Pattern Recognition 172 (2026) 112627

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Multi-year long-term person re-identification using gait and HAR features

David Freire-Obregón *, Oliverio J. Santana, Javier Lorenzo-Navarro, Daniel Hernández-Sosa, Modesto Castrillón-Santana

Universidad de Las Palmas de Gran Canaria, Institute of Intelligent Systems and Numeric Applications in Engineering, Las Palmas de Gran Canaria, 35007, Spain

ARTICLE INFO

Keywords:
Person re-identification
Biometrics
Gait
Human action recognition

ABSTRACT

We propose a two-stream person re-identification (Re-ID) framework that integrates gait and human action recognition (HAR) through cross-attention fusion. The model processes gait sequences via a BiLSTM-based encoder to capture temporal motion dynamics. At the same time, HAR embeddings are extracted using pre-trained video backbones and distilled into compact behavioral features. These two modalities are fused using a cross-attention mechanism, enriching gait-based identity representations with context-aware activity cues. We evaluate our method on a newly curated long-term spatio-temporal dataset of ultra-distance runners captured in natural outdoor settings across multiple locations spanning three years (2020 to 2023). Experimental results demonstrate that integrating HAR significantly enhances gait-based Re-ID performance. Compared to gait-only models, our approach yields a 12 % improvement in mean Average Precision (mAP) in cross-year scenarios and up to an 11.6 % gain in same-year evaluations. The HAR-enhanced models also exhibit faster convergence and higher Rank-1 accuracy, establishing the effectiveness of multi-modal motion-based representations for long-term, real-world person Re-ID.

1. Introduction

Humans can recognize familiar individuals across diverse contexts and times. Biometrics automates this process using physical, chemical, or behavioral traits [1]. Typically, biometric verification systems assume a known gallery of identities created during registration. In contrast, when identities are not pre-registered, the task shifts to linking observations of the same individual across time and space, regardless of their true name. This problem, known as person re-identification (Re-ID), involves retrieving an individual across different cameras or time spans [2]. In this work, we focus on supervised Re-ID, where labeled identities guide the learning of embeddings. Most research emphasizes short-term image-based Re-ID, where appearance remains stable. Real deployments, however, require long-term Re-ID, where clothing and visual cues change over time. Video provides richer temporal information, with gait serving as a stable biometric foundation. Nevertheless, benchmarks for long-term, video-based Re-ID remain scarce. DeepChange [3] is among the few, though its imbalance in identity frequency poses challenges, and its use in video-based scenarios is limited. Long-term Re-ID in crowded, dynamic environments is particularly difficult: individuals may look alike, appear briefly, or undergo occlusions, motion blur, and domain shifts. Appearance-based cues often fail under such conditions. Motion signals such as gait, together with human action recognition (HAR), offer complementary stability and behavioral context. Despite this, they are rarely combined in trainable, end-to-end systems.

Our main contributions are as follows. First, we propose a two-stream architecture that combines gait and HAR features for person Re-ID. A gait dynamics is learned in a structured way by a BiLSTM branch, and a light-weight and well-cited baseline in video recognition is obtained by the HAR branch, producing compact clip-level embeddings by global-temporal average pooling, in order to keep the auxiliary branch light-weight. Then, we incorporate a cross-attention scheme enabling interactive mixing of the two feature streams, allowing for a flexible combination of activity and motion cues. We utilized a triplet loss function to uncover discriminative and generalizable representations. Second, we introduce a new dataset collected in unconstrained environments, where ultra-distance race participants were recorded at two locations in 2020 and at two additional locations in 2023. This design facilitates evaluation under both short-term conditions (same-day/within-year) and long-term conditions (across multiple years), thereby capturing realistic variations in appearance, viewpoint, and environment.

Experiments show that fusing gait and HAR outperforms single-stream and visual-only baselines, yielding more robust embeddings. Results highlight the potential of motion-based multimodal representations for Re-ID in unconstrained conditions.

* Corresponding author.
E-mail address: david.freire@ulpgc.es (D. Freire-Obregón).

https://doi.org/10.1016/j.patcog.2025.112627
Received 2 June 2025; Received in revised form 12 September 2025; Accepted 18 October 2025
Available online 23 October 2025

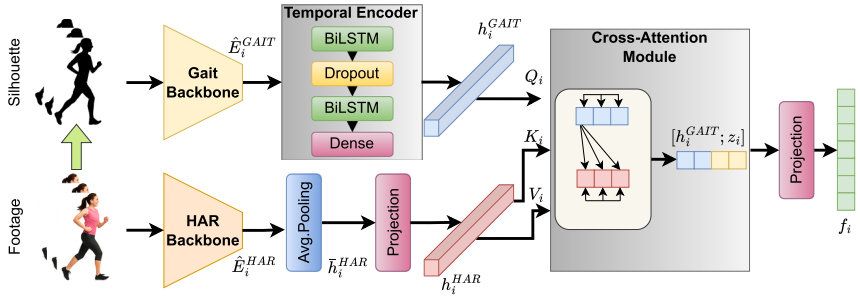


Fig. 1. Pipeline of the proposed two-stream architecture. Gait and HAR features are extracted independently, aligned in dimension, and fused via a cross-attention mechanism where gait attends to HAR. The fused embedding is used for identity representation. This design prioritizes gait as the primary cue while enriching it with complementary HAR context.

2. Related work

Person Re-ID has traditionally focused on visual appearance cues such as color, clothing, and facial features. However, in real-world long-term scenarios involving large-scale outdoor events or cross-year comparisons, appearance cues often fall short due to lighting variations, occlusions, and clothing changes [4]. To address these challenges, researchers have explored motion-based modalities such as gait as complementary or alternative identity signals [5]. Our work builds on this line of research by jointly modeling gait and HAR in a unified architecture for spatio-temporal Re-ID [6].

Re-ID in time and space. Spatio-temporal Re-ID aims to recognize individuals across significant temporal and spatial gaps, where appearance-based models are particularly vulnerable to domain shift and context changes [7]. Existing methods often address these issues by incorporating temporal modeling [8] or leveraging soft biometrics such as gait [9]. In this regard, gait is especially promising in long-term Re-ID due to its inherent stability and robustness against changes in clothing and viewpoint. However, many current models still struggle when faced with unconstrained real-world variability, such as that found in long-term datasets.

Appearance-based gait recognition. Gait recognition methods can be broadly categorized into skeleton-based and appearance-based approaches [9]. Our work considers 2D silhouettes used to extract discriminative motion patterns. More recent deep learning-based models, such as GaitSet [10], use set-based learning to handle unaligned input frames. GaitPart [11] introduces part-based modeling to capture local motion features better. These approaches have demonstrated strong performance under controlled conditions but often lack robustness in dynamic, real-world environments.

HAR and multi-modal cues. HAR provides a higher-level understanding of behavior that can complement gait in identity reasoning. Although HAR has been widely studied for activity classification tasks [12], its use in person Re-ID remains limited. Recently, pre-trained HAR models for person Re-ID in ultra-distance sports scenarios have been explored, emphasizing the impact of fatigue-induced movement changes on recognition performance [13]. However, their approach does not explicitly incorporate gait analysis. We aim to bridge this gap by jointly embedding HAR and gait cues, enabling richer, context-aware representations for spatio-temporal Re-ID.

