POLITECNICO DI TORINO
Repository ISTITUZIONALE

Multimodal Fusion of Face and Gait for Person Identification in Automotive Applications

Original
Multimodal Fusion of Face and Gait for Person Identification in Automotive Applications / Boscolo, F., Lamberti, F., Montuschi, P., Testa, M.. - In: IEEE INTERNET OF THINGS JOURNAL. - ISSN 2327-4662. - ELETTRONICO. - 13:2(2026), pp. 2438-2450. [10.1109/JIOT.2025.3631488]

Availability:
This version is available at: 11583/3004733 since: 2025-11-12T14:31:14Z

Publisher:
IEEE

Published
DOI:10.1109/JIOT.2025.3631488

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

19 November 2026

Multimodal Fusion of Face and Gait for Person Identification in Automotive Applications

Federico Boscolo, *Graduate Student Member, IEEE*, Fabrizio Lamberti, *Senior Member, IEEE*, Paolo Montuschi, *Fellow, IEEE*, and Mario Testa

***Abstract*—Smart and secure access to vehicles is a crucial aspect of the evolving automotive industry. This article focuses on the development of an end-to-end multimodal biometric recognition framework that identifies people walking toward a vehicle from an RGB video feed. The framework is based on a deep learning pipeline for person detection and tracking, face and gait feature extraction, and fusion of the two modalities at the score and feature level. Traditional face recognition (FR) systems can suffer from variations in lighting and occlusions. In order to deal with these issues, the proposed framework integrates face and gait features with the aim to enhance accuracy. The pipeline is modular, enabling seamless integration of new models for each step of person identification without the need for additional training. Baseline face and gait recognition (GR) models, as well as score- and feature-level fusion (FLF) techniques, are evaluated on subsets of the CASIA-A and CASIA-B datasets. Experimental results show that weighted mean score-level fusion (SLF) significantly improves both Rank-1 accuracy and verification accuracy (TAR@FAR = 10^{-5}) over unimodal baselines. Overall, the reported work provides insights into current limitations and suggests directions for future research about secure identity verification in vehicles.**

***Index Terms*—Face recognition (FR), feature fusion, gait recognition (GR), intelligent vehicles, multimodal biometrics.**

I. Introduction

IN THE era of cutting-edge technologies, the automotive industry is increasingly embracing intelligent, connected, and autonomous vehicles. This shift not only redefines the driving experience but also provides different, more convenient access methods over traditional key fobs, such as digital keys. Nevertheless, scenarios involving exceptional or emergency access, where physical or digital keys may be unavailable or not viable, as well as new services to be possibly activated at a distance before the driver's entry, necessitate innovative solutions.

The evolution of vehicle access methods passed from physical ignition keys to remote keyless entry (RKE) systems, which, despite advancements like rolling keys, remain vulnerable to cryptographic and replay attacks [1]. Advances continued with the integration of smartphone apps, Bluetooth, and near field communication (NFC) technologies, as seen in solutions from major automobile manufacturers [2]. In parallel, future 6G-ready vehicular networks are expected to rely on machine learning-driven perception and authentication services to meet stringent latency and security requirements [3]. Machine learning also addresses the challenges of vehicular networks such as heterogeneous connectivity and real-time constraints by enhancing perception and communication capabilities through data-driven models [4]. Recently, biometric identification techniques like fingerprints and face recognition (FR) have been implemented in some vehicles to provide additional access methods. However, fingerprints require contact with a sensor, while FR, besides suffering from variations in lighting and occlusions, is vulnerable to spoofing attacks in real-world scenarios, especially in the absence of additional depth information [5]. Alternative contactless modalities, like wireless sensing-based driver authentication [6] and behavioral biometrics via CAN-bus command patterns [7], have also been explored. The integration of biometric data from multiple sources, e.g., face and gait, represents a promising path to enhance driver identification at a distance and improve robustness of recognition systems, overcoming the limitations of existing FR approaches, as demonstrated in [8] and [9]. Similar multimodal frameworks, such as *FIMBISAE* by Ahmed et al. [10], applied in the medical field, demonstrate how fusing heterogeneous traits (e.g., fingerprints and ECG) can strengthen system security. Gait recognition (GR), for instance, allows for the analysis of walking patterns of distant subjects and is less susceptible to lighting variations and facial occlusions; it usually does not outperform FR, but it can provide more robustness to FR systems and address their shortcomings. In this article, we present a modular, multimodal framework that integrates face and gait biometrics to improve recognition performance in automotive access scenarios. Our framework is based on a plug-and-play pipeline that combines pretrained deep learning models within distinct modules for person detection, tracking, segmentation, FR, GR, and fusion at both score and feature levels. We evaluate our proposed framework using CASIA-A [11] and CASIA-B datasets [12]. Unlike previous approaches that combine face and gait modalities, such as [8], which rely on single-frame FR, our approach aggregates face feature vectors across multiple frames, enhancing robustness and yielding substantially higher recognition performance. Notably, prior literature lacks comprehensive evaluations comparing multiple fusion methods within a unified pipeline, particularly for

Received 20 June 2025; revised 5 September 2025; accepted 1 November 2025. Date of publication 11 November 2025; date of current version 8 January 2026. This publication is part of the project PNRR-NGEU which has received funding from the MUR - DM 117/2023. *(Corresponding author: Fabrizio Lamberti.)*

Federico Boscolo, Fabrizio Lamberti, and Paolo Montuschi are with the Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy (e-mail: federico.boscolo@polito.it; fabrizio.lamberti@polito.it; paolo.montuschi@polito.it).

Mario Testa was with the Department of Control and Computer Engineering, Politecnico di Torino, 10129 Turin, Italy. He is now with the ADS (AI, Data and Space), LINKS Foundation, 10138 Torino, Italy (e-mail: mariotesta1999@gmail.com).

Digital Object Identifier 10.1109/JIOT.2025.3631488

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

automotive-oriented recognition and verification. Our findings provide a comprehensive landscape of fusion techniques for face and gait on CASIA-A and CASIA-B, underlining the potential of fusion techniques to improve identity verification systems at a distance in future automotive applications.

Our contributions are summarized as follows.

1) The development of a modular multimodal biometric pipeline that leverages pretrained state-of-the-art models for vehicle owner recognition; this pipeline substantially improves Rank-1 accuracy upon single-modality approaches, demonstrating its effectiveness with a 99.40% Rank-1 accuracy on CASIA-B using the best fusion configuration, compared to 96.96% for the best unimodal face baseline and 90.51% for the best unimodal gait baseline.
2) The systematic evaluation of diverse fusion strategies, including algebraic score-level methods (i.e., weighted mean, sum, and multiplication) and more complex feature-level fusion (FLF) techniques such as direct concatenation, polynomial fusion, hierarchical fusion, and Transformer-based neural fusion. Our experiments demonstrate that weighted mean score-level fusion (SLF) achieved a Rank-1 accuracy of 99.40% and a true acceptance rate (TAR) at false acceptance rate (FAR) = 10^{-5} of 98.06% on CASIA-B, significantly outperforming unimodal baselines.
3) The introduction of preprocessing techniques such as dynamic weighted mean fusion based on subject–camera distance and input video upscaling to increase recognition robustness.

The remainder of this article is organized as follows. Section II reviews related work in FR, GR, and their fusion. Section III introduces our proposed multimodal framework, including the pipeline architecture, its modules, and considered fusion methods. Section IV presents the experimental setup, baseline model performance, and comprehensive fusion results on CASIA-A and CASIA-B datasets. Finally, Section V concludes this article, discussing limitations and outlining future research directions.

II. Related Work

In this section, we review key contributions related to FR, GR, and their fusion.

A. Face Recognition

FR is a widely studied and deployed task in computer vision, particularly in identity verification and authentication applications. It is commonly used in access control systems, where it enables secure entry based on facial features, and in surveillance, where it assists in monitoring and identifying individuals in real time. FR tasks can be split into two main categories: face verification and face identification. Face verification involves comparing two face images to determine whether they depict the same individual, typically needed for confirming identities. Face identification, in turn, involves comparing a face against multiple stored identities to find a match and is commonly used in surveillance and law enforcement.

Furthermore, FR can operate under two different protocols: the closed-set protocol, in which all testing identities are present in the training set, thereby collapsing the recognition task to a classification problem, and the open-set protocol, which involves different identities in the training and testing sets, resembling real-world scenarios. Most FR approaches today focus on the open-set protocol.

