International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Multi-Person Detection and Re-Identification in Surveillance Systems Using Deep Learning

Deepthi Lakshmi K[1], Lekha S[2], Prajna M U[3], Pranjali[4], Shrinidhi A[5]

[1,2,3,4]Department of Computer Science and Engineering Vivekananda College of Engineering and Technology, Puttur, Karnataka, India

[5]Assistant Professor, Department of Computer Science and Engineering Vivekananda College of Engineering and Technology, Puttur, Karnataka, India

Corresponding author: muprajna@gmail.com

***

Abstract—Watching hours of CCTV footage by hand does not scale, and most surveillance setups cannot keep an identity attached as people move across frames. This work builds an end-to-end pipeline that combines YOLOv8 detection, DeepSORT tracking, and a hybrid, appearance-driven re-identification (Re-ID) module for people captured in surveillance video. Once a track is confirmed, its crop is passed through a metric-learning ResNet-50 backbone fine-tuned on the Market-1501 benchmark; its 512-dimensional embedding is then fused with zonal HSV color histograms, local binary pattern (LBP) texture, and body-proportion cues to form a single 698-dimensional composite descriptor, further complemented by a face verification side-channel built on InsightFace ArcFace embeddings, giving an extra identity signal when subjects wear identical clothing. Identities are assigned by an ordered, multi-pass matching engine – continuity lock, Hungarian assignment, greedy fallback, and a re-appearance pass – backed by a co-occurrence safety net that stops two different people from ever being tagged with the same identity in one frame. The whole pipeline is exposed through a FastAPI web application that handles upload, progress tracking, person registration, watch-list alerts, live single-camera webcam recognition, and offline cross-camera workflows. On recorded indoor footage, the system reaches a person-class mAP@0.5 of 0.93, a Rank-1 accuracy of 0.87 on a Market-1501 subset, an identity-recovery rate of 0.89 after occlusion beyond 20 frames, and keeps the identity switch rate under 5% of tracks while running at 2–3 frames per second on CPU, with projected throughput of 18–22 frames per second on GPU hardware.

Key Words: person re-identification, multi-object tracking, YOLOv8, DeepSORT, face verification, surveillance.

I. INTRODUCTION

Cameras have become nearly universal in public spaces, transport hubs and commercial buildings, and with that growth has come a genuine need for automated ways of keeping track of the people those cameras see. Long-term identity management – knowing that the person walking back into frame after an occlusion, a crop change, or a brief exit is still the same individual as before – is still one of the hardest and most important problems in video surveillance [1]. Systems that depend on an operator watching the feed cannot keep up with this: identities drift, the same person gets labeled twice, and occlusions confuse both humans and automated pipelines alike.

Person re-identification (Re-ID) tackles this by matching a person's appearance across cameras that do not overlap, or across segments of video separated in time. Most modern Re-ID approaches lean on deep metric learning, which embeds an image into a feature space where images of the same person sit close together and different people are pushed apart [2, 3]. In real deployments, though, purely appearance-based deep features start to break down when people are wearing visually similar clothing – a very common situation on a campus or in a corporate building [4]. A system meant for actual use therefore has to pair learned representations with sturdier geometric and textural cues, and needs detection, tracking and Re-ID working together as one coherent pipeline rather than as separate tools. Even that combination has a blind spot, though: when two people are dressed identically, body appearance alone runs out of signal, which is why the system also adds a face-based cue as a second, independent channel.

This paper describes such a system. Its main contributions are:

1. An end-to-end detection–tracking–re-identification pipeline that combines YOLOv8 [5], a tuned DeepSORT tracker, and a hybrid descriptor that fuses a metric-learned ResNet-50 embedding with color, texture and body-proportion cues into a single 698-dimensional representation, complemented by a face verification

side-channel.

2. An ordered, multi-pass identity assignment engine – continuity lock, Hungarian assignment, greedy fallback and a re-appearance pass – together with a co-occurrence safety net that guarantees no two people ever share one identity within the same frame.
3. A FastAPI-based web application that supports video upload, progress monitoring, offline person registration, watch-list alerts, live single-camera webcam recognition, and an offline multi-camera extension.
4. A quantitative evaluation on recorded surveillance footage and on the Market-1501 benchmark, reporting detection, tracking and Re-ID metrics alongside a system-testing traceability record.