Datasets. Several benchmark datasets support gait-based person Re-ID research. CASIA-B [14] remains one of the most widely used datasets, featuring gait sequences from 124 subjects under varying viewpoints and conditions (e.g., normal, bag-carrying, and coat-wearing). The OU-ISIR gait datasets, including the large-scale OU-MVLP [15], provide extensive samples across different age groups, clothing variations, and viewpoints. OU-MVLP, in particular, includes over 10,000 subjects captured from 14 view angles, making it one of the most extensive multi-view gait datasets available. GREW [16] is a recent large-scale dataset collected from real-world surveillance footage containing over 26,000 subjects in unconstrained environments. Other datasets, such as FVG [17], contribute to advancing cross-view and in-the-wild gait recognition by offering high-quality visual gait data under naturalistic conditions.

However, these datasets typically lack long-term variability, as they do not include recordings of the same individuals across multiple years. In contrast, one of our key contributions is a new dataset comprising real-world race footage recorded over three years (2020 and 2023) and from different distant locations within the same year. This enables the evaluation of gait-based Re-ID under realistic conditions of appearance change, long-term temporal gaps, and spatial variability, a scenario not addressed in existing benchmarks.

3. Methodology

This section describes the proposed two-stream architecture for spatio-temporal person Re-ID (see Fig. 1), which integrates gait and human action information through pre-trained backbones. We first formalize the problem, then detail the structure and roles of the gait and HAR backbones, the fusion strategy based on cross-attention, and finally, the training objective using triplet loss with semi-hard negative sampling.

3.1. Problem formulation

Let $\mathcal{D} = \{(v_i, y_i)\}_{i=1}^{N}$ be a dataset of N raw video recordings, where:

- v_i denotes the ith input video captured at a particular time and location,
- $y_i \in \{1, \ldots, C\}$ is the identity label associated with v_i.

Each video v_i is processed through two modality-specific pipelines to extract temporal embeddings for gait and HAR.

Gait Processing Pipeline. The gait modality aims to model the subtle motion patterns unique to each individual. To extract these cues, the raw video v_i is first processed into a sequence of binary silhouettes by combining person detection and pose estimation. Specifically, we use YOLOv8 [18] and Bot-SORT [19] for robust multi-object tracking, enabling consistent localization of the subject across frames. Silhouettes are later generated using SAMURAI [20]. The resulting silhouette sequence serves as input to a pre-trained gait encoder $\mathcal{B}_{\text{GAIT}}$, which has been trained on large-scale public datasets such as CASIA-B, OUMVLP, and GREW.

The encoder produces a feature tensor from the silhouette sequence:

$$E_i^{\text{GAIT}} = \mathcal{B}_{\text{GAIT}}(v_i^{\text{silhouettes}}) \in \mathbb{R}^{D_g \times P}$$

Where D_g is the feature dimension and P represents the number of horizontal partitions of the body used during Horizontal Pyramid Pooling (HPP). This output captures part-level information rather than temporal dynamics; each row corresponds to a distinct horizontal region of the body (e.g., upper torso, lower legs), not a time step.

To ensure stability, each part embedding vector is first standardized:

$$\tilde{E}_i[:,p] = \frac{E_i^{\text{GAIT}}[:,p] - \mu_{\text{train}}}{\sigma_{\text{train}}}$$

and then L2-normalized:

$$\hat{E}_i[:,p] = \frac{\tilde{E}_i[:,p]}{\|\tilde{E}_i[:,p]\|_2}$$

This results in a sequence of normalized part-based embeddings $\hat{E}_i^{\text{GAIT}} \in \mathbb{R}^{D_g \times P}$. Although they form a sequence-like structure, the order of parts corresponds to spatial locations rather than chronological time.

Part-wise Encoding. To capture the structured spatial information encoded in the body parts, the normalized part descriptors $\hat{E}_i^{\text{GAIT}}$ are interpreted as a sequence and passed through a BiLSTM. While this sequence does not represent time, the recurrent architecture allows for context modeling across adjacent body regions. This allows the model to capture structured co-movement patterns (e.g., torso-leg coordination) that are informative for identity. We use a bidirectional LSTM to aggregate such spatial dependencies. Let $\tilde{E}_i = (\hat{E}_i^{\text{GAIT}})^\top \in \mathbb{R}^{P \times D_g}$, where the part dimension P is treated as the temporal axis.

$$h_i^{\text{GAIT}} = \phi\big(\text{BiLSTM}_{256}\big(\text{Dropout}\big(\text{BiLSTM}_{128}(\tilde{E}_i)\big)\big)\big)$$

Here, ϕ denotes a dense layer with ReLU activation. This configuration models spatial body part embeddings as a pseudo-temporal sequence, allowing the BiLSTM to capture part-to-part relational dynamics relevant to identity.

HAR Processing Pipeline. The HAR stream is designed to capture high-level activity patterns from the subject's movement. However, raw videos may contain multiple actors, visual clutter, or background distractions To ensure that the extracted embeddings focus solely on the subject of interest, we apply a context-constrained preprocessing step.

Step 1: Video Preprocessing - Context Constraint. We use the silhouettes previously computed to crop a tight region around the individual in each frame. The rest of the frame is suppressed by superimposing the silhouette onto a static mode frame $\overline{f}$. For an individual i at time $t \in [0, T]$, let $Sil^{(i)}(t)$ be the silhouette, and $F^{(i)}(t)$ the raw frame. The preprocessed frame is defined as:

$$F'^{(i)}(t) = \text{Crop}(F^{(i)}(t), Sil^{(i)}(t)) + \left(1 - 1_{Sil^{(i)}(t)}\right) \cdot \overline{f}$$

This produces a context-constrained video $v_i^{\text{context}} = \{F'^{(i)}(t)\}_{t=0}^{T}$ focused solely on the individual.

Step 2: HAR Feature Extraction. The processed video v_i^{context} is then passed through a pre-trained action recognition backbone $\mathcal{B}_{\text{HAR}}$, which outputs a sequence of temporal embeddings:

$$E_i^{\text{HAR}} = \mathcal{B}_{\text{HAR}}(v_i^{\text{context}}) \in \mathbb{R}^{T' \times D_h}$$

where T' is the number of temporal segments or frames output by the HAR backbone and D_h is the dimensionality of each HAR embedding vector.

After extracting the sequence of embeddings from the HAR backbone, we apply normalization frame-wise before pooling. First, the embeddings are standardized using the mean and standard deviation computed across the training set, and then, each frame is L2-normalized:

$$\tilde{E}_i^{\text{HAR}} = \frac{E_i^{\text{HAR}} - \mu_{\text{train}}^{\text{HAR}}}{\sigma_{\text{train}}^{\text{HAR}}}, \quad \hat{E}_i^{\text{HAR}} = \frac{\tilde{E}_i^{\text{HAR}}}{\|\tilde{E}_i^{\text{HAR}}\|_2}$$

Although temporal modeling techniques, such as those used in gait recognition, aim to capture patterns, we observed that naive temporal processing alone did not yield significant improvements in Re-ID performance. Then, average pooling is applied over the temporal dimension to aggregate frame-level features into a sequence-level representation:

$$\bar{h}_i^{\text{HAR}} = \text{AvgPool}(\hat{E}_i^{\text{HAR}}) \in \mathbb{R}^{D_h}$$

Since the output dimensions of the gait and HAR streams may differ ($D_g \neq D_h$), we apply a projection layer to transform the HAR representation into the common dimension D:

$$h_i^{\text{HAR}} = \phi_{\text{proj}}(\bar{h}_i^{\text{HAR}}) \in \mathbb{R}^D$$

Both $h_i^{\text{GAIT}} \in \mathbb{R}^D$ and $h_i^{\text{HAR}} \in \mathbb{R}^D$ are projected into a shared embedding space of dimension D, where D denotes the common projection size (i.e., $D_g = D$), ensuring compatibility for subsequent fusion. The resulting pair of embeddings $(h_i^{\text{GAIT}}, h_i^{\text{HAR}})$ are aligned in dimension and subsequently used for cross-attention-based fusion and identity embedding computation.