Convolutional neural networks (CNNs) became the most dominant approach in FR when DeepFace [13] demonstrated near-human performance on the labeled faces in the wild (LFW) [14] dataset in 2014. DeepFace used a deep CNN trained with a softmax-based classification loss, mapping face images into a discriminative feature space where intraclass distances were minimized and interclass distances were maximized. Following DeepFace, two primary paradigms emerged: softmax-based methods and contrastive learning methods. The former category refined DeepFace's approach by incorporating variations of softmax loss, treating FR as a multiclass classification problem, and proving effective in closed-set scenarios. Methods belonging to the latter category, such as triplet loss introduced by FaceNet in 2015 [15], directly optimized the embedding space for face verification and open-set recognition by grouping positive pairs and distancing negative pairs; despite requiring complex strategies for mining negative samples, this approach significantly improved FR in real-world applications.

To enhance the performance of softmax loss in open-set scenarios, techniques like SphereFace [16], CosFace [17], and ArcFace [18] were developed. SphereFace incorporated a margin penalty into the softmax loss to optimize angular distributions of features, though it faced stability issues. CosFace and ArcFace further improved accuracy by introducing additive cosine and angular margins, respectively; in particular, ArcFace demonstrated superior margin control and robustness, becoming the preferred loss function for recent FR architectures due to its high accuracy in challenging conditions such as lighting and pose variations.

One of the most widely used datasets for FR is the previously mentioned LFW dataset, which contains over 13 000 images of faces collected from the web under unconstrained settings, making it a benchmark for face verification in real-world conditions. Popular large-scale datasets include MS-Celeb-1M [19], featuring over 10 million images of 100 000 celebrities, and VGGFace2 [20], comprising 3.3 million images of over 9000 subjects, which includes a diverse range of poses, ages, and ethnicities. Several datasets have been devised to mitigate the effects of face occlusions such as masks. One example is the masked FR dataset (MFDD) by Wang et al. [21], which provides over half a million real and synthetically masked face images.

B. Gait Recognition

GR is another popular computer vision task, which offers a method to identify individuals based on their walking patterns. Unlike FR, which can be impacted by changes in expression or lighting, especially at a distance, gait is more robust to variations in appearance, such as clothing or accessories, and

can work despite face occlusions. This robustness makes it an attractive method for security systems, surveillance, and, in general, identification based on the overall appearance of a person.

Like FR, GR can be categorized into closed- and open-set protocols. In the closed-set protocol, all identities in the test set are present in the training data, whereas the open-set protocol handles unseen identities during testing. Most contemporary works focus on open-set scenarios due to their relevance in real-world applications such as video surveillance.

GR methods are primarily based on model-free or model-based approaches. Model-free techniques, such as gait energy image (GEI) [22], focus on silhouette-based representations, making them easier to implement but sensitive to variations in clothing and carrying conditions. Model-based methods, in contrast, construct explicit models of the human body and estimate joint movements, offering more robustness though, often, at the cost of higher computational complexity.

With the advent of deep learning, model-free approaches based on CNNs and recurrent neural networks (RNNs) have become predominant in GR due to their ability to capture both spatial and temporal information from gait sequences. CNN-based methods such as GaitSet [23] and GaitPart [24] leverage the hierarchical feature extraction capability of their architecture, representing gait as a set of spatial features aggregated across time. In particular, the two mentioned models use a sequence of binary silhouettes to extract spatial–temporal information, with GaitPart splitting gait sequences into parts and focusing on localized body parts.

Several datasets have emerged to support research in GR, with CASIA-A [11] and CASIA-B [12] being among the most widely used. CASIA-A, one of the earliest datasets for GR, contains sequences from 20 subjects walking in three directions (parallel, 45° and 90°) relative to the camera. CASIA-B is a larger dataset comprising 124 subjects, captured from 11 different viewpoints, and includes variations in clothing and carried objects to simulate real-world conditions. Other notable datasets are OU-ISIR [25], which offers extensive cross-view data, and TUM-GAID [26], focused on indoor and outdoor gait sequences under realistic conditions.

C. Fusion of Face and GR

Early works on fusion of face and gait, such as that by Kale et al. [27], explored both hierarchical and SLF strategies to combine the two biometric traits. Their hierarchical approach involved cascading multiple classifiers in increasing order of accuracy. Specifically, a gait classifier was applied first to narrow down the candidate pool, reducing the number of subjects passed to the face classifier, which operated on a smaller and more refined set. This progressive filtering minimized errors by decreasing the number of probes at the final stage. This approach was compared to SLF techniques like sum, product, and min rules, showing improved recognition accuracy on the NIST gait database.

Later, Zhou and Bhanu [28] focused on feature fusion to enhance recognition at a distance, working with side-view data. Their method involved linear normalization of features and dimensionality reduction, with subsequent concatenation of face and gait features. They experimented on a proprietary dataset of 46 subjects, which was not publicly released.

Geng et al. [29] introduced an adaptive fusion method designed to account for variations in view angle and subject-to-camera distance. By adjusting the fusion weights dynamically based on silhouette size and angle on five different views, they achieved improved performance over single modalities, especially with challenging view angles. Their approach, which was tested on a subset of the CASIA-A dataset, incorporated adaptive score fusion and demonstrated that optimized weighting schemes could lead to more robust identification.

More recently, research has expanded to cover the fusion of deep learning-based recognition approaches. For instance, Fu et al. [8] focused on face and gait fusion at the feature level, using pretrained face and gait models (ResNet50 with ArcFace loss and GaitSet) on a subset of the CASIA-B dataset. Their method combined min–max normalization with feature concatenation, achieving moderate gains in cross-view scenarios.

Prakash et al. [9], in turn, introduced an adaptive fusion approach using convolutional long short-term memory (LSTM) models and keyless attention mechanisms, adjusting fusion weights based on view angle. They experimented on CASIA-A, achieving improvements in accuracy over the simpler method in [29].

III. Methodology

To integrate FR and GR effectively, we designed a multimodal framework that leverages state-of-the-art models for each stage of the recognition process, capable of applying a wide range of fusion techniques. Sections III-A–III-C describe the framework's overall architecture, detailing the modular pipeline and the fusion strategies employed.

A. Pipeline for Joint Face and GR

This section presents a multimodal biometric recognition pipeline based on the combination of face and gait modalities, designed for the integration of models with a plug-and-play architecture. As discussed in Sections I and II, the primary motivation for this fusion is to improve recognition performance, especially in challenging scenarios with lack of face clarity and variations in subject appearance. Gait, being less affected by external factors, complements FR by mitigating its limitations. Our approach aims to improve both accuracy and robustness in out-of-vehicle identification tasks in the wild, by exploring fusion methods at both score and feature levels.

To achieve our goal, the devised pipeline, illustrated in Fig. 1, combines state-of-the-art models for each step of person identification, from the detection of subjects in RGB videos to the final recognition scores. The overall workflow is detailed in Algorithm 1, which orchestrates the entire process from gallery creation to final score generation. A crucial subroutine within this pipeline is the feature extraction stage, separately detailed in Algorithm 2.

An input video I is fed to the pipeline, whose first step is person detection and tracking, to extract the bounding boxes of each subject in the video. The resulting video sequences

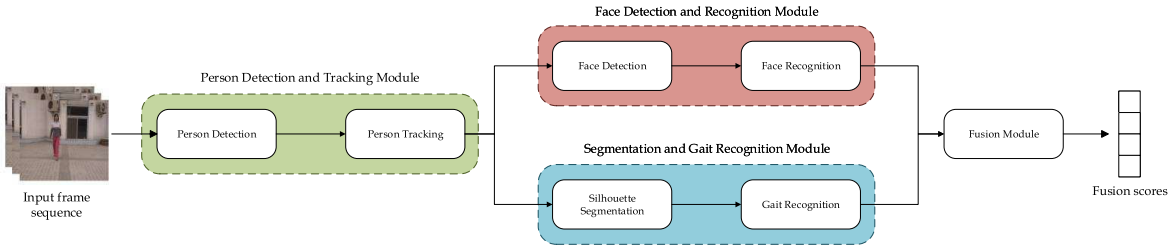


Fig. 1. Overview of the proposed pipeline. Given an input video sequence I, the system detects and tracks individuals, extracting cropped video sequences for each of them. Face and gait features are then computed separately and compared against their respective galleries using a similarity metric m. The resulting face and gait scores, along with their feature vectors, are passed to the fusion module, which performs SLF and FLF to generate the final prediction scores. The specific models and fusion methods used depend on the pipeline configuration.