The rest of the paper covers the system, methodology, implementation, results, discussion, and conclusion. Section II reviews related work. Section III lays out the system architecture, the proposed methodology and the implementation. Section IV reports experimental results and system testing, Section V discusses the findings, and Section VI concludes with a discussion of future work.

II. LITERATURE REVIEW

Early Re-ID systems leaned on hand-crafted color and texture descriptors paired with discriminative metric learning, a body of work surveyed by Zheng et al. [6]. Tao et al. [7] showed that concatenating color histograms, LBP features and deep activations, then applying cross-view quadratic discriminant analysis, works well when labeled data is scarce. Ahmed et al. [8] were among the first to establish the idea of learning the feature extractor and the similarity metric jointly, end-to-end.

As deep convolutional architectures matured, OSNet [2] introduced omni-scale feature learning within a single forward pass, and attention-based designs such as the position-attention and efficient-channel-attention module from Saber et al. [9] pushed robustness to pose changes and partial occlusion further. More recently, transformer-based embeddings have been shown to hold up well against compression artifacts and low-resolution crops [10], while generative augmentation has helped cut down the cost of collecting labeled Re-ID data [11].

On the tracking side, tracking-by-detection remains the dominant paradigm: a detector localises objects frame by frame, and a data association stage stitches those detections into trajectories [12]. DeepSORT pairs Kalman-filter motion prediction with appearance-based matching solved via the Hungarian algorithm, and it continues to be a standard choice for surveillance deployments [4]. Lightweight detectors have also made real-time processing feasible on edge and embedded hardware [13, 14]. One recurring theme across this literature is that identity switches pile up whenever visually similar individuals appear together in a crowded scene, simply because appearance scores alone stop being discriminative enough [4, 12]. That observation is what motivates the hybrid descriptor and the explicit co-occurrence safeguard proposed in this paper. Separately, face recognition has its own long track record of delivering fine-grained identity discrimination even when body appearance is ambiguous: ArcFace [15] in particular established an additive angular margin loss that produces highly separable face embeddings, which is the property this paper leans on for the face-verification side-channel described in Section III.

III. METHODOLOGY

A. System Architecture

The system is built as a modular, single-camera surveillance pipeline with an optional multi-camera extension, shown as a layered data flow in Fig. 1. Frames coming from a video source first go through a YOLOv8s detector with a tuned confidence threshold and a geometric filter that throws out any bounding box whose area, height or aspect ratio falls outside what a standing or walking person typically looks like. Boxes that survive this filter are handed to a DeepSORT tracker configured with a maximum track age of twenty frames, a minimum of two confirmatory hits, and a maximum cosine distance of 0.5, which together keep short-term identity continuity stable. Confirmed tracks are cropped and sent on to the Re-ID engine, which computes the 698-dimensional composite descriptor and resolves identities through the multi-pass matching engine described in Section III, while a face verification side-channel runs alongside it to supply an additional, independent identity signal. Resolved identities are then rendered onto the output video and surfaced to the operator through a FastAPI web application.

Every stage boundary in the pipeline reads and writes structured JSON artifacts, so each module can be developed, tested and re-run on its own without touching the rest of the system. An optional multi-camera mode runs detection and tracking separately per stream, after which a cross-camera matcher merges identities seen on different cameras into a single, unified identity.