Rationale for temporal aggregation. In the HAR stream, we use global temporal average pooling in order to obtain clip-level embeddings. This keeps the HAR stream lightweight, computationally robust, and easy to train in parallel with the combining module, aligning with typical practice in recent video recognition backbones (e.g., C2D [21], I3D [22], SlowFast [23], X3D [24]). Shallow recurrent layers and temporal 1D convolutions in our experiments did not show consistent improvement under long-term Re-ID, while average pooling kept accuracy with a modest cost. Even though more intricate aggregation methods are potentially available (e.g., self-attention, Transformer pooling), we intentionally select simplicity and robustness in order to uncover the isolated contribution of HAR cues towards gait-based recognition.

3.2. Cross-attention fusion

We apply a cross-attention mechanism to fuse both modalities, where the gait embedding attends to the HAR representation. Given the aligned embeddings $(h_i^{\text{GAIT}}, h_i^{\text{HAR}})$ for sample i, we define:

$$Q_i = h_i^{\text{GAIT}} \in \mathbb{R}^{1 \times D}$$

$$K_i = V_i = h_i^{\text{HAR}} \in \mathbb{R}^{1 \times D}$$

The attention weights and attended HAR embedding are computed as:

$$\alpha_i = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{D}}\right), \quad z_i = \alpha_i \cdot V_i$$

We then concatenate the original gait embedding with the attended HAR vector and apply a projection layer ϕ to obtain the final fused identity representation:

$$f_i = \phi_{\text{proj}}\big([h_i^{\text{GAIT}}; z_i]\big) \in \mathbb{R}^D$$

where $[\,;\,]$ denotes concatenation and ϕ_{proj} is a dense projection layer.

In this setup, gait is defined as the querying modality (serving as the Query in the attention mechanism) because the primary objective of the framework is identity recognition rather than activity categorization. Gait embeddings are explicitly trained to discriminate identities and thus form the base representation to be preserved. HAR features, in turn, provide the Keys and Values, offering complementary contextual cues that enrich the gait signal without shifting the focus toward action recognition. Through the attention mechanism, the model selectively integrates relevant HAR information into the gait embedding. The resulting attended representation is then combined with the original gait embedding, ensuring that gait remains the dominant identity signal. Finally, a projection layer fuses both sources into a unified identity representation in a common embedding space, where samples can be directly compared using L2 distance for Re-ID.

3.3. Triplet loss for metric learning

To teach the model how to tell different people apart, we use a *triplet loss* function. It compares three examples at a time:

- an anchor (a), a reference video of a person,
- a positive sample (p), another video of the same person,

- a negative sample (n), a video of a different person.

The goal is to make sure that the model places the anchor closer to the positive than to the negative in the embedding space, with some extra margin α for safety. The loss is defined as:

$$\mathcal{L}_{\text{triplet}} = \sum_{j=1}^{N} \max\left(\|f(a_j) - f(p_j)\|_2^2 - \|f(a_j) - f(n_j)\|_2^2 + \alpha,\ 0\right)$$

Here:

- $f(\cdot)$ is the embedding function (i.e., the model's output for each input),
- $\|\cdot\|_2$ is the L2 distance between embeddings,
- α is a fixed margin (e.g., 0.5),
- The $\max(\cdot,\ 0)$ makes sure we only penalize the model when the negative is too close to the anchor.

If the model already separates the anchor and negative correctly, we don't penalize it. If it doesn't, the loss becomes positive and the model learns from that mistake.

To construct effective triplets, we apply semi-hard negative mining, selecting negatives that satisfy:

$$\|f(a) - f(p)\|_2^2 < \|f(a) - f(n)\|_2^2 < \|f(a) - f(p)\|_2^2 + \alpha$$

We define a *modality-aware distance* used both in the loss function and during negative sample selection:

$$d_{\text{combined}}(x, y) = \lambda \cdot \|h_x^{\text{GAIT}} - h_y^{\text{GAIT}}\|_2 + (1 - \lambda) \cdot \|h_x^{\text{HAR}} - h_y^{\text{HAR}}\|_2$$

This combined distance selects negative samples across identities and checkpoints during mining, with $\lambda = 0.5$ balancing the contribution of gait and HAR modalities equally.

4. Dataset

Our study is based on a dataset collected at four different locations during the 2020 and 2023 editions of the Transgrancanaria ultra-distance running competition. In this event, athletes compete on a 128-kilometer course that typically takes between 12 and 30 h to complete. Each sample in the dataset consists of a short video clip, typically lasting no more than ten s, recorded at 25 frames per second. The dataset does not exhibit a symmetric structure in which each runner is recorded at every location. Instead, most runners appear in two locations, while a subset of fifteen runners appear in three within the training set and seventeen in three within the test set. As a result, location pairs differ in how much they share subjects. Some pairs involve identical sets of runners, while others involve partially overlapping or entirely distinct subsets. This necessitates a pair-level analysis when interpreting training and test splits, as seen in Table 1.

The training set of the dataset comprises a total of 423 labeled videos, each corresponding to a unique observation of a runner at a specific Recording Point (RP), identified by a location-year code (*RPloc_year*). These observations are distributed across four distinct points in different years: RP0_20, RP1_20, RP2_23, and RP3_23. Runners are most often recorded in exactly two of these locations, forming 234 unique unordered pairs, which serve as the fundamental unit of co-occurrence in the analysis. The pair RP0_20 ↔ RP1_20 emerges as the most frequent, appearing in 158 instances, indicating a strong connection between these two 2020 locations. Similarly, the 2023 pair RP2_23 ↔ RP3_23 appears 39 times, reflecting a notable but less dominant co-location pattern.

The training dataset includes 15 runners who appear in three distinct locations. Each of these runners contributes three unique location pairs, increasing the number of co-occurrence pairs without a proportional increase in video samples, as explained in Table 1. Their presence adds structural complexity to the network of relationships by connecting more location pairs per individual rather than inflating the dataset size.

Table 1
Number of runners and videos per location pair in the training and test sets. The training set contains 204 runners: 189 observed in two different locations and 15 observed in three, resulting in a total of 423 labeled videos. Because some runners contribute to multiple location pairs, the sum of runner and video counts across pairs exceeds the total number of unique training videos. The test set comprises 17 runners observed in three different locations, resulting in 51 labeled videos.

Set	Location Pair	# Runners	# Videos
Train	RP0_20 ↔ RP1_20	158	316
Train	RP2_23 ↔ RP3_23	39	78
Train	RP1_20 ↔ RP3_23	15	30
Train	RP0_20 ↔ RP3_23	11	22
Train	RP2_23 ↔ RP1_20	6	12
Train	RP0_20 ↔ RP2_23	5	10
Test	RP1_20 ↔ RP2_23	17	34
Test	RP1_20 ↔ RP3_23	17	34
Test	RP2_23 ↔ RP3_23	17	34

Fig. 2. These dataset samples illustrate short-term and long-term Re-ID scenarios. Each column corresponds to the same individual, while each row shows data captured at different locations and years. This figure has been anonymized for privacy purposes.