Algorithm 1 Recognition and Fusion Pipeline

Input: Probe video set P_{set}, Gallery video set G_{set}, Metric m
Output: Set of fused score pairs for the probe set $\mathcal{S}_{probe}$

// -- 1. Gallery Creation --
$G_f \leftarrow \emptyset$; $G_g \leftarrow \emptyset$;
foreach *video* $I_g \in G_{set}$ **do**
  // Call Alg1 to get gallery feat.
  $(\mathcal{V}_{f_g}, \mathcal{V}_{g_g}) \leftarrow \text{FeatureExtraction}(I_g)$;
  Append features from $\mathcal{V}_{f_g}$ to G_f;
  Append features from $\mathcal{V}_{g_g}$ to G_g;
end
$G_{fused} \leftarrow \text{FuseGalleries}(G_f, G_g)$;
// -- 2. Probe Processing --
$\mathcal{S}_{probe} \leftarrow \emptyset$;
foreach *video* $I_p \in P_{set}$ **do**
  // Call Alg1 to get probe feat.
  $(\mathcal{V}_{f_p}, \mathcal{V}_{g_p}) \leftarrow \text{FeatureExtraction}(I_p)$;
  // Process feat. for each subject in the probe video
  for *each pair* (V_{f_k}, V_{g_k}) *from* $(\mathcal{V}_{f_p}, \mathcal{V}_{g_p})$ **do**
    // Score-Level Fusion
    $S_{f_k} \leftarrow \text{Compare}(V_{f_k}, G_f, m)$;
    $S_{g_k} \leftarrow \text{Compare}(V_{g_k}, G_g, m)$;
    $S_{SLF} \leftarrow \text{FuseScores}(S_{f_k}, S_{g_k})$;
    // Feature-Level Fusion
    $V_{fused_k} \leftarrow \text{FuseFeatures}(V_{f_k}, V_{g_k})$;
    $S_{FLF} \leftarrow \text{Compare}(V_{fused_k}, G_{fused}, m)$;
    Append (S_{SLF}, S_{FLF}) to $\mathcal{S}_{probe}$;
  end
end
return $\mathcal{S}_{probe}$;

Algorithm 2 Multimodal Feature Extraction

Input: Video sequence I
Output: Set of face features $\mathcal{V}_f$, Set of gait features $\mathcal{V}_g$

$\mathcal{V}_f \leftarrow \emptyset$; $\mathcal{V}_g \leftarrow \emptyset$;
// Get tracked subject sequences
$\hat{\mathcal{I}} \leftarrow \text{PersonDetectorTracker}(I)$;
foreach *sequence* $\hat{I}_k \in \hat{\mathcal{I}}$ **do**
  // -- Face Feature Extraction --
  $A_{f_k} \leftarrow \text{DetectAndAlignFaces}(\hat{I}_k)$;
  $F_{f_k} \leftarrow \text{ExtractFaceFeatures}(A_{f_k})$;
  // Aggregate face feat. for the sequence
  $V_{f_k} \leftarrow \text{Mean}(F_{f_k})$;
  // -- Gait Feature Extraction --
  $M_{g_k} \leftarrow \text{ExtractSilhouettes}(\hat{I}_k)$;
  $V_{g_k} \leftarrow \text{ExtractGaitFeatures}(M_{g_k})$;
  if *NormalizeEmbs* **then**
    $V_{f_k} \leftarrow \text{Normalize}(V_{f_k})$;
    $V_{g_k} \leftarrow \text{Normalize}(V_{g_k})$;
  end
  Append V_{f_k} to $\mathcal{V}_f$;
  Append V_{g_k} to $\mathcal{V}_g$;
end
return $(\mathcal{V}_f, \mathcal{V}_g)$;

$\hat{I} = \{\hat{I}_0, \ldots, \hat{I}_{n_s}\}$, cropped to each subject's bounding box for all n_s detected subjects, are processed in order by the rest of the pipeline. Each processed sequence $\hat{I}$ is passed to the face detection and recognition module as well as to the segmentation and GR module, which perform their respective tasks. The face feature vector V_f extracted from $\hat{I}$ is compared to the face gallery $G_f = \{G_{f_0}, \ldots, G_{f_n}\}$, comprised of the face feature vectors for all n subjects in the gallery, using the preferred similarity metric m. In parallel, the gait feature vector V_g is compared to the gait gallery $G_g = \{G_{g_0}, \ldots, G_{g_n}\}$ using the same metric. The resulting gait prediction scores $S_g = \{S_{g_0}, \ldots, S_{g_n}\}$, along with V_g and V_f and the face scores $S_f = \{S_{f_0}, \ldots, S_{f_n}\}$, are then fed into both SLF and FLF modules. These modules perform fusion at their respective levels and produce the final prediction scores for the particular subject.

B. Pipeline Modules

As anticipated, the pipeline consists of the following modules, which cover each step of recognition from RGB images as follows.

1) *Person Detection and Tracking*: Detect individuals approaching the vehicle and track their movement across frames.
2) *Face Detection and Recognition*: Locate the face of detected subjects in the frame and then extract face features and recognition scores.
3) *Segmentation and GR*: Extract subject silhouettes and perform GR, extracting gait features and recognition scores.
4) *Score/FLF*: Combine face and gait outputs either at the score level or feature level, determining the overall prediction and the final decision for each detected subject.

The first three modules are encapsulated within the *Feature-Extraction* routine detailed in Algorithm 2. The fusion module corresponds to the logic in Algorithm 1, which consumes the extracted features to produce recognition scores. The pipeline modules are described in detail in the remainder of this section, while the specific models employed for each module are discussed in Section IV-A.

1) Person Detection and Tracking Module: The person detection and tracking module handles the initial stage of the person identification process and is split into two steps as follows.

1) *Person Detection*: From an input video, the module identifies individuals and outputs the corresponding bounding boxes along with confidence scores.
2) *Multiperson Tracking*: Using the detected bounding boxes, the module assigns a unique identifier (ID) to each subject based on their appearance.

Each step is executed using a pretrained deep learning model. The framework automatically crops frames, creates sequences for each subject, and links them to the corresponding track IDs. For a given input video, the final output is a set of image sequences cropped to the bounding boxes of each tracked individual. Each sequence is then passed to the face and GR branches for further processing. This corresponds to the *PersonDetectorTracker* routine in Algorithm 2. Processing of the input sequence is conducted frame by frame, which allows the module to work on a live recording with minimal latency, depending on the complexity of the detection and tracking models employed.

Advantages: This module ensures that all subjects detected on camera are processed. Its real-time, frame-by-frame processing allows for minimal latency, key for live applications.

Limitations: Performance can be affected by factors like subject distance, extreme occlusions, or harsh lighting conditions, which may lead to missed detections or tracking errors.

2) Segmentation and GR Module: The segmentation and GR module is responsible for generating gait embeddings from a sequence of video frames. This process consists of two steps as follows.

1) *Segmentation*: Semantic segmentation is applied to the input frames to create binary masks (i.e., silhouettes).
2) *GR*: The extracted silhouettes are fed into a GR model, which produces temporally aggregated gait feature embeddings.

These two steps are represented by the *ExtractSilhouettes* and *ExtractGaitFeatures* routines in Algorithm 2. Both steps are performed using pretrained models; different model combinations were tested.

Advantages: GR is highly robust to variations in facial appearance (e.g., occlusions, expressions, and lighting) and can identify subjects at a distance. It is also less susceptible to spoofing attacks compared to FR.

Limitations: GR can be sensitive to changes in clothing, carrying conditions, and camera view angles, which might alter walking patterns. The accuracy may also decrease significantly with degraded silhouettes.

3) Face Detection and Recognition Module: The face detection and recognition module is responsible for generating face embeddings from each frame in a sequence. This process is divided into two steps as follows.

1) *Face Detection and Alignment*: The module identifies and locates human faces and normalizes them to a standard position using detected landmarks.
2) *FR*: The aligned faces are passed through a feature extractor, generating embeddings.

In Algorithm 2, these operations are represented by *DetectAndAlignFaces* and *ExtractFaceFeatures* routines, followed by a mean aggregation to produce a single feature vector per subject sequence. These steps are implemented using two pretrained models; different FR models were tested.

Advantages: FR is a highly mature and widely accepted biometric modality, offering high accuracy for identity verification under ideal conditions. It is intuitive and directly verifiable by users.

Limitations: Its performance significantly degrades with variations in lighting, pose, facial expressions, and partial occlusions (e.g., masks and hats). It is also known to be vulnerable to spoofing attacks that use photos or videos.

4) Fusion Module: The fusion module is designed to apply a variety of techniques for the fusion of face and gait modalities. In general, fusion of modalities can be performed at the decision level, score level, or feature level. In this work, we do not consider approaches at the decision level because they treat FR and GR as independent classifiers, preventing the system from leveraging their combined discriminative power at a more granular level. Therefore, the architecture of the fusion module currently supports score and FLF methods only.