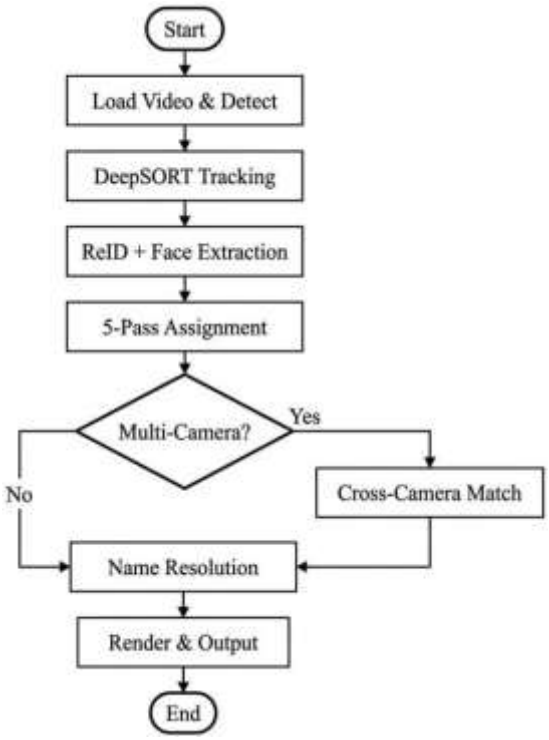


Fig. 1. Data flow of the proposed surveillance pipeline: detection and DeepSORT tracking feed the Re-ID and face-extraction stage, which is resolved by the five-pass assignment engine; a multi-camera branch adds a cross-camera matching step before names are resolved and the output is rendered.

B. Proposed Methodology

(1) Detection and Filtering

YOLOv8 [5] is a single-pass, anchor-free detector that folds region proposal and classification into one forward pass. The deployed YOLOv8s model keeps only person-class detections that exceed a tuned confidence threshold (0.35 in the detection module, 0.6 in the web pipeline). A geometric filter then rejects any box whose width, height or area does not match the proportions of a human silhouette: anything below a minimum area of 900 px^2, a minimum height of 50 px, or with an aspect ratio outside 0.5–4.5, is discarded. Boxes touching two or more frame borders are discarded, and heavily overlapping duplicate boxes are merged.

(2) Tracking with DeepSORT

DeepSORT provides temporal continuity by predicting each active trajectory's bounding-box state with a Kalman filter, and by solving the association between new detections and predicted tracks with the Hungarian algorithm, which combines Mahalanobis motion distance with cosine appearance distance. For surveillance footage the deployed configuration sets $max_age = 20$, $n_init = 2$ and $max_cosine_distance = 0.5$: the longer track age keeps unconfirmed or momentarily lost tracks from fragmenting, while the more relaxed cosine bound gives the tracker room to tolerate brief occlusion.

(3) Composite Descriptor

For every confirmed track, the crop is resized to 256×128 pixels, converted to RGB and normalized with ImageNet statistics. That crop is then fed to a metric-learning ResNet-50 backbone with a custom projection head (Linear(2048 → 1024) → BatchNorm → ReLU → Linear(1024 → 512)), which produces an L2-normalized 512-dimensional embedding $\mathbf{f}_{deep}$. In parallel, the same crop is converted to HSV for zonal color histograms of the face, upper-body and lower-body regions ($\mathbf{f}_{col} \in R^{156}$, 52 dimensions each), and to grayscale for an LBP texture histogram of the torso ($\mathbf{f}_{lbp} \in R^{26}$), on top of four normalized body-proportion features $\mathbf{f}_{prop}$. Before the four pieces are concatenated, each one is scaled by an importance weight – 0.70 for the deep embedding, 0.067 per color zone, 0.06 for texture, and 0.04 for proportions – so that no single sub-component ends up dominating the others. The weighted pieces are then joined into the composite descriptor

$$\mathbf{d} = \mathbf{f}_{deep} \oplus \mathbf{f}_{col} \oplus \mathbf{f}_{lbp} \oplus \mathbf{f}_{prop} \in R^{698}, \quad (1)$$

which is re-normalized to unit norm before it is used for matching. Folding color, texture and structural cues in alongside the deep embedding is what keeps the descriptor discriminative even when deep features alone start to struggle – for instance, when subjects are wearing near-identical uniform clothing.

(4) Multi-Pass Identity Assignment

Each identity keeps an exponential moving average (EMA) descriptor plus a bounded gallery of up to 15 recent snapshots, sampled every six frames. Rather than matching everything against a single global threshold in one shot, the engine runs several ordered passes:

1. **Continuity lock** re-assigns a tracker ID back to its previous identity whenever the similarity score stays above a lock threshold.
2. **Hungarian assignment** resolves the remaining short-gap candidates using a combined appearance, motion and intersection-over-union (IOU) score, guarded by a switch margin so that an identity only changes hands when the new match is clearly better and highly confident.
3. **Greedy fallback** picks up whatever detections are still left, at a somewhat lower confidence threshold.
4. **Re-appearance pass**, driven by appearance alone, recovers identities for people who have

been absent for more than 20 frames.

