AI-Based Missing Person Identification Using YOLO and Deep Facial Embeddings

Student Name
Department of Computer Science and Engineering
University / College Name
Email: student@email.com

***Abstract*—Missing person identification using surveillance imagery remains a challenging problem due to adverse visual conditions, limited availability of reference images, and strict privacy constraints surrounding real-world data. This paper presents a hybrid deep-learning framework for missing person identification that combines YOLO-based person detection with deep face embedding models, specifically FaceNet and ArcFace. To address ethical and privacy limitations, a manually curated composite synthetic dataset is constructed by combining publicly available crowd and in-the-wild face datasets with additional curated images to realistically emulate CCTV conditions, including low illumination, occlusions, visually similar individuals, accessories, and clothing-matched decoys, while restricting each identity to only three to four reference images. The proposed system is evaluated across five YOLO variants (v8–v12) and a wide range of cosine similarity thresholds to analyze detection sensitivity, false positive behavior, and overall identification robustness. Extensive experiments conducted on 780 group images demonstrate that the YOLO–ArcFace pipeline achieves superior performance, reaching a peak identification accuracy of 97.50**

***Index Terms*—Missing person identification, YOLO, facial recognition, deep learning, surveillance systems, privacy-preserving AI**

I. Introduction

The rise in the development of closed-circuit television (CCTV) networks in both public and semi-public areas has presented a unique challenge for the automation of human identification. Public transport stations, campuses, market places, and even streets have been placed under constant surveillance. This results in the collection of huge datasets. In the midst of this automation explosion, the search for the lost continues to be a manual process. This process is not only inefficient but also filled with errors when dealing with crowded environments and sub-standard video quality coupled with the possibility of the missing person not being prominent enough.

Face recognition automation provides a very natural solution to this issue, but their use in practical surveillance settings is challenging in various aspects and specifications. Unlike carefully managed face recognition datasets, surveillance images are typical representatives of uncontrolled lighting conditions, blur induced by motion or camera shake, partial occlusions, turned-away views, and compressed formats in most cases. Moreover, in cases related to lost persons, only a handful of sample images are available in most cases, usually captured in various environments profoundly differed from those present in typical surveillance videos.

In addition to the limitations posed by technology, the application of real missing person data for algorithm development also raises many privacy and ethical issues. The data of biometric information of vulnerable people cannot be shared or published freely, and therefore it is difficult to build research data that is open and reproducible. As a result, many existing studies evaluate their methods on idealized or unrelated face datasets, which fail to capture the operational realities of surveillance-based identification. This gap between academic evaluation and real-world deployment motivates the need for privacy-preserving experimental frameworks that still reflect realistic surveillance conditions.

In this work, we address these challenges by proposing a hybrid deep-learning framework that integrates person detection and face recognition for missing-person identification under constrained and privacy-aware settings. The system first localizes all individuals in a scene using YOLO, a state-of-the-art real-time object detector, and then performs identity matching using deep face embeddings. Two alternative recognition backends are investigated: FaceNet, a widely used metric-learning approach based on triplet loss, and ArcFace, a more recent angular-margin-based embedding model known for its strong inter-class separability. By comparing these two paradigms within the same detection framework, we analyze how embedding discriminability affects performance in realistic surveillance scenarios.

To ensure ethical compliance while maintaining realism, a synthetic missing-person dataset is manually constructed to emulate CCTV-like conditions. The dataset deliberately includes low-light imagery, occlusions, visually similar individuals, and clothing-matched decoys, while restricting each identity to only three to four reference images, closely mirroring real investigative constraints. Using this dataset, we perform a comprehensive evaluation across five YOLO versions and eleven similarity thresholds, allowing us to study both detection sensitivity and false-positive behavior in a principled manner.

The main contributions of this paper are: i) an privacy-preserving experimental framework that faithfully reflects surveillance conditions for the identification of missing persons; ii) a systematic comparison between YOLO–FaceNet and YOLO–ArcFace pipelines across several operating points

and variants of detectors; and iii) an empirical analysis on how threshold selection and embedding choices affect accuracy and falsealarm rates. While the current study is focused on frame-level identification, the framework is designed to be easily extendable toward continuous video-based tracking and re-identification, enabling the persistent monitoring across CCTV footage.