In the case of SLF, the scores S_f of the comparison between the face embedding V_f with the face gallery G_f, along with the corresponding gait scores S_g obtained from comparing V_g with G_g, are combined in one of multiple ways (e.g., using sum, multiplication, and weighted mean of scores).

FLF, in turn, operates directly on the embeddings V_f and V_g to integrate their feature representations into a unified embedding space (e.g., by means of concatenation). To be able to compare the newly fused feature vector $V_{f\oplus g}$ with a gallery, the same operations must be performed on both face and gait galleries, resulting in a combined face and gait gallery $G_{f\oplus g}$ that can be compared with the feature vector to obtain the combined score S'. Algorithm 1 explicitly implements both fusion paths. It first generates separate galleries (G_f, G_g) and a fused gallery (G_{fused}). Then, for each probe subject, it

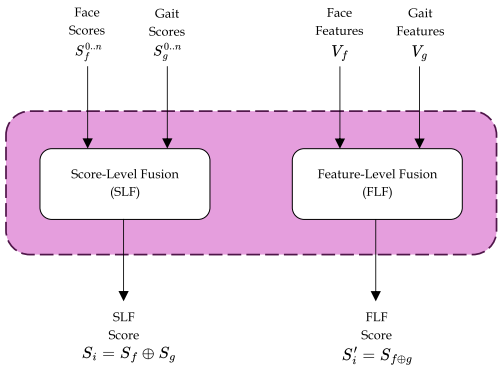


Fig. 2. Fusion module's different modes of operation. The final prediction scores are calculated using both SLF and FLF techniques. In the former case, prediction scores are combined after comparison with their respective galleries, while in the latter case, extracted features are combined and compared with a fused gallery in order to obtain the scores. The specific models and fusion methods used depend on the pipeline configuration.

computes scores via the SLF path (FuseScores) and the FLF path (FuseFeatures followed by Compare). Fig. 2 provides a diagram of the fusion module's operation in the two different fusion modes.

Advantages: Fusion significantly enhances overall recognition accuracy and robustness by combining complementary information from FR and GR, mitigating their limitations. It improves performance in challenging scenarios and increases resistance to spoofing attacks.

Limitations: The effectiveness of fusion depends on the chosen strategy and the quality of individual modalities. FLF can be computationally more complex and may require model training for optimal performance. The choice of the similarity metric for the fused features also impacts accuracy.

C. Fusion Methods

A variety of different methods can be used to perform fusion at the score or feature level. This section describes the main fusion techniques that were considered in this work.

1) Score-Level Fusion: Fusion at the score level is typically used for combining recognition scores in a fast yet effective manner. This approach aggregates the prediction scores of individual models before making the final recognition decision. Common SLF techniques include the following.

1) *Sum*: This method calculates the final prediction score as the sum of individual scores from each modality

$$S_{\text{sum}} = S_f + S_g. \tag{1}$$

2) *Multiplication*: By multiplying scores, this approach emphasizes cases where both modalities strongly indicate the same identity. However, it is more sensitive to low-confidence scores, as a single low score can lower the overall score significantly

$$S_{\text{mult}} = S_f \cdot S_g. \tag{2}$$

3) *Weighted Mean*: To account for the varying reliability of different modalities, the weighted mean fusion applies predefined weights to each score before averaging

$$S_{\text{mean}} = \alpha \cdot S_f + (1 - \alpha) \cdot S_g. \tag{3}$$

We employ all of these methods in our pipeline to determine which one is best suited for recognition, as well as to provide different fusion options in different scenarios.

2) Feature Concatenation: Fusion at the feature level involves the combination of raw or processed feature vectors from each modality before computing the prediction scores. Feature concatenation represents a simple approach to FLF.

First, the face feature vectors are averaged over all frames of the input sequence to obtain a mean face feature vector, which is combined with the gait feature vector. The two vectors are then normalized using either min–max normalization or L2 normalization. Next, the feature vectors are flattened and concatenated into a single composite feature vector. The same operations are applied to the gallery features to enable comparison.

3) Dynamic Weighted Mean: The dynamic weighted mean approach represents an alternative to the weighted mean SLF. Instead of combining scores using the same weight α for all frames, the weight can vary based on external factors (e.g., time and subject distance).

For instance, when the subject is further from the camera, gait is expected to be more reliable due to low face resolution. As the subject approaches, the face modality becomes more informative. We adopt this approach in our pipeline to dynamically emphasize the most reliable modality in each section of the input sequence.

4) Hierarchical Feature Fusion: This approach, first described in [27], sequentially filters samples by performing FR and GR in a sequential manner.

The process begins by applying the less accurate recognizer (gait), ranking samples by gait scores. A subset of top-ranked candidates is selected, and the corresponding face scores are computed only for this subset. Final predictions are made using face scores, reducing the computational load.

5) Automated Feature Fusion: Inspired by recent advances in Transformer-based architectures, we introduce a deep fusion method that leverages a Transformer encoder to model relationships between face and gait embeddings.

Face and gait embeddings are first normalized with L2 normalization. Given their dimensionalities (512 for face and 1184 for gait), we project both into a common 848-D space using linear layers. These projections serve as two input tokens to a Transformer encoder with four layers, eight attention heads, and 2048-D feedforward layers.

Through self-attention, the network learns intramodal and cross-modal features. We aggregate the output token embeddings by concatenation, for a fused embedding of 1696 dimensions.

We trained the fusion model using both contrastive and triplet losses, with triplet loss yielding better performance. Training lasted 50 epochs with early stopping on a validation set of ten subjects drawn from the training data.

D. Dataset Selection

To evaluate the performance of our pipeline and fusion techniques, we needed a dataset including both face and gait information in RGB video format, simulating realistic out-of-vehicle identification scenarios where subjects approach a

TABLE I
SUMMARY OF THE CONSIDERED STATE OF THE ART DATASETS. ONLY CASIA-A AND CASIA-B OFFER BOTH CLEAR, NONBLURRED FACE DATA AND FRONTAL VIEW ANGLES, MAKING THEM SUITABLE FOR TASKS INVOLVING JOINT FACE AND GR

Name	Type	Face data	Front views
CASIA-A [11]	RGB/Silh.	Yes	Yes
CASIA-B [12]	RGB/Skel./Silh.	Yes	Yes
NIST/USF [33]	RGB	No	No
TUM-GAID [26]	RGB	Yes	No
CCGR [31]	RGB/Skel./Silh.	No	No
CASIA-E [34]	RGB/Silh.	N/A	N/A
FVG [30]	RGB	No	Yes
SUSTech-1K [32]	RGB/Point-cloud	No	Yes

vehicle (i.e., the camera). For our task, a combination of clear, nonblurred faces and frontal or semi-frontal viewpoints was required.

Given these constraints, we surveyed several available gait datasets to determine their suitability for our task. A summary of this evaluation is presented in Table I.

Some of the surveyed datasets ultimately had to be discarded as they did not fully meet our requirements. TUM-GAID [26] only provides side-to-side views, lacking any frontal or semi-frontal angles for vehicle approaching scenarios. Datasets such as FVG [30], CCGR [31], and SUSTech-1K [32] contain blurred faces, which completely impair our FR Module. The NIST database related to the USF HumanID Gait Challenge [33] was excluded due to incompatible dataset characteristics, namely, limited frontal views and the lack of sufficiently clear face images. CASIA-E [34] appears to be a promising dataset for our purpose, but it has not been released for research use at the time of writing.

We ultimately selected CASIA-A [11] and CASIA-B [12], deeming them most suitable for our task.

CASIA-A is one of the earliest gait datasets and contains RGB video sequences of 20 subjects walking at different angles (parallel, 45°, and 90°) relative to the camera. It provides clear frontal and semi-frontal face views in a controlled setting. CASIA-B extends the previous dataset with 124 subjects captured from 11 different viewpoints and includes three walking conditions: normal (NM), with a bag (BG), and with a coat (CL). These conditions are valuable for evaluating robustness to variations in appearance. For experimental purposes, we only selected view angles at 0°, 18°, and 36°, which correspond to typical angles for approaching a vehicle.

It is worth remarking that these datasets have some limitations as well. The image resolution is relatively low compared to modern camera standards, and both were recorded in a single background environment, which may not fully represent the variability of real-world conditions. Ideally, a domain-specific dataset tailored for out-of-vehicle identification would address these gaps, but to our knowledge, such a dataset is not yet available.

IV. EXPERIMENTS

In this section, we present the experimental evaluation of our multimodal recognition pipeline, starting with the selection of baseline models for each module.