Moreover, the training dataset contains both same-year pairs, such as RP0_20 ↔ RP1_20 and RP2_23 ↔ RP3_23, and cross-year pairs, including examples like RP0_20 ↔ RP3_23 and RP1_20 ↔ RP2_23. These temporal pairings provide insight into longitudinal movement patterns, highlighting transitions and tracking continuity across years (see Fig. 2). This structure offers a rich foundation for spatio-temporal analysis and deeper modeling of athlete behavior across time and locations.

The test partition contains 51 videos from 17 runners, each recorded in three locations: RP1_20, RP2_23, and RP3_23. Evaluation is performed through pairwise comparisons between locations in both directions (e.g., RP1_20 to RP2_23 and RP2_23 to RP1_20), yielding six directional evaluation scenarios. For each direction, the 17 runners from the source location are matched against the 17 runners from the target location, producing 289 comparisons per direction. This results in a total of 1734 comparisons across the test set. The design ensures a comprehensive and balanced evaluation of generalization across spatial and temporal dimensions while maintaining a subject-disjoint protocol.

5. Gait & HAR backbones: From motion to identity

In this section, we provide a detailed description of the backbones used for gait and HAR, highlighting their architectural design and the type of information each captures. These modality-specific backbones form the foundation of our two-stream framework, with the gait backbone focusing on fine-grained motion dynamics and the HAR backbone capturing high-level activity patterns.

Gait Backbones. Silhouette-based gait recognition methods focus on extracting discriminative features from silhouette sequences to identify individuals based on their walking patterns. In our work, we employ several representative models-GaitBase, GLN Phase 1, GLN Phase 2, GaitGL, GaitPart, and GaitSet-which share the common objective of leveraging silhouette information yet differ significantly in their architectural designs. GaitSet treats gait sequences as unordered frame sets, using temporal pooling to aggregate frame-level features without modeling explicit spatial relationships. Building on this, GaitPart [11] incorporates part-based modeling using Focal Convolution (FConv) to extract local features across horizontal body regions, enhancing fine-grained spatial detail but potentially introducing sensitivity to alignment errors. GaitGL [25] refines this approach by combining global and local feature extraction branches with 3D convolutional layers to capture holistic and localized temporal-spatial dynamics. However, its added complexity does not always yield consistent improvements in real-world scenarios. GLN Phase 1 introduces a grouped latent representation strategy to disentangle feature learning, while GLN Phase 2 extends this with a refinement stage to progressively enhance feature granularity across network layers [26]. In contrast, GaitBase [27] adopts a deep residual network architecture to serve as a strong baseline backbone, demonstrating the benefits of increased network capacity for capturing robust gait patterns.

As previously mentioned in Section 3, the models have been trained and evaluated on three benchmark datasets-OU-MVLP, CASIA-B, and GREW-which vary significantly in terms of scale, environmental conditions, and overall complexity. OU-MVLP [15] is a large-scale constrained dataset collected in controlled indoor conditions with consistent camera viewpoints, offering extensive identity coverage but limited environmental variation. Although smaller in scale, CASIA-B [14] introduces covariates such as clothing and carrying conditions in a multi-view indoor setup, making it valuable for analyzing specific intra-subject variations. In contrast, GREW [16] is a real-world, in-the-wild dataset captured in unconstrained environments with diverse backgrounds, lighting conditions, and occlusions, reflecting practical deployment challenges and more related to the running scenario of our dataset. Due to these differences, not all models are trained across all datasets. Some architectures, particularly those relying on sensitive spatial modeling (e.g., GaitPart, GaitGL), may struggle to generalize on GREW without extensive reconfiguration. Furthermore, training high-capacity models such as GLN Phase 2 or GaitBase on GREW can be computationally demanding due to the dataset's scale and complexity. Consequently, model training choices are influenced by dataset characteristics and the architectural robustness and scalability of the gait backbones.

HAR Backbones. We employed a diverse set of backbone architectures with varying capacities to model spatial and temporal dynamics, including C2D, I3D, Slow8x8, Slow4x16, SlowFast8x8, SlowFast4x16, and X3D variants (L, M, S, XS). The C2D model [21] employs 2D convolutions on individual frames, treating the video as a sequence of static images. While efficient, it lacks explicit temporal modeling. I3D [22] overcomes this by inflating 2D filters to 3D and using two streams (RGB and optical flow) to capture appearance and motion jointly. SlowNet [28] processes fewer frames at high spatial resolution to model long-term patterns; we use Slow8x8 and Slow4x16 to vary temporal coverage.

SlowFast [23] adds a high-frame-rate path for fast motion, complementing the slow path's semantic focus. We evaluate SLF_8x8 and SLF_4x16. The X3D family [24] expands a 2D base model into spatiotemporal variants via progressive scaling. The four X3D versions differ in complexity and cost: XS applies multiple scaling stages, S reduces the frame rate, M increases spatial resolution, and L adds depth through deeper residual blocks. This structured design strikes a balance between efficiency and performance.

Several models (I3D, Slow, SlowFast) are enhanced with Non-local Networks (NLN) [29], which compute global pairwise dependencies across space and time. All backbones are pre-trained on Kinetics-400 [30], enabling robust and transferable HAR performance across varied conditions.

6. Experimental setup

Baselines. As baseline models for person Re-ID, we consider OSNet, DenseNet-121, MobileNet V2, and AlignedReID. All models are trained on the Market1501 dataset, a widely used benchmark in the Re-ID community. OSNet (Omni-Scale Network) [31] is designed explicitly for person Re-ID, featuring a multi-stream architecture that captures local and global features through dynamic, omni-scale feature aggregation while remaining lightweight and efficient. DenseNet-121 [32] is a popular backbone known for its densely connected layers, which encourage feature reuse and efficient gradient propagation, offering robust performance at a higher computational cost. MobileNet V2 [33] is a compact and efficient model optimized for mobile and real-time applications, utilizing depthwise separable convolutions to significantly reduce model size and inference time, with a trade-off in accuracy. AlignedReID [34] builds upon the ResNet backbone and incorporates local feature alignment through part-based matching, enhancing the model's ability to handle misalignment and pose variation-two common challenges in Re-ID scenarios. Together, these baselines offer a range of architectural styles, from lightweight mobile networks to specialized Re-ID solutions. These models are adapted for our task by leveraging tracklets extracted from the input videos. For each individual, we isolate sequences of bounding boxes (tracklets) corresponding to their appearances across the video timeline. To ensure robust and fair evaluation, we perform multiple comparisons between different runners, averaging the results over several runs to mitigate variance introduced by temporal sampling or environmental noise.

Implementation details. First, the gait pipeline is pre-trained; once fusion is enabled, we first train the fusion layers (lr = 6e-5) and then fine-tune the last gait layers with a reduced learning rate (lr = 1e-5). We trained the network with Adam. The training was conducted using a batch size of 32, and each iteration involved generating semi-hard triplets to enhance convergence and model robustness. The model was trained for 2000 iterations, with evaluations performed every 500 to monitor performance. We employed the Triplet Loss with a margin (α) 0.6 to encourage a clear separation between positive and negative pairs in the embedding space. Performance was assessed using standard retrieval metrics, including Cumulative Match Characteristic (CMC) curves and mean Average Precision (mAP), reported per location segment to capture spatial consistency across different checkpoints.