II. Literature Survey

Early works on missing person identification primarily relied on conventional facial recognition pipelines and controlled image settings. Traditional approaches using handcrafted features and basic CNN-based recognition demonstrated feasibility but lacked robustness under real-world surveillance conditions. Systems proposed by Pratyush Raj et al. [6] and Rayabarapu Alekya et al. [10] employed standard face detection followed by embedding extraction or VGG-Face with SVM classification. While these methods achieved acceptable accuracy for frontal and well-lit images, they performed poorly when exposed to occlusions, non-frontal poses, aging effects, and crowded scenes, mainly due to the absence of an explicit person detection stage and limited discriminative capability of the feature representations.

However, to counter the weaknesses of only having recognition in images, many research works tried to integrate facial recognition using deep learning approaches along with automated matching concepts. In fact, D. Sattibabu et al. [5] developed a CNN model which tried to match facial images to centralized databases, thereby decreasing the need for manual detection in investigations. But at the same time, their model failed to identify facial images from surveillance cameras, which had a blurred and reduced resolution. Likewise, Mari Selvan et al. [1] designed an automated image recognition system powered by AI, focusing on automation and smart surveillance. But their paper failed to judge the efficiency of their system in adverse situations like insufficient illumination,-obstructed facial features, or an insufficient number of recognition images.

With the advancement of object detection techniques, YOLO-based approaches became prominent for surveillance-oriented applications. Ha Yeon Kim et al. [2] demonstrated the feasibility of integrating YOLOv3 for human detection followed by facial recognition to locate missing persons in real-time surveillance footage. While this work established the importance of person detection in crowded scenes, its performance degraded significantly under low-light conditions and heavy occlusions, and the study lacked detailed false-positive analysis. To address partial occlusions, G. Pandiya Rajan et al. [11] combined YOLO-based person detection with a custom CNN for feature extraction, achieving improved recognition when faces were partially visible. However, the system failed under extreme occlusions and did not explore similarity threshold tuning or false alarm control, which are critical for operational deployment. Several researchers explored enhancements to facial recognition robustness through preprocessing and auxiliary models. Muhammad Fadhly Sani et al. [7] integrated YOLO with FaceNet and super-resolution models to improve recognition from poor-quality images. Although the approach improved detection accuracy, the additional image enhancement introduced significant computational overhead, limiting real-time applicability. Moreover, the work did not thoroughly analyze false-positive behavior, which is particularly important in missing person identification where incorrect matches may have serious consequences.

Even cloud-based and secure infrastructure solutions have also been explored. Hemadharshini et al. in [3] presented an assisted search solution on Amazon Web Services using Rekognition and Lambda functions for scalable identification and alerts.Similarly, Abid Faisal Ayon et al. in [8]concentrated on secure and encrypted data exchange between agencies for locating missing persons. Even though the solutions include scalable deployment and data security, they are mostly internet-dependent and vendor-based technologies and fail to consider the challenges of vulnerable image quality from CCTV cameras in addition to ethical factors in using biometric information.

Some recent works have proved the relevance of having more discriminative and robust embedding learning models. Jing Zhang et al. [9] presented an extensive survey of face recognition methods, proving that the adoption of embedding learning outperformed previous approaches regarding accuracy and robustness. Nevertheless, this work remained mostly theoretical and was not validated experimentally in more realistic surveillance settings. Khader Basha et al. [4] addressed the issue of identification of missing children using long-term identification solutions and age progresisons, but it remained vulnerable in disguise, accessories, and crowded settings.

It is clear from the existing literature that, although deep learning and YOLO-based detection have made tremendous progress in the area of missing person identification, most existing systems have one or more important drawbacks, such as vulnerability to distortions in the surveillance environment, a lack of analysis in the false positives, a requirement for actual biometric information, and an inadequate analysis in the reference constraints. There is, therefore, a need for a privacy-preserving and surveillance-realistic platform with in-depth analysis in the area of missing person identification.

III. Methodology

In this section, we elaborate on the complete experimental setting, which includes system architecture, pipelines of face recognition, dataset, and evaluation setting. We aim to make sure that our solution can be replicated, and our experimental results are properly grounded.

A. System Architecture

The proposed system operates in two stages. First, a person detection model identifies all individuals present in a surveillance frame. Second, a face recognition module extracts facial embeddings from each detected person and matches them against a gallery of known missing persons. The processing pipeline for a given image is as follows: 1. The

input surveillance image is processed by a YOLO model to detect all persons. 2. Each detected person bounding box is extracted as a region of interest. 3. A face detector is applied within each region to localize faces. 4. For each detected face, a deep embedding is generated. 5. Cosine similarity is computed between the embedding and all gallery embeddings. 6. The identity with the highest similarity is selected as the candidate match, subject to a similarity threshold. Two alternative recognition backends are evaluated within this same detection framework: FaceNet and ArcFace.