A. Model Selection

After defining the pipeline modules, we selected state-of-the-art models for each step. This section describes the chosen models and discusses alternative approaches for the key components of the pipeline: FR, GR, and segmentation.

1) Person Detection and Tracking: In the proposed pipeline, YOLOv8 [35] is employed as the baseline model for person detection due to its combination of speed and accuracy. YOLOv8 leverages an advanced backbone and head architecture to deliver state-of-the-art performance across detection tasks. For person tracking, we utilize the ByteTrack tracking algorithm [36]. ByteTrack combines high-confidence detections with low-confidence ones, using a robust association strategy to enhance tracking continuity. The combination of ByteTrack with YOLOv8 ensures efficient detection and tracking, allowing for reliable cropping of the input RGB sequences to the bounding box of the subject for each frame.

2) Segmentation and GR: Three alternatives are considered for segmentation or silhouette extraction. The HumanSegV2 model from Baidu's PaddlePaddle framework [37] was selected as the first alternative, due to its reliability and integration with the OpenGait framework. Mediapipe Selfie Segmenter [38] demonstrated effectiveness in extracting human silhouettes across diverse contexts. Ultralytics YOLOv8's segmentation was also considered. GR is implemented using models available in the OpenGait framework [39], including GaitSet, GaitPart, and GaitBase. These models were trained on the CASIA-B and OU-MVLP datasets, and we employed them with their pretrained weights.

3) Face Detection and Recognition: YOLOv5Face [40] is used as the baseline model for the face detection task. Trained on the WiderFace dataset, YOLOv5Face demonstrates robust performance in real-world scenarios.

For recognition, three alternative models were considered. GhostFaceNets [41] was selected for its balance of accuracy and efficiency. ArcFace [18] was chosen for its handling of pose and expression variation. Facebook DeepFace [13] was also considered for its scalability and robustness.

4) Image Upscaling: The quality of input images could significantly impact the performance of biometric recognition systems. Low-resolution frames, as in the case of the CASIA gait datasets, can degrade feature extraction and potentially reduce the accuracy of FR and GR models. To address this issue, we incorporated a state-of-the-art image upscaler, real-world enhanced super-resolution generative adversarial network (Real-ESRGAN) [42], to increase image resolution before processing. Real-ESRGAN is a deep learning-based super-resolution model designed to handle real-world image degradations, such as noise, blur, and compression artifacts.

For this work, we employed the Real-ESRGAN model using pretrained weights and applied $2\times$ image upscaling to subject frames, bringing their resolution from 320×240 pixels to 640×480, before face and gait feature extraction. While this preprocessing step does not improve FR performance significantly due to the low resolution of faces in the input images, it ensures a higher clarity of silhouettes before GR, improving the quality of segmentation and subsequent gait feature extraction.

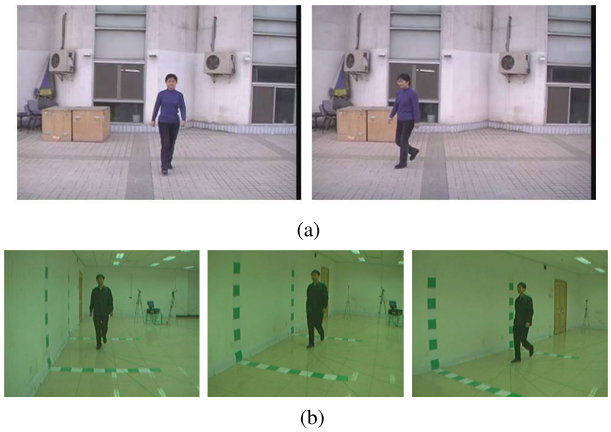


Fig. 3. Samples from the CASIA datasets used in this work. (a) Samples from CASIA-A (angles: 45° and 90°). (b) Samples from CASIA-B (angles: 0°, 18°, and 36°).

B. Experimental Setup

Experiments were conducted on CASIA-A and CASIA-B using only RGB video data. Fig. 3 depicts sample frames of sequences from both datasets.

The first 74 subjects of CASIA-B were used for training. For testing, the first NM sequence (NM01) per angle was used as gallery; all others (NM02-06, CL01-02, and BG01-02) were used as probes. For CASIA-A, 0° view was used as gallery; other 0° and two 45° views were used as probes. This results in 150 gallery samples and 1350 probes for CASIA-B, and 20 gallery samples and 60 probes for CASIA-A.

Both Rank-1 accuracy and TAR@FAR = 10^{-5} were evaluated. Two metrics were used to compute the similarity between the gallery and probes. The first is cosine similarity, defined for two feature vectors $\mathbf{A}$ and $\mathbf{B}$ as

$$\text{cosine similarity} := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}.$$

The second one is Euclidean distance, with the formula for two feature vectors $\mathbf{A} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{B} = (b_1, b_2, \ldots, b_n)$ given by

$$d(\mathbf{A}, \mathbf{B}) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}.$$

These metrics are commonplace in both FR and GR domains.

C. Baseline Model Performance

The baseline performance of FR and GR models was evaluated independently to establish reference points for subsequent fusion experiments. Tables II and III report Rank-1 accuracy and TAR@FAR = 10^{-5}, respectively.

GhostFaceNetsV1 was selected as the face model for subsequent experiments due to its superior performance. All gait models were retained for fusion evaluations, given their competitive baseline accuracy.

TABLE II

BASELINE RANK-1 ACCURACY RESULTS FOR THE CONSIDERED FACE AND GR MODELS ON THE CASIA-B TEST SET. "COS" REFERS TO THE COSINE SIMILARITY METRIC, WHEREAS "EUC" REFERS TO THE EUCLIDEAN DISTANCE METRIC

Model	Rank-1 Acc. (%)	
	cos	euc
GhostFaceNetsV1	**96.96**	**96.66**
ArcFace	88.22	85.70
DeepFace	87.77	86.14
GaitBase	89.25	90.51
GaitSet	88.00	88.07
GaitPart	89.25	90.00

TABLE III

BASELINE VERIFICATION RESULTS FOR THE BEST PERFORMING FACE MODEL (GHOSTFACENETSV1) AND THE CONSIDERED GAIT MODELS, USING BOTH COSINE SIMILARITY AND EUCLIDEAN DISTANCE

Model	TAR@FAR=10^{-5}	
	cos	euc
GhostFaceNetsV1	**70.05**	**51.49**
GaitBase	46.72	48.11
GaitSet	45.22	49.62
GaitPart	13.63	33.22

TABLE IV

SLF RESULTS FOR THE SELECTED COMBINATIONS OF MODELS

Model combination	Rank-1 Acc. (%)		TAR@FAR=10^{-5}	
	cos	euc	cos	euc
GaitBase+GhostFaceNetsV1 (Sum)	99.33	97.11	95.37	67.20
GaitSet+GhostFaceNetsV1 (Sum)	98.96	98.81	88.32	92.95
GaitPart+GhostFaceNetsV1 (Sum)	98.88	98.81	93.93	92.95
GaitBase+GhostFaceNetsV1 (Mult)	98.22	98.96	92.38	90.64
GaitSet+GhostFaceNetsV1 (Mult)	98.81	99.03	88.15	91.17
GaitPart+GhostFaceNetsV1 (Mult)	98.74	**99.33**	93.92	92.69
GaitBase+GhostFaceNetsV1 (Mean)	**99.40**	98.88	**98.06**	90.41
GaitSet+GhostFaceNetsV1 (Mean)	98.81	98.88	88.32	91.66
GaitPart+GhostFaceNetsV1 (Mean)	**99.40**	**99.33**	91.35	**91.94**

D. Fusion Results

To assess the effectiveness of multimodal integration, we evaluated different fusion strategies for joint face and GR and examined their impact on Rank-1 and verification accuracy. Sections IV-D1–IV-D3 present results for SLF, feature concatenation, and more advanced fusion techniques.

1) Score-Level Fusion: In SLF, the individual similarity scores produced by the face and gait modules are combined to generate a unified decision metric. As described in Section III-B.4, three different SLF methods are employed: sum, multiplication, and weighted mean. A linear search was conducted on parameter α for the weighted mean. The best model combinations for all fusion methods are reported in Table IV.

2) Feature Concatenation: Feature concatenation combines the embeddings extracted from the face and gait modules. Table V details the performance using the best model combinations.