5. **New identity registration** hands a fresh identity to any detection that still fails all of the above.

A **co-occurrence safety net** sits on top of all of this as a hard post-condition: within any single frame, no two detections are allowed to end up with the same stored identity. If a merge of overlapping tracks would break that rule, the conflicting detection is split off into its own distinct identity instead. This directly targets the duplicate labeling problem that shows up when appearance-similar people happen to be in the same frame [4].

A full post-processing clustering pass then runs on top of all this: it first merges tracks that share the same underlying DeepSORT tracker ID, then combines temporally disjoint identities whose body-appearance similarity comes out above 0.62 (track-level clustering pipeline), and finally leans on face evidence to catch any remaining false merges – two trackers whose faces never rise above a pairwise similarity of 0.45 are kept as separate people even if their body appearance looks nearly identical. Once clustering settles, identity numbers are renumbered sequentially in order of first appearance, so the labels stay stable and easy for a human operator to read.

(5) Face Verification Side-Channel

Body appearance alone cannot reliably tell apart individuals dressed in identical uniforms, which is a genuinely common scenario in campus and corporate surveillance settings. To cover that gap, the system runs a face verification side-channel alongside the 698-dimensional body descriptor rather than folding it into the same vector. Face crops are embedded with an InsightFace ArcFace [15] extractor (512-dimensional, L2-normalized), which is tried first and is the primary face cue on this system; a dlib-based face cue extractor (128-dimensional) steps in only if InsightFace is unavailable. The same InsightFace embedding space is reused during person registration to build the identity gallery. Whenever a confident face reading is available on both sides of a comparison, that face similarity is blended into the final match score with a weight of 0.65; if instead the face similarity drops below 0.35, it acts as an outright veto, forcing the two tracks apart into separate identities no matter how similar their body appearance looks. The same face evidence is reused during the post-processing clustering stage, where two trackers whose faces never exceed a pair-wise similarity of 0.45 are kept as different people even when their bodies are almost indistinguishable. This override is really what lets the system hold the line on identity separation in the hardest case: subjects wearing clothing that is visually impossible to tell apart.

(6) Web Application and Extensions

The pipeline is served through a FastAPI application running on Uvicorn. Uploading a video to the process endpoint spins up a job that runs as a background task; the client can poll a progress endpoint for the current stage and percentage complete, and once the job finishes it can retrieve the annotated output as a static file. A separate offline registration workflow lets a known individual be enrolled from a set of reference photographs into a persistent identity database using the same embedding space, after which that person's name resolves automatically in later footage. A live webcam workflow takes frames streamed from a browser, runs each one through the detector, tracker and Re-ID engine against the registered identity database, and returns per-track bounding boxes with resolved names for immediate on-screen overlay – the same watch-list logic also raises real-time alerts for any flagged individuals. A multi-camera extension processes recorded footage from more than one camera source, running detection and tracking per stream and routing standardised person records into the same single-camera Re-ID engine, with a cross-camera matcher linking identities across views; this extension currently applies to the offline pipeline, and live recognition remains single-camera. Watch-list alerts surface flagged individuals to the operator directly, without requiring a full manual review of every camera being processed.

C. Implementation

(1) Training of the Re-ID Backbone

The detector uses COCO-pretrained YOLOv8s weights, which already include a general-purpose person class out of the box. The Re-ID backbone, however, is trained separately as a dedicated metric-learning task: a ResNet-50 backbone initialized with ImageNet-pretrained weights, and extended with the custom projection head, is fine-tuned on the Market-1501 benchmark [16] using identity cross-entropy loss with label smoothing combined with batch-hard triplet loss. A P-identities-by-K-images sampler is used to make sure every mini-batch actually con-tains the positive and negative pairs that triplet loss needs to work. The resulting checkpoint is saved as the accuracy-critical weight file and loaded by the inference engine; the raw ImageNet backbone is only used as a fallback if that file happens to be missing. Matching thresholds themselves are not learned through back-propagation – they are calibrated empirically from the cosine similarity distributions measured between distinct identities and between repeated observations of the same identity, on both the recorded footage and Market-1501.