Experimental protocol. We performed a five times 10-fold split strategy on the training set. In each run, eight folds were used for training, one for validation, and one for internal testing. This approach thoroughly evaluated various hyperparameter configurations and fusion strategies while keeping the external test set completely untouched. The configuration that achieved the best performance was then selected to train the final model using the entire training set. The results reported in this work are based exclusively on evaluating this final model on the untouched test set, ensuring a fair and unbiased assessment of the system's generalization capabilities.

7. Long-Term Re-ID experiments

Long-term Re-ID is a key focus of this work, as it reflects real-world challenges where individuals must be re-identified after extended tem-

Table 2
Long-term Re-ID using only gait features. Cross-year evaluations between RP1_20 and RP2_23. Metrics reported as mAP / Rank-1 / Rank-5. Baseline performances are shown in gray, and the best-performing backbone is highlighted in blue.

Gait Backbone	RP1_20 → RP2_23			RP2_23 → RP1_20		
	mAP↑	R1↑	R5↑	mAP↑	R1↑	R5↑
MobileNet V2 [33]	11.6 %	0.0 %	11.8 %	20.9 %	5.9 %	29.4 %
DenseNet 121 [32]	16.9 %	0.0 %	29.4 %	24.5 %	5.9 %	35.3 %
OSNet [31]	31.9 %	17.7 %	52.9 %	28.8 %	11.8 %	47.1 %
AlignedReID [34]	28.3 %	11.8 %	47.1 %	30.1 %	17.6 %	41.2 %
GaitBase_C [27]	45.3 %	29.4 %	64.7 %	43.3 %	23.5 %	64.7 %
GaitBase_O [27]	28.4 %	11.8 %	52.9 %	34.9 %	23.5 %	47.1 %
GLNp1_C [26]	39.0 %	23.5 %	58.8 %	32.5 %	17.6 %	41.2 %
GLNp2_C [26]	33.5 %	17.6 %	52.9 %	39.1 %	29.4 %	47.1 %
GaitGL_C [25]	**49.4 %**	29.4 %	64.7 %	45.6 %	29.4 %	58.8 %
GaitGL_G [25]	46.2 %	23.5 %	76.5 %	**46.6 %**	29.4 %	64.7 %
GaitGL_O [25]	29.4 %	17.6 %	35.3 %	29.0 %	11.8 %	47.1 %
GaitPart_C [11]	40.2 %	23.5 %	47.1 %	34.9 %	17.6 %	64.7 %
GaitPart_G [11]	30.9 %	11.8 %	52.9 %	45.6 %	35.3 %	58.8 %
GaitPart_O [11]	30.2 %	11.8 %	47.1 %	28.1 %	11.8 %	52.9 %
GaitSet_C [10]	38.3 %	17.6 %	58.8 %	39.7 %	17.6 %	70.6 %
GaitSet_G [10]	30.8 %	11.8 %	52.9 %	35.3 %	17.6 %	47.1 %
GaitSet_O [10]	26.0 %	11.8 %	35.3 %	36.2 %	23.5 %	47.1 %

Table 3
Long-term Re-ID using only gait features. Cross-year evaluations between RP1_20 and RP3_23. Metrics reported as mAP / Rank-1 / Rank-5. Baseline performances are shown in gray, and the best-performing backbone is highlighted in blue.

Gait Backbone	RP1_20 → RP3_23			RP3_23 → RP1_20		
	mAP↑	R1↑	R5↑	mAP↑	R1↑	R5↑
MobileNet V2 [33]	14.6 %	0.0 %	29.4 %	14.7 %	0.0 %	26.5 %
DenseNet 121 [32]	19.2 %	0.0 %	35.3 %	23.6 %	11.8 %	23.5 %
OSNet [31]	19.9 %	0.0 %	29.4 %	17.8 %	0.0 %	29.4 %
AlignedReID [34]	27.5 %	17.6 %	29.4 %	24.1 %	11.8 %	35.3 %
GaitBase_C [27]	38.8 %	23.5 %	47.1 %	34.5 %	17.6 %	52.9 %
GaitBase_O [27]	26.8 %	5.9 %	47.1 %	32.6 %	17.6 %	35.3 %
GLNp1_C [26]	35.6 %	23.5 %	47.1 %	34.4 %	17.6 %	52.9 %
GLNp2_C [26]	35.0 %	17.6 %	52.9 %	39.9 %	23.5 %	58.8 %
GaitGL_C [25]	34.1 %	17.6 %	41.2 %	39.5 %	29.4 %	47.1 %
GaitGL_G [25]	**53.7 %**	35.3 %	82.4 %	**54.7 %**	35.3 %	70.6 %
GaitGL_O [25]	34.1 %	17.6 %	52.9 %	31.7 %	11.8 %	47.1 %
GaitPart_C [11]	49.3 %	35.3 %	64.7 %	50.1 %	35.3 %	76.5 %
GaitPart_G [11]	47.3 %	35.3 %	58.8 %	51.1 %	41.2 %	58.8 %
GaitPart_O [11]	33.3 %	17.6 %	47.1 %	34.8 %	23.5 %	47.1 %
GaitSet_C [10]	40.9 %	23.5 %	70.6 %	46.3 %	29.4 %	64.7 %
GaitSet_G [10]	29.4 %	11.8 %	41.2 %	32.1 %	17.6 %	47.1 %
GaitSet_O [10]	29.0 %	5.9 %	58.8 %	35.4 %	17.6 %	58.8 %

poral gaps, often under substantial changes in appearance, context, and recording conditions.

Hereafter, the experiments are presented in two stages. First, we evaluate different gait backbones independently to identify the most suitable architecture for the Re-ID task. This allows us to establish a performance baseline and analyze the standalone effectiveness of each gait model. Second, we assess the impact of incorporating HAR embeddings by comparing performance before and after their integration. As described in Section 3, the gait-only experiments use the h_i^{GAIT} (please refer to Fig. 1) outputs as identity embeddings for discrimination.

7.1. Gait analysis

The gait backbones analyzed in these experiments follow the nomenclature *BackboneID_DB*, where *BackboneID* refers to the specific appearance-based gait backbone used, and *DB* denotes the dataset it was pre-trained on: O for OU-MVLP, C for CASIA-B, and G for GREW.

Tables 2 and 3 present the performance of various gait and appearance-based models in cross-year Re-ID scenarios, specifically between recordings from 2020 and 2023. As expected, this setting is significantly more challenging than same-year evaluations due to long-term appearance changes, domain shifts, and real-world variations in clothing, posture, and lighting. Appearance-based baselines such as MobileNet V2, DenseNet 121, OSNet, and AlignedReID (highlighted in gray) exhibit low performance across all metrics. For instance, MobileNet V2 achieves 0.0 % Rank-1 in several configurations, and DenseNet never exceeds 11.8 % in Rank-1, underscoring their limited capacity to generalize temporally.