3) Advanced Fusion Techniques: After initial experimentation, the best model configuration (GaitBase + GhostFaceNetsV1) and weighted mean fusion were defined as the

TABLE V

FEATURE CONCATENATION FUSION RESULTS FOR THE SELECTED COMBINATIONS OF MODELS. IN PARENTHESES, THE TYPE OF NORMALIZATION USED FOR THE FEATURES

Model combination	Rank-1 Acc. (%)		TAR@FAR=10^{-5}	
	cos	euc	cos	euc
GaitBase+GhostFaceNetsV1 (L2)	**98.96**	**98.96**	**94.61**	**94.61**
GaitSet+GhostFaceNetsV1 (L2)	98.88	98.88	94.01	94.01
GaitPart+GhostFaceNetsV1 (L2)	**98.96**	**98.96**	88.24	88.24
GaitBase+GhostFaceNetsV1 (Min-Max)	94.22	94.59	62.81	46.98
GaitSet+GhostFaceNetsV1 (Min-Max)	89.70	89.55	46.65	45.32
GaitPart+GhostFaceNetsV1 (Min-Max)	97.25	97.18	87.58	88.49

TABLE VI

COMPARISON OF ADVANCED FUSION APPROACHES WITH DIFFERENT PREPROCESSING STRATEGIES

Technique	Rank-1 Acc. (%)		TAR@FAR=10^{-5}	
	cos	euc	cos	euc
Benchmark (Mean)	99.40	98.88	98.06	90.41
Polynomial	99.18	99.18	97.00	97.00
Hierarchical	98.14	92.45	96.47	96.15
Transformer-based	98.37	98.59	85.76	85.42
Benchmark (DYN)	99.40	99.40	92.70	92.70
Polynomial (DYN)	99.48	99.48	96.80	96.80
Hierarchical (DYN)	98.00	98.59	96.47	93.57
Transformer-based (DYN)	98.44	98.52	76.52	92.93
Benchmark (UPS)	99.40	99.63	95.24	92.86
Polynomial (UPS)	99.40	99.40	97.09	97.09
Hierarchical (UPS)	99.40	99.11	**98.28**	72.27
Transformer-based (UPS)	98.44	98.52	76.37	92.93
Benchmark (DYN+UPS)	99.40	**99.70**	90.38	95.60
Polynomial (DYN+UPS)	**99.63**	99.63	97.24	**97.24**
Hierarchical (DYN+UPS)	99.03	99.40	88.93	95.30
Transformer-based (DYN+UPS)	98.44	98.52	76.60	92.93

benchmark. Table VI presents the results for polynomial, hierarchical, and Transformer-based fusion, with or without preprocessing techniques ("UPS" refers to upscaling and "DYN" refers to dynamic weighted mean).

E. Cross-View Performance on CASIA-A

To further verify the robustness of our pipeline, we conducted additional experiments on CASIA-A, which presents a more challenging cross-view setting compared to CASIA-B. Unlike CASIA-B, where probes could be directly compared with the corresponding gallery of the same angle, in CASIA-A, there are only two sequences per angle with the subject facing the camera: 0° and 45°. This setup introduces significant appearance variations, resembling a general scenario where a subject may approach the vehicle at a substantially different angle than the recorded gallery.

We replicated the evaluations performed on CASIA-B, first by comparing the baseline face and gait approaches and then applying SLF and feature concatenation techniques. The results of these experiments are reported in Table VII.

Table VIII provides a comparison of a subset of the advanced fusion techniques on the CASIA-A dataset.

The advanced fusion approaches evaluated on CASIA-A achieved consistent Rank-1 accuracies of 96.67%, matching the feature concatenation benchmark, except for the hierarchical approach. This inconsistency is likely due to a poor generalization of the top results selection threshold for the gait classifier. The Rank-1 accuracy of 96.67% corresponds to only 2 incorrect predictions out of a total of 60 probes, demonstrating high performance. Consequently, we did not deem it necessary to further evaluate the dynamic weighted mean and upscaling preprocessing techniques, previously effective on CASIA-B, due to the limited improvement potential. The small size of the CASIA-A dataset, however, remains a significant bottleneck, preventing a fully comprehensive assessment of the performance of the tested fusion approaches.

TABLE VII

COMPARISON OF BASELINE, SLF, AND FEATURE CONCATENATION RESULTS ON THE CASIA-A DATASET

Model combination	Rank-1 Acc. (%)		TAR@FAR=10^{-5}	
	cos	euc	cos	euc
GaitBase (baseline)	58.33	60.00	57.14	55.55
GhostFaceNetsV1 (baseline)	90.00	85.00	**85.19**	60.78
GaitBase+GhostFaceNetsV1 (Sum)	95.00	85.00	77.19	64.70
GaitBase+GhostFaceNetsV1 (Mult)	95.00	90.00	70.18	62.96
GaitBase+GhostFaceNetsV1 (Mean)	95.00	90.00	70.18	62.96
GaitBase+GhostFaceNetsV1 (Concat_L2)	**96.67**	**96.67**	72.41	**72.41**
GaitBase+GhostFaceNetsV1 (Concat_MinMax)	86.67	88.33	38.46	58.49

TABLE VIII

COMPARISON OF ADVANCED FUSION APPROACHES WITH THE FACE BASELINE AND THE BEST-PERFORMING APPROACH, REFERRED TO AS "BENCHMARK," ON CASIA-A

Technique	Rank-1 Acc. (%)		TAR@FAR=10^{-5}	
	cos	euc	cos	euc
Face baseline (GhostFaceNetsV1)	90.00	85.00	85.19	60.78
Benchmark (Concat_L2)	**96.67**	**96.67**	72.41	**72.41**
Polynomial	**96.67**	**96.67**	65.52	65.52
Hierarchical	86.67	81.67	**88.88**	60.78
Transformer-based	**96.67**	**96.67**	56.89	55.17

F. Discussion

The experiments discussed in Sections IV-D and IV-E evaluated various combinations of FR and GR models as well as fusion techniques to determine the most effective pipeline for out-of-vehicle person identification under the considered conditions. The pipeline was tested on a subset of the CASIA-B gait dataset and then validated on the CASIA-A dataset to evaluate cross-view generalization.

The baseline results served to demonstrate the performance of individual FR and GR models. GhostFaceNetsV1 achieved the highest accuracy among the FR models, with a Rank-1 accuracy of 96.96% and a TAR@FAR = 10^{-5} of 70.05%. For GR, GaitBase outperformed the other models, with a Rank-1 accuracy of 90.51% and a TAR@FAR = 10^{-5} of 48.11%. Rank-1 accuracy was already high for these approaches, but the verification score could be improved.

For what it concerns SLF, the weighted mean proved to be the most robust method, consistently achieving high performance on both CASIA-A and CASIA-B. The benchmark combination of GhostFaceNetsV1 and GaitBase (with PP-HumanSeg) using weighted mean fusion reached a Rank-1 accuracy of 99.40% and TAR@FAR = 10^{-5} of 98.06% on CASIA-B, improving the former score and greatly improving the latter, establishing it as the most effective fusion method for our pipeline.

The simpler FLF method, i.e., feature concatenation, provided improvements to the baseline comparable to SLF on CASIA-B but did not quite reach the weighted mean method. On CASIA-A, it proved to be more effective than the weighted mean method in terms of Rank-1 accuracy.

In addition, several advanced fusion methods were evaluated, including polynomial SLF, hierarchical fusion, and a deep Transformer-based FLF method, as well as preprocessing techniques such as input image upscaling and dynamic weighted mean of features. The dynamic weighted mean approach, which assigns weights to face frames based on their distance from the camera, provided modest improvements to accuracy and verification scores, particularly when combined with polynomial fusion. The upscaling of input frames boosted Rank-1 accuracy but did not consistently enhance verification scores. However, the combination of dynamic weighted mean and image upscaling yielded slight improvements across both metrics.

The overall best configuration identified in this study was the combination of GhostFaceNetsV1 for FR, GaitBase for GR, and SLF using the weighted mean approach. This setup performed better than baseline models as well as other fusion approaches on CASIA-B, with a Rank-1 accuracy of 99.40% and a TAR@FAR = 10^{-5} of 98.06%. The weighted mean approach generalized well to CASIA-A, achieving high performance, being outperformed only by feature concatenation with min–max normalization on CASIA-A.

Regarding the advanced fusion techniques, their variable performance deserves a closer analysis. The Transformer-based model's suboptimal verification accuracy can be attributed to its significant data requirements; the relatively small scale of the CASIA datasets is insufficient for the Transformer to learn a truly generalizable fused embedding space, likely leading to overfitting on the training subjects. The performance of the hierarchical approach can be explained by the limited accuracy of the first gait-based filtering stage. An incorrect filtering decision at this step cannot be corrected by the face classification step, causing a propagation of error that particularly impacts challenging cross-view scenarios. These factors highlight that, for the specific conditions and datasets tested, simpler and more robust fusion rules offered a better tradeoff between complexity and generalization.