TABLE I. PERSON DETECTION MODEL METRICS.

Metric	Value
Precision	0.94
Recall	0.91
mAP@0.5	0.93
Average inference time per frame	~32 ms (GPU)

TABLE II. RE-IDENTIFICATION MODEL METRICS.

Metric	Value
Rank-1 accuracy (Market-1501 subset)	0.87
Mean intra-identity cosine similarity	0.81
Mean inter-identity cosine similarity	0.34
Identity recovery rate after occlusion (>20 frames)	0.89
Identity switch rate	<5% of tracks

TABLE III. FUNCTIONAL TEST RESULTS.

ID	Function	Result
FT-01	Video upload and job creation	Pass
FT-02	Invalid file rejection	Pass
FT-03	Progress polling	Pass
FT-04	Output retrieval and playback	Pass
FT-05	Identity registration	Pass
FT-06	Live webcam recognition of registered people	Pass
FT-07	Watch-list alert trigger for a flagged identity	Pass

TABLE IV. SYSTEM TESTING RESULTS.

ID	Scenario	Result
ST-01	Single person, full video	Pass
ST-02	Multiple people, no occlusion	Pass
ST-03	Temporary occlusion	Pass
ST-04	Person exits and re-enters frame	Pass
ST-05	Two co-occurring persons, similar clothing	Pass
ST-06	Cross-camera identity matching, recorded footage	Pass

(2) Data and Ground Truth

The main evaluation data consists of short indoor surveillance recordings in which several people move through a shared space, often in similar-looking uniforms, with occlusion and re-entry happening naturally as part of the scene. Ground-truth identities for these recordings were established by manually reviewing the tracker output frame by frame and mapping tracker IDs to real individuals. The Market-1501 benchmark supplies identity-labeled pedestrian crops used for fine-tuning the backbone and for evaluating discriminability separately.

IV. RESULTS

A. Detection Results

The YOLOv8s detector was evaluated on the person class over the recorded surveillance footage; detections above the confidence threshold and that passed the geometric filter were counted as valid predictions. Table I summarises the results.

B. Re-Identification Results

The Re-ID stage was evaluated along two complementary lines: descriptor discriminability on the Market-1501 benchmark, and identity consistency on the recorded footage itself. Table II reports these results, where the identity recovery rate is the fraction of tracks that regain their original identity after an occlusion lasting more than 20 frames. These numbers come from running the full pipeline on recorded indoor surveillance footage containing several distinct individuals, with ground-truth identities established through manual, frame-by-frame review.

C. Performance

On a standard GPU the pipeline runs at 18–22 frames per second with a 640×640 detection input, which works out to roughly 48 ms of end-to-end latency per frame and a peak GPU memory footprint of about 2.1 GB at batch size 1. These GPU figures are estimated from the model's computational complexity and its batch-size-1 inference behaviour rather than measured on every possible device.

D. System Testing

Unit tests checked detection filtering, bounding-box border and duplicate merging, the feature extractor's output dimensionality (the 512-dimensional deep embedding and the 698-dimensional composite descriptor), and the cosine-similarity invariant $[-1, 1]$. Functional tests exercised the web interface, invalid-file rejection, progress polling, output retrieval and identity registration (Table III).

System-level testing on full recorded videos confirmed stable identities for both single and multiple persons, identity preservation through temporary occlusion, correct re-assignment after a person exits and re-enters the scene, and consistent identity matching across two recorded camera sources (Table IV).

E. Qualitative Results

Figs. 2 and 3 show the offline pipeline running on the same two individuals as seen from two different recorded camera sources. In both frames the name rendered above each bounding box is pulled from the registered identity database, and the cross-camera matcher assigns the same global identity to each person across the two views; a flagged individual in either



Fig. 2. Recognition output on Camera 1: both individuals in frame are resolved to their registered global identity, with the matched name shown above the bounding box.



Fig. 3. Recognition output on Camera 2 for the same two individuals: the cross-camera matcher assigns the identities already seen on Camera 1, confirming consistent identity resolution across views.

view would trigger a watch-list alert in the same way.