B. YOLO-Based Person Detection

There are five YOLO models employed in this research: YOLOv8n, YOLOv9s, YOLOv10n, YOLOv11n, and YOLOv12n. All models are applied to only the "person" class for detecting people in each image. They act as the first-stage detector to focus on people in the image. Applying various YOLO models enables us to study how detection accuracy and the stability of bounding box detection contribute to face recognition.

C. Face Recognition Pipelines

1) YOLO–ArcFace Pipeline: The ArcFace-based pipeline uses the InsightFace buffalo-l model, which integrates a RetinaFace/SCRFD-style face detector, facial landmark alignment, and ArcFace-based embedding extraction into a single module. For each detected person region, faces are localized and converted into 512-dimensional embeddings. Cosine similarity is then computed between these embeddings and the gallery of known identities. Due to the strong class separation properties of ArcFace, only a minimal acceptance threshold (0.25) is required to suppress noise.

2) YOLO–FaceNet Pipeline: The FaceNet pipeline uses MTCNN to detect faces inside each person region, followed by InceptionResNetV1 (pretrained on VGGFace2) to generate 512-dimensional embeddings. To handle poor lighting and low contrast typical of surveillance imagery, each face crop is processed in two forms: the original image and an enhanced version produced using CLAHE-based contrast normalization, brightness and contrast adjustment, and sharpening. Both versions are evaluated and the best similarity score is selected. This multi-version strategy improves robustness under adverse conditions but also increases the risk of false matches, making careful threshold selection essential.

D. Dataset Construction

First, they have combined multiple public sources and manually curated images to create a composite synthetic surveillance dataset to build the realistic evaluation set with privacy preservation in mind. Negative samples came from Group and Crowd Images, specifically from the CrowdHuman and WiderPerson datasets that have highly variable pose, scale, and occlusion. Positive group images known to have the identity embedded came from the CelebA In-The-Wild dataset. Such a dataset provided controlled conditions for inserting target individuals into crowd scenes while avoiding real missing-person data. Additional publicly available images were manually selected from the internet and integrated with the rest of the existing dataset to offer a better realism situation, with great diversity in lighting conditions, camera quality, accessories, and background clutter. The final set was designed to simulate real-world CCTV conditions such that there is low lighting, some occlusions, similar-looking individuals, and ambiguity based on clothing. The strategy here adopts a balanced approach towards data construction and maintains privacy and ethical considerations while being able to fully test the proposed system under real-world conditions.

TABLE I
EXPERIMENTAL DATASET PARAMETERS

Parameter	**Value**
Number of known persons	4
Reference images per person	3–4
Total group images	780
True images (contain target)	240
False images (no target)	540

The group images include low-light scenes, visually similar individuals, occlusions, accessories such as hats, and clothing-matched decoys. This design intentionally increases ambiguity, making the task significantly harder than standard face recognition benchmarks.

E. Similarity Threshold Protocol

The list of possible values for in the FaceNet pipeline would be comprised of 0.65, 0.60, 0.55, 0.50, 0.45, 0.40, 0.35, 0.30, 0.25, 0.20, 0.15. With a high value of the tolerance level, it would be efficient in terms of FP. But it would result in detection failure for partially occluded faces. It would result in both detection failure and FP for a low value of the tolerance level. As the difference in the distribution in the ArcFace pipeline is large in that case, it would not need to perform an exhaustive search. The minimum value for the tolerance level would be 0.25.

F. Evaluation Metrics

There are two forms of performance on which assessments are made: • True Positive Rate (TPR) = Percentage of images containing a target person that are successfully paired. • False Positive Rate (FPR): The percentage of images with no target person who were incorrectly matched. For ArcFace, mAP@0.5 is further provided to give a comprehensive assessment on recognition quality on different versions of YOLO.

G. Experimental Configuration

All these experiments are carried out within Google Colab environments equipped with an NVIDIA T4 GPU. The same testing set and testing procedure are applied for all versions of YOLO and recognition pipelines. This approach allows for a controlled investigation of the effects of detection quality, embedding models, and similarity thresholds on the performance of missing person retrieval.

IV. Expected Results

This section presents a detailed quantitative evaluation of the proposed YOLO–FaceNet and YOLO–ArcFace pipelines on the privacy-preserving synthetic surveillance dataset. The experiments were designed to measure not only identification accuracy but also false positive behavior under different YOLO versions and similarity thresholds, which is critical for safety-critical missing-person search.