G. Comparison With State of the Art

In this section, we present an analysis of our proposed method's performance against established state-of-the-art approaches on CASIA-B and CASIA-A. To ensure a fair and direct comparison, we adopted the same experimental protocols employed by the authors of the original works.

For the CASIA-B dataset, we benchmarked our model against the results presented by Fu et al. [8]. We replicated their exact experimental setup, which involves using walking sequences with view angles ranging from 0° to 90°. For the gallery set, we included the first four "Normal" (NM) sequences (NM#01–04) for each subject; the remaining NM sequences (NM#05 and 06), along with all "Bag" (BG) and "Coat" (CL) sequences, were used as the probe set.

TABLE IX
COMPARISON WITH STATE OF THE ART ON CASIA-B

Method	NM#05-06	BG#01-02	CL#01-02	Average
Face baseline (from [8])	9.90	9.04	7.60	8.85
Gait baseline (from [8])	95.57	88.87	74.77	86.40
Fusion (Fu et al. [8])	95.87	90.23	76.70	87.60
Face (Ours, GhostFaceNetsV1)	98.17	97.33	95.33	96.94
Gait (Ours, GaitBase)	97.83	92.17	66.17	85.39
Fusion (Ours, Weighted Mean)	**99.67**	**98.83**	**97.16**	**98.55**

TABLE X
COMPARISON WITH STATE OF THE ART ON CASIA-A

Method	Rank-1 Accuracy (%)
Face baseline(from [9])	80.00
Gait baseline (from [9])	70.00
Fusion (Prakash et al. [9])	90.00
Fusion (Geng et al. [29])	86.67
Face (Ours, GhostFaceNetsV1)	78.33
Gait (Ours, GaitBase)	99.16
Fusion (Ours)	**100.00**

TABLE XI
END-TO-END COMPUTATIONAL PERFORMANCE OF THE PIPELINE

Metric	Value
Latency (ms/frame)	5.84
GFLOPs/frame	3.76
Peak VRAM (MB)	2819
Average Power (W)	160

As illustrated in Table IX, the results clearly demonstrate the superiority of our approach. Our unimodal face module, leveraging GhostFaceNetsV1, achieves an average Rank-1 accuracy of 96.94% compared to 8.85% reported in [8]. Similarly, our gait module based on GaitBase reaches 85.39% on average. Our fusion strategy is able to outperform the best method described in [8].

For the CASIA-A dataset, we adhered to the evaluation protocol by Prakash et al. [9], also previously adopted in [29]. This protocol involves, for each of the three angles (0°, 45°, and 90°), selecting two sequences per subject for the gallery and using the remaining two sequences as the probe.

The results are presented in Table X. The unimodal face and gait approaches already provide competitive results, while our multimodal fusion approach is able to achieve a perfect Rank-1 accuracy of 100%, effectively solving the recognition task under this specific protocol and setting a new state-of-the-art benchmark for this dataset.

H. Computational Efficiency Analysis

The computational efficiency of the proposed pipeline is a critical aspect for its deployment in real-world applications (which is already planned), especially on resource-constrained embedded systems. This section presents an analysis of the pipeline's latency, computational cost, and memory consumption to evaluate its overall computational requirements.

The analysis was conducted on a high-performance desktop workstation equipped with an NVIDIA RTX 3090 GPU, an

TABLE XII
MODULE-BY-MODULE COMPUTATIONAL BREAKDOWN (LATENCY, FLOPS, AND VRAM)

Module	Latency (ms)	FLOPs (G)	Peak VRAM (MB)
Person Detection & Tracking	120	132.0	694.0
Face Detection & Recognition	472	201.7	588.0
Segmentation & Gait Recognition	41	80.1	1536.0
Fusion	9	N/A	1.0
Total	**642**	**413.8**	**2819.0**

Intel Core i7-12700K CPU, and 32 GB of RAM. While this setup does not represent a low-power embedded platform, it serves to establish a benchmark for the pipeline's raw performance and identify computational bottlenecks. The pipeline was implemented using PyTorch and evaluated in a Python 3.9 environment.

The following performance metrics were measured as follows.

1) *Inference Time (Latency)*: Reported in milliseconds (ms), representing the time taken for an inference pass through a pipeline component or the entire system.
2) *Computational Cost (FLOPs)*: Expressed in giga floating-point operations (GFLOPs), serving as a hardware-agnostic measure of the computational workload.
3) *GPU Memory Consumption (VRAM)*: Measured in megabytes, indicating the peak video RAM used.
4) *Power Consumption*: Power consumption in watts was monitored using `nvidia-smi` to provide an indication of the load on a desktop GPU.

Table XI presents the end-to-end computational performance of the full multimodal pipeline. These metrics represent the average requirements to process a single frame.

The pipeline achieves an average latency of 5.84 ms/frame, corresponding to a throughput of approximately 171 frames/s on the test hardware. The total peak memory footprint is 2819 MB, with an average power draw of 160 W. While these results demonstrate feasibility on high-end hardware, a more detailed breakdown is required to identify optimization targets for deployment on embedded systems.

To identify specific performance bottlenecks, a module-by-module breakdown of latency, FLOPs, and VRAM consumption was conducted. Table XII summarizes these results, highlighting the computational burden of each component for an entire probe sequence.

From the results, it is evident that the pipeline's performance is not uniformly distributed across its components. The face detection and recognition module is the primary bottleneck in terms of processing time, accounting for 472 ms, or approximately 73% of the total 642-ms latency. This module also represents the largest computational workload, requiring 201.7 GFLOPs.

Conversely, the segmentation and GR module is the most demanding in terms of memory. It requires a peak of 1536 MB of VRAM, which is more than double the consumption of the tracking module and nearly three times that of the face module. The person detection and tracking module presents a moderate load across all metrics, while the final fusion step is computationally negligible.

Furthermore, the impact of including the Real-ESRGAN upscaling module was evaluated. The additional overhead in terms of latency and memory was found to be negligible in the context of the entire pipeline's workload, and for this reason, its specific metrics are not detailed in the performance tables. These findings suggest that future optimization efforts for embedded deployment should prioritize reducing the latency of the FR pathway and minimizing the memory footprint of the segmentation and GR models.

I. Limitations and Future Work

While the proposed pipeline achieves high accuracy on standard benchmarks, it faces two main limitations: dataset generalization and computational cost. The CASIA-A and CASIA-B datasets do not fully reflect the complexity of real-world vehicle access conditions, which often include challenging scenarios such as heavy occlusion, extreme lighting variations, and diverse subject–camera interactions. To our knowledge, no publicly available dataset adequately simulates these specific conditions.

Furthermore, as established in the computational analysis, the current implementation has room for optimization. The FR module represents a computational bottleneck, while the segmentation and GR module has the largest memory footprint. This presents a practical limitation for deployment on the low-power or real-time embedded systems typical in intelligent vehicles.

Future work will focus on addressing these limitations. We plan to create a new, realistic dataset that reflects real-world vehicle access conditions, featuring high intersubject and intrasubject variability, diverse lighting, and occlusions. We are also exploring the creation of a synthetic dataset for access scenarios using state-of-the-art generative models (e.g., Stable Diffusion [43], Wan 2.1 [44], and ControlNet [45]).

Concurrently, a critical direction is the optimization of the pipeline for embedded hardware, such as the NVIDIA Jetson Orin family. This will involve leveraging frameworks like PyTorch and the TensorRT framework to address the identified computational and memory bottlenecks, moving the system from a high-performance benchmark to a feasible real-world application.

V. CONCLUSION

This article introduced a modular, multimodal person recognition pipeline for out-of-vehicle scenarios. Our primary

contribution is a system that effectively combines face and gait modalities, demonstrating substantial improvements over unimodal approaches in both recognition (Rank-1 accuracy) and verification (TAR@FAR), an aspect often overlooked in the literature.

Through extensive experimentation, we identified the combination of GhostFaceNetsV1 and GaitBase, fused via an SLF weighted mean, as the optimal configuration. This pipeline achieved state-of-the-art performance on CASIA-B with 99.40% Rank-1 accuracy and 98.06% TAR@FAR = 10^{-5}. It also showed strong generalization on CASIA-A, where our fusion approach achieved a perfect 100% Rank-1 accuracy under the tested protocol.

In conclusion, this work lays the foundation for out-of-vehicle multimodal person recognition systems and demonstrates the power of combining state-of-the-art recognition modes with innovative fusion techniques. By addressing the identified limitations, future research could extend the applicability and reliability of this approach in real-world scenarios, paving the way for reliable biometric vehicle access systems.