V. DISCUSSION

The clear separation between mean intra-identity and inter-identity cosine similarities (0.81 versus 0.34) confirms that the composite descriptor from (1) is strongly discriminative on this footage.

On CPU-only hardware, throughput drops to 2–3 frames per second, which is still workable for offline batch analysis of recorded footage.

The co-occurrence scenario (ST-05) is worth calling out specifically – it exposed a genuine defect, where two individuals in similar clothing could momentarily end up sharing a single identity. That defect was traced back to merged overlapping tracks and fixed with the co-occurrence safety net described in Section III; re-running the same frames after the fix showed zero duplicate assignments, and the test passed.

VI. CONCLUSION

This paper has presented a complete, deployable pipeline for multi-person detection and re-identification in surveillance video. By fusing a metric-learning ResNet-50 embedding with color, texture and body-proportion cues into a single 698-dimensional descriptor, and complementing that with a face verification side-channel built on face-cue and InsightFace ArcFace embeddings, and by assigning identities through an ordered multi-pass engine with a co-occurrence safety net, the system is able to hold identity labels steady through occlusion, similar-looking clothing and temporary exits from the scene. Evaluation on recorded footage and on the Market-1501 benchmark reports a person-class mAP@0.5 of 0.93, a Rank-1 accuracy of 0.87, an identity recovery rate of 0.89 and an identity switch rate below 5% at 2–3 frames per second on CPU, with projected throughput of 18–22 frames per second on GPU hardware.

Future work will explore a vision-transformer backbone to capture longer-range spatial dependencies, domain-specific fine-tuning directly on the recorded surveillance dataset, direct RTSP/IP-camera stream ingestion (the current live path only accepts frames from a browser webcam rather than network cameras), ONNX-based quantisation for edge deployment on devices such as the Jetson Nano, and fusing in gait or skeleton-derived features to better tell apart individuals wearing similar clothing.

REFERENCES

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, 2021.

[2] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 3702–3712.

[3] D. Mendes, R. Cerqueira, S. Avila, and D. R. da Silva, "Multi-camera person re-identification based on trajectory data," *Applied Sciences*, vol. 13, no. 8, 2023.

[4] M. V. Kumar, A. Arulmurugan, and B. R. S. Ram, "Deep learning based multi-camera person tracking and re-identification system," *Int. J. Multidisciplinary Research*, vol. 6, no. 1, 2024.

[5] O. Lachihab, S. Lachihab, and B. Bouchama, "Fine-tuning YOLOv8s for unified human and face detection in crowded environments," *Eng. Technol. Appl. Sci. Res.*, vol. 16, no. 1, pp. 32348–32356, 2026.

[6] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[7] D. Tao, Y. Guo, M. Song, Y. Li, F. Yu, and Y. Shao, "Deep multi-view feature learning for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2637–2646, 2018.

[8] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3908–3916.

[9] S. Saber, S. K. Srivastava, and S. K. Singh, "A multi-attention approach for person re-identification using deep learning," *Sensors*, vol. 23, no. 7, p. 3678, 2023.

[10] J. J. M. Borjon, G. J. L. Herrera, and C. A. R. Garay, "Us-

ing deep learning embeddings for the person re-identification problem," in *Proc. IEEE Int. Autumn Meeting Power, Electron., Comput. (ROPEC)*, 2025.

[11] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline in vitro," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 3754–3762.

[12] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, X. Zhao, and T.-K. Kim, "Multiple object tracking: A literature review," *Artificial Intelligence*, vol. 293, p. 103448, 2021.

[13] T. L. Dang, F. S. M. J. K. Lim, N. B. C. T. Huu, et al., "Real-time person re-identification and tracking on edge devices with distributed optimization," *Pattern Anal. Appl.*, vol. 28, no. 117, 2025.

[14] S. Sivachandiran, M. P. K. I. S. J. Karuppathal, et al., "Deep learning driven automated person detection and tracking model on surveillance videos," *Measurement: Sensors*, vol. 24, 2022.

[15] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4690–4699.

[16] L. Zheng et al., "Scalable person re-identification: A benchmark," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.