In contrast, gait-based models show substantially stronger performance. In the RP1_20 → RP2_23 evaluation, GaitGL_C achieves the highest mAP at 49.4 % and ties with GaitBase_C for the best Rank-1

Table 4
Long-term Re-ID fusing gait and HAR. Performance on cross-year evaluations between RP1_20 and RP2_23 (probe → gallery). Metrics reported as mAP / Rank-1 / Rank-5. Baseline appearance-based methods are shown in gray, the gait-based backbone under consideration is shown in blue, and the best-performing HAR model is highlighted in green.

HAR Backbone	RP1_20 → RP2_23			RP2_23 → RP1_20		
	mAP↑	R1↑	R5↑	mAP↑	R1↑	R5↑
MobileNet V2 [33]	11.6 %	0.0 %	11.8 %	20.9 %	5.9 %	29.4 %
DenseNet 121 [32]	16.9 %	0.0 %	29.4 %	24.5 %	5.9 %	35.3 %
OSNet [31]	31.9 %	17.7 %	52.9 %	28.8 %	11.8 %	47.1 %
AlignedReID [34]	28.3 %	11.8 %	47.1 %	30.1 %	17.6 %	41.2 %
GaitGL_G [25]	46.2 %	23.5 %	76.5 %	46.6 %	29.4 %	64.7 %
C2D [21]	50.2 %	35.3 %	64.7 %	50.4 %	35.3 %	64.7 %
I3D [22]	49.1 %	35.3 %	58.8 %	48.7 %	35.3 %	64.7 %
$I3D_{NLN}$ [29]	48.8 %	29.4 %	82.4 %	48.6 %	35.3 %	70.6 %
SL_4x16 [28]	53.8 %	41.2 %	76.5 %	**51.4 %**	35.3 %	64.7 %
SL_4x16_{NLN} [29]	49.2 %	35.3 %	76.5 %	50.0 %	35.3 %	64.7 %
SL_8x8 [28]	45.9 %	29.4 %	76.5 %	45.3 %	29.4 %	70.6 %
SL_8x8_{NLN} [29]	54.6 %	41.2 %	76.5 %	47.1 %	29.4 %	70.6 %
SLF_4x16 [23]	54.5 %	41.2 %	64.7 %	46.8 %	29.4 %	64.7 %
SLF_4x16_{NLN} [29]	49.5 %	29.4 %	70.6 %	47.0 %	29.4 %	82.4 %
SLF_8x8 [23]	51.9 %	29.4 %	76.5 %	50.3 %	35.3 %	76.5 %
SLF_8x8_{NLN} [29]	48.3 %	29.4 %	70.6 %	48.2 %	35.3 %	58.8 %
X3D_L [24]	**55.2 %**	41.2 %	70.6 %	**51.4 %**	35.3 %	70.6 %
X3D_M [24]	48.2 %	35.3 %	58.8 %	47.9 %	35.3 %	52.9 %
X3D_S [24]	46.5 %	29.4 %	76.5 %	47.3 %	35.3 %	58.8 %
X3D_XS [24]	48.5 %	35.3 %	76.5 %	48.0 %	35.3 %	58.8 %

Table 5
Long-term Re-ID fusing gait and HAR. Performance on cross-year evaluations between RP1_20 and RP3_23 (probe → gallery). Metrics reported as mAP / Rank-1 / Rank-5. Baseline appearance-based methods are shown in gray, the gait-based backbone under consideration is shown in blue, and the best-performing HAR model is highlighted in green.

HAR Backbone	RP1_20 → RP3_23			RP3_23 → RP1_20		
	mAP↑	R1↑	R5↑	mAP↑	R1↑	R5↑
MobileNet V2 [33]	14.6 %	0.0 %	29.4 %	14.7 %	0.0 %	26.5 %
DenseNet 121 [32]	19.2 %	0.0 %	35.3 %	23.6 %	11.8 %	23.5 %
OSNet [31]	19.9 %	0.0 %	29.4 %	17.8 %	0.0 %	29.4 %
AlignedReID [34]	27.5 %	17.6 %	29.4 %	24.1 %	11.8 %	35.3 %
GaitGL_G [25]	53.7 %	35.3 %	82.4 %	54.7 %	35.3 %	70.6 %
C2D [21]	51.1 %	35.3 %	76.5 %	56.5 %	41.2 %	76.5 %
I3D [22]	51.3 %	29.4 %	76.5 %	60.0 %	47.1 %	70.6 %
$I3D_{NLN}$ [29]	51.3 %	29.4 %	82.4 %	53.9 %	35.3 %	82.4 %
SL_4x16 [28]	50.1 %	29.4 %	82.4 %	53.8 %	41.2 %	70.6 %
SL_4x16_{NLN} [29]	52.2 %	35.3 %	76.5 %	56.1 %	41.2 %	76.5 %
SL_8x8 [28]	56.2 %	35.3 %	88.2 %	54.4 %	35.3 %	76.5 %
SL_8x8_{NLN} [29]	58.0 %	35.3 %	88.2 %	63.3 %	47.1 %	76.5 %
SLF_4x16 [23]	53.8 %	35.3 %	76.5 %	57.6 %	41.2 %	82.4 %
SLF_4x16_{NLN} [29]	57.7 %	41.2 %	82.4 %	56.4 %	41.2 %	76.5 %
SLF_8x8 [23]	53.8 %	29.4 %	88.2 %	58.7 %	41.2 %	76.5 %
SLF_8x8_{NLN} [29]	54.1 %	29.4 %	88.2 %	61.4 %	47.1 %	76.5 %
X3D_L [24]	**63.8 %**	47.1 %	82.4 %	**65.7 %**	52.9 %	82.4 %
X3D_M [24]	51.4 %	35.3 %	76.5 %	54.3 %	41.2 %	76.5 %
X3D_S [24]	58.3 %	41.2 %	88.2 %	58.3 %	41.2 %	82.4 %
X3D_XS [24]	50.9 %	29.4 %	76.5 %	60.3 %	47.1 %	70.6 %

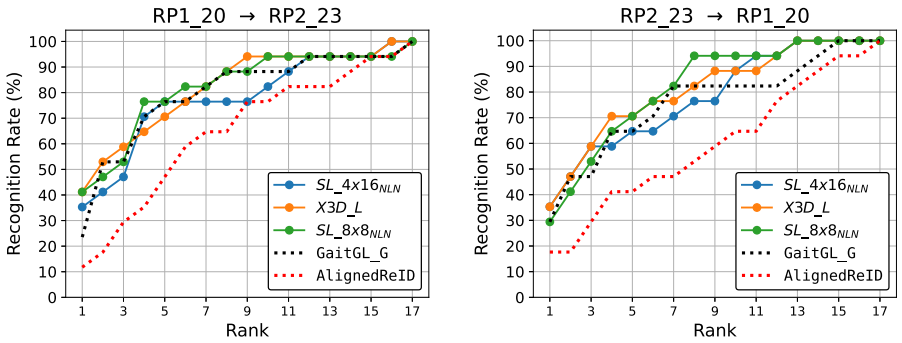


Fig. 3. CMC curves for the long-term evaluation between RP1_20 and RP2_23, comparing HAR-enhanced models. GaitGL_G is shown in black dotted, and AlignedReID (baseline) in red dotted. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

score (29.4 %), while GaitGL_G obtains the highest Rank-5 accuracy (76.5 %) and a competitive mAP (46.2 %). In the reverse direction (RP2_23 → RP1_20), GaitGL_G slightly outperforms GaitGL_C in mAP (46.6 %) and matches its Rank-1 score (29.4 %), confirming its robustness when trained on in-the-wild GREW data.

