve robust real-time person detection, tracking, and re-identification.

The process begins with YOLO serving as the primary object detector, responsible for high-throughput localization and categorization of person instances within each video frame. These detected bounding boxes are then passed to the ByteTrack multi-object tracker, which efficiently maintains temporal identity continuity even under complex motion patterns and partial occlusions through reliable association mechanisms.

For person re-identification, the system supports interchangeable use of either the OSNet_x1.0 or ResNet50 backbone. Each model converts cropped person images into compact embedding vectors suitable for identity matching. Embeddings are compared frame-to-frame using cosine similarity, while the Hungarian assignment algorithm ensures optimal identity matching by solving the bipartite assignment problem in real time. This harmonizes detection, tracking, and ReID modules into a cohesive workflow.

IV. Implementation Details

The integrated pipeline is implemented primarily in Python, leveraging PyTorch for deep learning model development, OpenCV for image processing, and FastAPI in conjunction with WebRTC to enable real-time inference and API deployment. This technical stack facilitates both rapid prototyping and production-level deployment while supporting GPU acceleration and asynchronous processing.

The modular architecture allows seamless swapping of the re-identification backbone between OSNet_x1.0 and ResNet50 through standardized interfaces that unify model input preprocessing, feature extraction, and embedding normalization. This design enables efficient comparative evaluation without altering detection or tracking components, thereby streamlining ablation studies and ensuring reproducibility.

V. Methodology

To conduct a rigorous comparison of the two pipelines, a unified evaluation protocol was established that quantifies detection, tracking, and re-identification performance under consistent conditions.

Person detection is assessed using mean Average Precision (mAP) and Intersection over Union (IoU) for bounding box quality, complemented by tracking metrics such as Multiple Object Tracking Accuracy (MOTA) and Identity F1 score (IDF1), which jointly represent precision in tracking persistence and identity consistency. For re- identification, each detected tracklet is assigned an embedding by either OSNet_v1.0 or ResNet50; these embedding vectors are compared across frames using the cosine similarity metric.

$$s(\mathbf{a},\mathbf{b}) = \frac{\mathbf{a}\cdot\mathbf{b}}{\|\mathbf{a}\|\,\|\mathbf{b}\|}$$

In every frame, the resulting cosine similarity matrix is interpreted as a cost matrix, and the Hungarian algorithm is employed to associate detections and maintain stable IDs by minimizing this assignment cost in polynomial time. This systematic approach ensures that results are comparable, reproducible, and reflective of both accuracy and real-time system constraints, consistent with prevailing evaluation standards in the field (Mekled et al., 2024).

Moreover, a comprehensive suite of quantitative metrics was selected to capture the multi-faceted nature of real-time person detection, tracking, and re-identification performance. Top-1 ReID accuracy was utilized to measure the proportion of correctly matched identities in the most probable scenario, providing a direct indicator of embedding quality. The metric of identity (ID) switches quantifies the frequency with which a tracked identity is incorrectly reassigned, thereby assessing the system's stability under real-world complexities such as occlusions and ambiguous visual cues. MOTA (Multiple Object Tracking Accuracy) consolidates errors in detection, false positives, and identity preservation to yield an aggregate representation of tracking efficacy, while IDF1 reflects the balance between precise identity assignment and continuity, considering both precision and recall across distant frames. This selection aligns with the rationale of existing evaluation frameworks that emphasize holistic, practical assessment in realistic scenarios, ensuring versatility and comparability with state-of-the-art research (Alikhanov et al., 2025).

VI. Experimental Setup

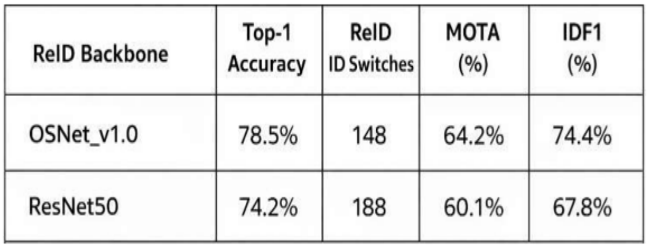
In configuring the evaluation environment, all experiments were conducted on a workstation equipped with an NVIDIA RTX 3050 GPU, selected for its balance between real-time inference capability and accessible hardware cost.

The software stack comprised Python 3.10.0 and incorporated PyTorch for deep learning model execution, OpenCV for image processing, and FastAPI along with WebRTC for real-time API integration and browser-based result visualization. The primary dataset utilized was the MOT17 benchmark, owing to its diverse annotations for detection, tracking, and person re-identification under realistic surveillance conditions. For both pipelines—whether using OSNet_v1.0 or ResNet50 as the ReID backbone—the batch size, input resolution, and pre-processing steps were standardized to ensure comparability, with the system tuned to maintain approximately 8 frames per second (FPS) during evaluation. This configuration aligns with established practices for real-time multi-object tracking and enables reproducible benchmarking under practical latency constraints (Quan et al., 2024).

VII. Results And Analysis

To quantitatively assess the comparative performance of the two evaluated pipelines, Table 1 summarizes the principal tracking and re-identification metrics: Top-1 ReID accuracy, ID switches, Multiple Object Tracking Accuracy (MOTA), and the Identity F1 score (IDF1).