ACKNOWLEDGMENT

The authors would like to thank Pandeli Borodani, Franck Guillemard, and Thibaud Miquel for their valuable support and contributions to this research. They also thank Giuseppe Scarso and Alessandro Cirillo for the implementation and evaluation of various models and techniques relevant to this study.

REFERENCES

[1] F. D. Garcia, D. Oswald, T. Kasper, and P. Pavlidès, "Lock it and still lose it—On the (in)security of automotive remote keyless entry systems," in *Proc. 25th USENIX Conf. Secur. Symp.*, Mar. 2016, pp. 929–944.

[2] C. Busold et al., "Smart keys for cyber-cars: Secure smartphone-based NFC-enabled car immobilizer," in *Proc. 3rd ACM Conf. Data Appl. Secur. Privacy*, Feb. 2013, pp. 233–242.

[3] F. Tang, Y. Kawamoto, N. Kato, and J. Liu, "Future intelligent and secure vehicular network toward 6G: Machine-learning approaches," *Proc. IEEE*, vol. 108, no. 2, pp. 292–307, Feb. 2020.

[4] F. Tang, B. Mao, N. Kato, and G. Gui, "Comprehensive survey on machine learning in vehicular network: Technology, applications and challenges," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 2027–2057, 3rd Quart., 2021.

[5] M. Wang and W. Deng, "Deep face recognition: A survey," *Neurocomputing*, vol. 429, pp. 215–244, Mar. 2021.

[6] S. D. Regani, Q. Xu, B. Wang, M. Wu, and K. J. R. Liu, "Driver authentication for smart car using wireless sensing," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 2235–2246, Mar. 2020.

[7] Y. Xun, J. Liu, N. Kato, Y. Fang, and Y. Zhang, "Automobile driver fingerprinting: A new machine learning based authentication scheme," *IEEE Trans. Ind. Informat.*, vol. 16, no. 2, pp. 1417–1426, Feb. 2020.

[8] H. Fu, W. Kang, Y. Zhang, and M. S. Shakeel, "Fusion of gait and face for human identification at the feature level," in *Proc. 16th Chin. Conf. Biometric Recognit.*, Feb. 2022, pp. 475–483.

[9] A. Prakash, S. Thejaswin, A. Nambiar, and A. Bernardino, "Multimodal adaptive fusion of face and gait features using keyless attention based deep neural networks for human identification," 2023, *arXiv:2303.13814*.

[10] T. Ahmed, S. Samima, M. Zuhair, H. Ghayvat, M. A. Khan, and N. Kumar, "FIMBISAE: A multimodal biometric secured data access framework for Internet of Medical Things ecosystem," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 6259–6270, Apr. 2023.

[11] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1505–1518, Dec. 2003.

[12] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 4, Aug. 2006, pp. 441–444.

[13] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[14] G. B. Huang, V. Jain, and E. Learned-Miller, "Unsupervised joint alignment of complex images," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[15] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," 2015, *arXiv:1503.03832*.

[16] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6738–6746.

[17] H. Wang et al., "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[18] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[19] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 87–102.

[20] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[21] Z. Wang, B. Huang, G. Wang, P. Yi, and K. Jiang, "Masked face recognition dataset and application," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 5, no. 2, pp. 298–304, Apr. 2023.

[22] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[23] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," 2018, *arXiv:1811.06186*.

[24] C. Fan et al., "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14213–14221.

[25] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1511–1521, Oct. 2012.

[26] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, "The TUM gait from audio, image and depth (GAID) database: Multimodal recognition of subjects and traits," *J. Vis. Commun. Image Represent.*, vol. 25, no. 1, pp. 195–206, Jan. 2014.

[27] A. Kale, A. K. Roychowdhury, and R. Chellappa, "Fusion of gait and face for human identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 5, May 2004, p. 901.

[28] X. Zhou and B. Bhanu, "Feature fusion of face and gait for human recognition at a distance in video," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 529–532.

[29] X. Geng, L. Wang, M. Li, Q. Wu, and K. Smith-Miles, "Adaptive fusion of gait and face for human identification in video," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2008, pp. 1–6.

[30] Z. Zhang et al., "Gait recognition via disentangled representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4705–4714.

[31] S. Zou, C. Fan, J. Xiong, C. Shen, S. Yu, and J. Tang, "Cross-covariate gait recognition: A benchmark," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 7855–7863.

[32] C. Shen, F. Chao, W. Wu, R. Wang, G. Q. Huang, and S. Yu, "LiDARGait: Benchmarking 3D gait recognition with point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1054–1063.

[33] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanID gait challenge problem: Data sets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.

[34] C. Song, Y. Huang, W. Wang, and L. Wang, "CASIA-E: A large comprehensive dataset for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 2801–2815, Mar. 2023.

[35] G. Jocher, J. Qiu, and A. Chaurasia. (2023). *Ultralytics YOLO*. [Online]. Available: https://github.com/ultralytics/ultralytics

[36] Y. Zhang et al., "ByteTrack: Multi-object tracking by associating every detection box," 2021, *arXiv:2110.06864*.

[37] L. Chu et al., "PP-HumanSeg: Connectivity-aware portrait segmentation with a large-scale teleconferencing video dataset," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 202–209.
[38] C. Lugaresi, "Mediapipe: A framework for perceiving and processing reality," in *Proc. 3rd Workshop Comput. Vis. (AR/VR) IEEE Comput. Vis. Pattern Recognition*, Mar. 2019.
[39] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, and S. Yu, "OpenGait: Revisiting gait recognition toward better practicality," 2022, *arXiv:2211.06597*.
[40] D. Qi, W. Tan, Q. Yao, and J. Liu, "YOLO5Face: Why reinventing a face detector," 2021, *arXiv:2105.12931*.
[41] M. Alansari, O. A. Hay, S. Javed, A. Shoufan, Y. Zweiri, and N. Werghi, "GhostFaceNets: Lightweight face recognition model from cheap operations," *IEEE Access*, vol. 11, pp. 35429–35446, 2023.
[42] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1905–1914.
[43] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021, *arXiv:2112.10752*.
[44] T. Wan et al., "Wan: Open and advanced large-scale video generative models," 2025, *arXiv:2503.20314*.
[45] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023, *arXiv:2302.05543*.



Federico Boscolo (Graduate Student Member, IEEE) received the M.Sc. degree in computer engineering from the Politecnico di Torino, Turin, Italy, in 2023, with a specialization in AI and data analytics. He is currently pursuing the Ph.D. degree with the GRaphics and INtelligent Systems (GRAINS) Research Group, Politecnico di Torino.

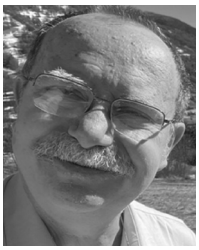
His research interests include machine learning algorithms for biometrics and person recognition at a distance. His work focuses on face and gait-based recognition systems, their multimodal fusion, and integration into automotive applications.



Fabrizio Lamberti (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in computer engineering from the Politecnico di Torino, Turin, Italy, in 2000 and 2005, respectively.

He is a Full Professor with the Department of Control and Computer Engineering, Politecnico di Torino. His research interests are in the areas of computer graphics, computer vision, human–machine interaction, and intelligent systems.

Dr. Lamberti is serving as an Associate Editor for IEEE TRANSACTIONS ON CONSUMER ELECTRONICS, IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, *IEEE Consumer Electronics Magazine*, and the *International Journal of Human-Computer Studies*. He has been appointed an Editor-in-Chief for IEEE *(Consumer Electronics Magazine)* from 2026 to 2027.



Paolo Montuschi (Fellow, IEEE) received the Laurea degree in electronic engineering and the Ph.D. degree in computer engineering from the Politecnico di Torino, Turin, Italy, in 1984 and 1989, respectively.

He is a Full Professor with the Department of Control and Computer Engineering, Politecnico di Torino, where he is also the Vice Rector for IT Digital Transition. He is a Former Member of the Board of Governors of the university. His research interests include computer arithmetic, computer architectures, and intelligent systems.

Dr. Montuschi is serving as the Editor-in-Chief for IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTING.



Mario Testa received the M.Sc. degree in artificial intelligence and data analytics from the Politecnico di Torino, Turin, Italy, in 2024.

He is currently an AI Engineer with the ADS (AI, Data, and Space) Department, LINKS Foundation, Torino, working on EU Horizon-funded projects. His research focuses on NLP, recommendation systems, RAG, and deep learning for image classification, segmentation, and multimodal data integration.

Open Access funding provided by 'Politecnico di Torino' within the CRUI CARE Agreement