A. FaceNet False-Positive Analysis

Table II reports the percentage of false images (540 negatives) that were incorrectly identified as containing a known person for different YOLO versions and similarity thresholds. This table directly measures how aggressively the FaceNet pipeline produces false alarms when the acceptance threshold is lowered.

Interpretation. At high thresholds (0.60), the FaceNet pipeline produces almost no false positives but also misses many true matches. As decreases, recall improves but false positives increase rapidly, stabilizing around 1.5–1.7

B. FaceNet True-Positive Performance

Table III reports the true-positive rate on 240 images that contain at least one known individual.

Interpretation. Reducing is always beneficial for recall for all versions of YOLO. But to judge the effectiveness of increase in true positives, the rise in false positives, as shown in Table IV-A, has to be considered. For YOLO11n, the value of = 0.30 is optimal, and this results in maximum recalls of 95.42

C. ArcFace Performance Across YOLO Variants

ArcFace results are summarized using mAP@0.5 on the 240 positive images and false positive rate on the 540 negative images.

Interpretation. ArcFace consistently achieves very high identification accuracy across all YOLO variants, with YOLOv9s and YOLOv12n producing zero false positives on all 540 negative images. This indicates that ArcFace embeddings are highly discriminative even under visually confusing CCTV-like conditions.

D. Threshold–Accuracy Relationship for FaceNet

Figure 1 (Threshold vs Detection Accuracy) should be placed immediately after Tables II and III. This plot shows how FaceNet accuracy increases as decreases, illustrating the recall–precision trade-off inherent to FaceNet embeddings. Interpretation. The curve reveals that accuracy improves steadily as is lowered, but Table II shows that this gain is accompanied by a growing false-positive rate. The knee of the curve around = 0.30 corresponds to the best practical operating point.

E. ROC Analysis

Figure 2 (ROC Curve) should be placed after Tables IV and V. The ROC curve plots false positive rate versus true positive rate for FaceNet (YOLO11n) across thresholds, with the ArcFace operating point overlaid. Interpretation. The FaceNet curve exhibits a gradual trade-off between recall and false alarms, whereas ArcFace appears as a point in the upper-left corner, indicating simultaneously high recall and near-zero false positives. This visually demonstrates ArcFace's superior separability.

F. Biometric Operating Point Comparison

Figure 3 (ArcFace vs FaceNet Operating Point) should be placed at the end of this section. This plot directly compares the best FaceNet configuration (YOLO11n, = 0.30) against the best ArcFace configuration (YOLOv9s). Interpretation. ArcFace obtains a true-positive rate of 97.50

V. Discussion

The experiments show that the choice of face embedding model has a substantially greater impact on missing-person identification performance than the choice of YOLO detector. While all YOLO variants provided sufficiently accurate person localization, the reliability of identity matching was dominated by how well the embedding space separated different individuals under surveillance-specific distortions.

FaceNet exhibits a strong dependence on threshold selection. As demonstrated by the threshold–accuracy relationship and the ROC curve, higher recall is achieved only by accepting an increasing number of false matches. This reflects the limited inter-class separation of FaceNet embeddings when faces are captured under low light, occlusion, or from non-frontal angles. Even after applying image enhancement and multi-version evaluation, visually similar individuals and clothing-matched decoys frequently produce overlapping similarity scores. As a result, FaceNet requires careful and context-dependent calibration of the similarity threshold to remain operationally usable.

ArcFace behaves fundamentally differently. Its angular-margin loss produces compact and well-separated identity clusters, which translates into both high recall and extremely low false-positive rates without the need for aggressive threshold tuning. The operating-point comparison places ArcFace in a region of the ROC space that FaceNet cannot reach, indicating that improved discrimination rather than additional preprocessing is the key factor in robust surveillance identification.

The limited influence of YOLO version on final recognition accuracy suggests that, beyond a baseline level of person detection quality, further improvements in missing-person identification are primarily driven by embedding quality rather than detector sophistication. This finding is important for deployment scenarios, where computational constraints may favor smaller YOLO models while still maintaining high recognition reliability when paired with a strong embedding model. The outcome of the results presented here shows the