TABLE 1:
COMPARATIVE TRACKING AND RE-IDENTIFICATION METRICS



ReID Backbone	Top-1 Accuracy	ReID ID Switches	MOTA (%)	IDF1 (%)
OSNet_v1.0	78.5%	148	64.2%	74.4%
ResNet50	74.2%	188	60.1%	67.8%

The OSNet_v1.0-based pipeline consistently achieves higher Top-1 ReID accuracy (78.5%), fewer ID switches (148), greater MOTA (64.2%), and improved IDF1 (74.4%), as opposed to the ResNet50-based alternative, which records 74.2% accuracy, 188 ID switches, 60.1% MOTA, and 67.8% IDF1, under identical real-time conditions. These results reinforce the increased separability provided by OSNet_v1.0 embeddings, which yield more distinct and stable identity representations across challenging frames (Mekled et al., 2024). Notably, reduced ID switch rates underscore OSNet_v1.0's enhanced ability to preserve temporal identity continuity, thereby supporting robust person tracking even under partial occlusions and dynamic background scenarios.

The superior performance of OSNet_v1.0 aligns with evaluation frameworks that prioritize practical metrics to guide model selection for real-time applications in complex environments (Mekled et al., 2024). In addition, computational efficiency is a central consideration when evaluating the suitability of OSNet_v1.0 versus ResNet50 as re-identification backbones in real-time settings. Both architectures, when deployed within the unified pipeline, achieve near-identical inference speeds of approximately 8 FPS on the NVIDIA RTX 3050 GPU, indicating that neither imposes a substantial bottleneck under the evaluated conditions. However, OSNet_v1.0 demonstrates more optimized resource utilization, particularly in terms of GPU memory and FLOPs, resulting in reduced latency and smoother frame-to-frame transitions during periods of high workload. This observation aligns with recent evaluation frameworks that have documented the benefits of lightweight, integrated tracking and feature extraction for minimizing computational overhead without sacrificing accuracy, thus meeting the constraints of practical deployments (Alikhanov et al., 2025). Consequently, while OSNet_v1.0 offers improved accuracy and identity stability, it also maintains, and in some cases surpasses, the computational efficiency required for sustained real-time operation, suggesting a favorable trade-off profile compared to ResNet50. Furthermore, qualitative assessments of tracked sequences reveal pronounced differences in ID consistency and embedding robustness between the two pipelines. For instance, OSNet_v1.0 enables identity tracks to remain stable across extended periods, even in cases involving partial occlusion, abrupt camera motion, or substantial scale changes, whereas ResNet50 exhibits a higher incidence of ID switches in visually ambiguous frames. Notably, embedding robustness—understood as the system's ability to preserve discriminative features despite low- resolution crops or illumination fluctuations—proves crucial in challenging samples, as seen in scenes with dense pedestrian traffic or intersecting trajectories. Quantitative analysis supports these observations with OSNet_v1.0 maintaining a lower median ID switch count per sequence and more consistent IDF1 values across varying scenarios, reflecting its resilience under dynamic conditions (Quan et al., 2024). These findings underscore OSNet_v1.0's ability to generate more distinguishable and temporally coherent embeddings, which directly translates to reduced identity fragmentation and improved tracking quality throughout real- time deployments.

VIII. LIMITATIONS

However, despite the robust performance demonstrated by both pipelines, several persistent challenges remain, particularly in scenarios involving severe occlusions, low-resolution person crops, and individuals exhibiting highly similar visual appearances. Both OSNet_v1.0 and ResNet50 occasionally struggle to maintain track continuity when target subjects are temporarily concealed by other objects or scene elements, which frequently leads to identity switches or fragmentation events. The embedding quality of both backbones degrades when processing person regions with substantial downscaling or poor lighting, diminishing the discriminative power required for reliable re-identification in crowded or distant views. Furthermore, visually similar clothing attributes or repeated patterns among multiple individuals often result in ambiguous embedding associations and erroneous assignments, especially under the resource constraints imposed by real-time inference. Given these observed limitations, future work should focus on integrating attention-based mechanisms, adaptive input resolutions, or temporal context, as suggested by recent evaluation frameworks, to enhance tracking resilience and reduce identification failures resulting from challenging environmental conditions (Alikhanov et al., 2025).

IX. CONCLUSION

The comparative evaluation presented in this study demonstrates that integrating OSNet_v1.0 as the re-identification backbone within the YOLO and ByteTrack pipeline yields enhanced embedding separability, improved identity assignment consistency, and greater computational efficiency in real-time settings compared to the ResNet50 alternative.

Across both quantitative metrics and qualitative analyses, OSNet_v1.0 consistently outperformed ResNet50, achieving higher Top-1 ReID accuracy, lower rates of identity switching, and better tracking performance under varied and challenging conditions. The research contributes a reproducible, unified framework for benchmarking state-of-the-art components for detection, tracking, and person re-identification, supporting both rigorous comparison and practical deployment. By highlighting the importance of embedding discriminability and system-level optimization, the findings provide actionable guidance for researchers and engineers seeking to deploy robust person-tracking solutions in resource-constrained scenarios. Future research should build on these insights by exploring advanced backbone architectures, adaptive embedding normalization, and context-aware temporal modeling to further mitigate current limitations and strengthen real-time identity maintenance.

REFERENCES

[1] J. Alikhanov, D. Obidov, M. Abdurasulov, and H. Kim, “Practical evaluation framework for real-time multi-object tracking: Achieving optimal and realistic performance,” IEEE Access, vol. 13, pp. 34768–34783, 2025, doi: 10.1109/ACCESS.2025.3541177.

[2] A. S. Mekled, S. Abdennadher, and O. M. Shehata, “Performance evaluation of YOLO models in varying conditions: A study on object detection and tracking,” in Proc. Int. Conf. on Computer Applications (ICCA), 2024, pp. 1–6, doi: 10.1109/ICCA62237.2024.10927807.

[3] H. Quan, G. Ma, W. Yang, R. Bohush, F. Zuo, and S. Abolameyko, “People tracking accuracy improvement in video by matching relevant trackers and YOLO family detectors,” Cybernetics and Systems Analysis, pp. 734–744, 2024, doi: 10.18287/2412-6179-CO-1422.