Performance generally declines in the second half of the table, covering the RP1_20 ↔ RP3_23 setting, indicating greater difficulty in this long-term pair. Still, GaitGL_G emerges as the top performer in both directions, achieving the highest mAP (53.7 % and 54.7 %), Rank-1 (35.3 %), and Rank-5 (82.4 % and 70.6 %). GaitPart_C and GaitPart_G follow closely but remain consistently behind GaitGL_G across all metrics. These findings highlight the importance of both model architecture and training dataset, with GREW-trained backbones offering superior generalization in long-term, real-world scenarios. This establishes GaitGL_G, i.e., GaitGL trained on GREW, as the most robust and reliable backbone across evaluation protocols and the most suitable candidate for further integration with HAR embeddings.

7.2. HAR integration

Tables 4 and 5 present the performance of the proposed two-stream model on the long-term Re-ID tasks, where HAR features are integrated with the gait backbone via cross-attention, as detailed in Section 3. Compared to the gait-only baseline GaitGL_G (highlighted in blue), adding HAR consistently improves performance across all metrics, particularly in mAP and Rank-1 accuracy. In the RP1_20 → RP2_23 and RP2_23 → RP1_20 scenarios, GaitGL_G achieves mAP scores of 46.2 % and 46.6 %, respectively. When fused with HAR, models such as SL_8x8$_{NLN}$, SLF_4x16, and X3D_L surpass these baselines. The best performer, X3D_L, improves mAP by nearly nine percentage points (55.2 % and 51.4 % mAP), with corresponding Rank-1 gains of up to 11.8 % on average. These results highlight the value of integrating temporal and activity-level cues for long-term person Re-ID.

In the RP1_20 → RP3_23 and RP3_23 → RP1_20 evaluations, the gap widens further. The GaitGL_G baseline yields 53.7 % and 54.7 % mAP, while X3D_L again achieves the highest scores at 63.8 % and 65.7 % mAP, representing a gain of more than ten percentage points. Similarly, Rank-1 improves from 35.3 % to 52.9 % in the RP3_23 → RP1_20 direction. Other strong HAR models like SLF_8x8$_{NLN}$, SLF_4x16$_{NLN}$, and SL_8x8$_{NLN}$ also show consistent improvements over the gait-only setting.

We also present CMC curves to analyze how HAR integration impacts person Re-ID performance. These curves visualize the rank-based retrieval accuracy of the most promising HAR-enhanced configurations (e.g., *X3D_L*, *SL_4x16*$_{NLN}$, and *SL_8x8*$_{NLN}$) in comparison with the best-performing appearance-based baseline (AlignedReID) and the gait-only backbone (GaitGL_G).

The CMC curves cover all four long-term settings: (RP1_20 ↔ RP2_23, RP1_20 ↔ RP3_23). By highlighting the probability of correctly identifying the target at various ranks, these visualizations provide deeper insight into where HAR contributes most through early rank improvements (e.g., Rank-1 and Rank-5) or enhanced overall retrieval consistency.

Fig. 3 shows the CMC curves for the long-term Re-ID between RP1_20 and RP2_23. In this more challenging setting, the integration of HAR demonstrates consistent improvements over both the gait-only and appearance-only baselines.

In the RP1_20 → RP2_23 direction, X3D_L achieves the best overall performance, with 41.2 % Rank-1 accuracy and steady improvements across higher ranks, reaching 94.1 % by Rank-9 and 100.0 % by Rank-16. Slow_8x8$_{NLN}$ and Slow_4x16$_{NLN}$ also perform well, reaching 76.5 % Rank-5 and converging to 100.0 % by Rank-17. GaitGL_G lags at Rank-1 (23.5 %) but catches up quickly by Rank-4 and matches the HAR-enhanced models in the final ranks. AlignedReID performs the worst across all ranks, achieving only 11.8 % at Rank-1 and remaining below 50 % until Rank-5.

In the reverse direction (RP2_23 → RP1_20), all three HAR-enhanced models maintain their advantage. X3D_L leads early with 35.3 % Rank-1 and reaches 88.2 % by Rank-9. Slow_4x16$_{NLN}$ performs similarly, while Slow_8x8$_{NLN}$ converges faster after Rank-7. GaitGL_G shows slightly weaker early-rank performance than the HAR models (29.4 % Rank-1) but steadily closes the gap at higher ranks. AlignedReID falls behind again, showing low early-rank accuracy (17.6 % Rank-1) and slower convergence toward 100.0 %.

These results demonstrate that incorporating HAR features usually boosts early-rank retrieval performance while contributing to more stable performance across higher ranks. The performance gap between GaitGL_G and HAR-fused models is more pronounced in the cross-year setting than in the same-year evaluation, reinforcing the robustness and long-term Re-ID capacity of HAR-enhanced architectures.

Fig. 4 illustrates the most challenging cross-year evaluation setting: RP1_20 ↔ RP3_23. Despite the increased difficulty caused by the three-year gap and different locations, HAR-enhanced models outperform the gait-only and appearance-based baselines.

In the RP1_20 → RP3_23 scenario, X3D_L achieves the strongest early-rank performance, with 47.1 % at Rank-1 and 76.5 % by Rank-3. Slow_8x8$_{NLN}$ quickly catches up at Rank-3 (82.4 %) and holds steady through Rank-9 before converging to 100.0 % at Rank-15. Slow_4x16$_{NLN}$ demonstrates slightly weaker early-rank accuracy but converges similarly by Rank-15. GaitGL_G, while showing relatively solid performance (35.3 % Rank-1), trails behind the HAR models at most ranks and only aligns with them around Rank-13. AlignedReID remains the least effective, with just 17.6 % at Rank-1 and only reaching 70.6 % by Rank-13.

In the reverse direction, RP3_23 → RP1_20, the overall trend persists. X3D_L leads with 52.9 % Rank-1, outperforming all others across early ranks. Slow_8x8$_{NLN}$ and Slow_4x16$_{NLN}$ also deliver competitive results, achieving 70.6 % and 52.9 % Rank-1, respectively. GaitGL_G main-

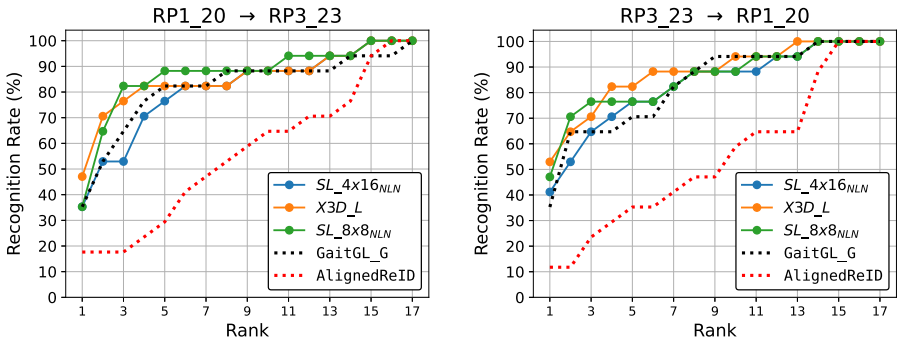


Fig. 4. CMC curves for the long-term evaluation between RP1_20 and RP3_23, comparing HAR-enhanced models. GaitGL_G is shown in black dotted, and AlignedReID (baseline) in red dotted. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

tains a solid but slightly lower trajectory, starting at 35.3 % Rank-1 and only catching up from Rank-9 onward. AlignedReID again lags across early and mid ranks, only surpassing 50 % accuracy after Rank-9.