TABLE II
FALSE POSITIVE RATE (%) FOR FACENET ON 540 NEGATIVE IMAGES

YOLO Version	τ=0.65	0.60	0.55	0.50	0.45	0.40	0.35	0.30	0.25	0.20	0.15
YOLOv8n	0.000	0.000	0.370	0.926	1.111	1.111	1.481	1.481	1.667	1.667	1.667
YOLOv9s	0.000	0.000	0.185	0.556	0.926	1.296	1.296	1.667	1.667	1.667	1.667
YOLOv10n	0.000	0.000	0.370	0.556	0.926	0.926	1.111	1.481	1.667	1.667	1.667
YOLO11n	0.000	0.000	0.370	0.741	1.111	1.111	1.481	1.667	1.667	1.667	1.667
YOLO12n	0.000	0.000	0.185	0.370	1.111	1.111	1.296	1.481	1.667	1.667	1.667

TABLE III
TRUE POSITIVE RATE (%) FOR FACENET ON 240 POSITIVE IMAGES

YOLO Version	τ=0.65	0.60	0.55	0.50	0.45	0.40	0.35	0.30	0.25	0.20	0.15
YOLOv8n	88.75	89.17	89.58	90.00	91.67	93.75	95.83	97.08	97.08	97.50	97.92
YOLOv9s	89.58	89.58	90.00	90.42	92.08	92.92	95.00	96.25	97.08	98.33	98.75
YOLOv10n	89.17	89.17	90.00	90.42	91.25	92.50	92.50	94.17	95.42	97.08	97.92
YOLO11n	89.17	89.58	89.58	90.83	92.92	93.33	95.00	95.42	96.67	97.92	98.33
YOLO12n	89.17	89.17	89.58	90.42	91.25	93.33	94.58	96.25	96.67	98.33	98.75

TABLE IV
ARCFACE TRUE-POSITIVE PERFORMANCE

YOLO Version	**mAP@0.5 (%)**
YOLOv8n	97.08
YOLOv9s	97.50
YOLOv10n	96.67
YOLOv11n	97.50
YOLOv12n	97.50

TABLE V
ARCFACE FALSE POSITIVE RATE

YOLO Version	**False Positive (%)**
YOLOv8n	0.370
YOLOv9s	0.000
YOLOv10n	0.185
YOLOv11n	0.185
YOLOv12n	0.000

capability of ArcFace to provide a more stable and secure platform for missing-person detection based on surveillance.

VI. CONCLUSION

This work has offered a privacy-preserving solution for missing person identification in surveillance videos through the integration of YOLO-based person detection and deep face embeddings. To simulate real-world scenarios, a realistic synthetic dataset was created, incorporating factors like low light, partial occlusions, similar persons, and few reference samples, making a valid comparison possible without requiring the use of actual data. With multiple versions of YOLO and different similarity margins, this study clearly established that the YOLO-ArcFace system outperforms the YOLO-FaceNet system in terms of identification accuracy and, more importantly, has lower false positives. This goes to prove that the discriminability of embeddings has a predominant effect on identification for surveillance, which transcends improvements in detection.

Future work will extend the proposed frame-based system to continuous CCTV video by incorporating multi-object tracking methods such as DeepSORT, enabling persistent tracking of detected individuals across time and camera views. This transition from isolated frames to temporal tracking is expected to further improve reliability by aggregating evidence over multiple observations, bringing the framework closer to practical deployment in real surveillance environments.

REFERENCES

[1] R. Mari Selvan, "AI-Powered Image-based Missing Person Identification System," IEEE, 2025.
[2] H. Y. Kim, "Implementation of YOLO-based Missing Person Search AI Application System," ResearchGate, 2023.
[3] H. S. Hemadharshini, "AI Based – Assisted Search for Missing Person," ResearchGate, 2023.
[4] K. Basha Sk, "Facial Recognition and Neural Networks for Missing Child Identification – A Smart Approach," 2025.
[5] D. Sattibabu, "Tracing Missing Person Through Facial Recognition Using Deep Learning," Springer Nature, 2024.
[6] P. Raj, "Facial Recognition System for Identification of Missing Person," EasyChair, 2024.
[7] M. F. Sani, "Improving Real-Time Attendance System Based on YOLO and FaceNet Using Super Resolution Models," IEEE, 2025.
[8] A. F. Ayon, "Toward Digitalization: A Secure Approach to Find a Missing Person Using Facial Recognition Technology," arXiv, 2024.
[9] J. Zhang, "Accuracy and Robustness Evaluation of Deep Learning Algorithms in Facial Recognition Systems," Elsevier, 2025.
[10] R. Alekya, "Missing Child Identification System Using Deep Learning and Multiclass SVM," MLSoft, 2023.
[11] G. P. Rajan, "Person Re-Identification Using Deep Learning-Based YOLO Network with Partially Occluded Faces," Springer Nature, 2025.