Comparing these results with both the RP1_20 ↔ RP2_23 evaluation and the same-year setting, the improvements introduced by HAR are more substantial in these highly unconstrained, long-term scenarios. HAR-enhanced models offer better Rank-1 and Rank-5 performance and show faster convergence in the CMC curves. Among them, X3D_L emerges as the most consistent across all settings, confirming its robustness in both short-term and long-term Re-ID. These findings reinforce the value of combining gait and action-level information, mainly when dealing with significant spatio-temporal gaps and real-world variability.

8. Short-term Re-ID experiments

As expected, the short-term Re-ID experiment performs better than the long-term results discussed in the previous section. Evaluating the model across different locations on the same day provides a more favorable setting with less variation in appearance, environment, and recording conditions.

8.1. Gait analysis

Table 6 presents the results for the short-term Re-ID scenario, where the objective is to re-identify individuals across different locations within the same day. The metrics reported are mAP, Rank-1, and Rank-5 accuracy. Baseline models based on appearance cues (e.g., MobileNet V2, DenseNet, OSNet, and AlignedReID) are shown in gray. Among these, AlignedReID stands out with strong performance, indicating the strength of appearance features when clothing and context remain relatively consistent. However, when evaluating gait-based models, several backbones outperform these baselines, especially those trained on more extensive or diverse datasets. GaitGL_G, trained on the GREW dataset, achieves the highest overall performance, with an mAP of 84.7 % and a perfect Rank-5 score in the RP3_23 → RP2_23 direction. This suggests that training on unconstrained, real-world data provides a significant advantage for cross-location generalization. Other high-performing backbones include GaitPart_G and GaitBase_O, further reinforcing the importance of architecture choice and pre-training data.

Comparing Tables 2, 3 and 6, all models show a noticeable drop in performance under long-term Re-ID (Tables 2 and 3), particularly the appearance-based baselines. For example, AlignedReID, which reached 84.4 % mAP and 76.5 % Rank-1 in the short-term setting, drops below 30 % mAP and 20 % Rank-1 across most long-term evaluations. This stark contrast demonstrates the limitations of appearance cues in long-term Re-ID tasks. Gait-based methods, while also affected, maintain relatively stable performance over time. Notably, GaitGL_G is the only backbone to consistently perform well in both short-term and long-term settings, achieving 73.6 % / 64.7 % / 94.1 % in the short-term scenario and up to 54.7 % / 35.3 % / 82.4 % in the long-term scenario.

Table 6
Short-term Re-ID using only gait features. Performance on same-year cross-location evaluations (probe → gallery). Metrics reported as mAP / Rank-1 / Rank-5. Baseline performances are shown in gray, and the best-performing backbone is highlighted in blue.

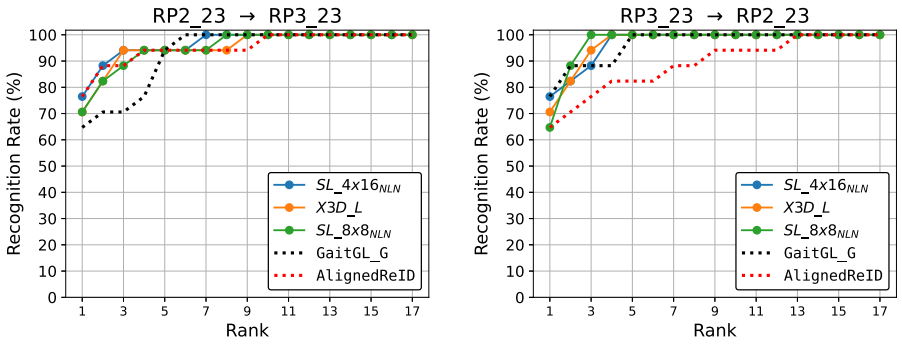
Gait Backbone	RP2_23 → RP3_23			RP3_23 → RP2_23		
	mAP↑	R1↑	R5↑	mAP↑	R1↑	R5↑
MobileNet V2 [33]	32.5 %	17.7 %	47.1 %	21.6 %	5.9 %	29.4 %
DenseNet 121 [32]	20.0 %	5.9 %	23.5 %	25.2 %	11.8 %	47.1 %
OSNet [31]	30.8 %	11.8 %	64.7 %	21.6 %	5.9 %	29.4 %
AlignedReID [34]	**84.4 %**	76.5 %	94.1 %	73.0 %	70.6 %	82.4 %
GaitBase_C [27]	42.3 %	23.5 %	64.7 %	44.6 %	29.4 %	76.5 %
GaitBase_O [27]	57.7 %	35.3 %	76.5 %	65.5 %	47.1 %	88.2 %
GLNp1_C [26]	55.2 %	41.2 %	70.6 %	46.8 %	23.5 %	82.4 %
GLNp2_C [26]	56.1 %	35.3 %	76.5 %	57.8 %	29.4 %	88.2 %
GaitGL_C [25]	59.2 %	47.1 %	82.4 %	63.2 %	47.1 %	82.4 %
GaitGL_G [25]	73.6 %	64.7 %	94.1 %	**84.7 %**	76.5 %	100.0 %
GaitGL_O [25]	49.8 %	35.3 %	70.6 %	58.8 %	41.2 %	70.6 %
GaitPart_C [11]	38.3 %	23.5 %	52.9 %	39.9 %	23.5 %	52.9 %
GaitPart_G [11]	60.5 %	47.1 %	76.5 %	76.0 %	64.7 %	94.1 %
GaitPart_O [11]	49.9 %	35.3 %	76.5 %	55.8 %	35.3 %	88.2 %
GaitSet_C [10]	58.7 %	47.1 %	82.4 %	56.9 %	41.2 %	76.5 %
GaitSet_G [10]	50.8 %	29.4 %	76.5 %	56.4 %	41.2 %	70.6 %
GaitSet_O [10]	51.9 %	35.3 %	64.7 %	51.6 %	35.3 %	70.6 %

8.2. HAR integration

Table 7 presents the results of the short-term Re-ID experiments when HAR features are integrated into the pipeline through the cross-attention mechanism described in Section 3. In this configuration, the fused embeddings f_i discriminate between identities (please refer to Fig. 1). The gait-based model GaitGL_G is used as the backbone and is highlighted in blue, while each HAR model is tested in combination with this backbone. The results demonstrate that incorporating HAR significantly enhances Re-ID performance compared to using gait alone.

For example, GaitGL_G by itself achieves 73.6 % mAP and 64.7 % Rank-1 in the RP2_23 → RP3_23 direction, and an impressive 84.7 % mAP and 76.5 % Rank-1 in the reverse direction. When HAR is added, many models exceed or match these results. Notably, the best overall performance is achieved by SL_4x16$_{NLN}$, which obtains 85.2 % mAP and 76.5 % Rank-1 in the RP2_23 → RP3_23 case, and ties with GaitGL_G at 84.7 % mAP and 76.5 % Rank-1 in reverse. Other top-performing HAR models include X3D_XS (84.3 % / 76.5 %) and X3D_L (81.0 % / 70.6 %) in the forward direction, all of which outperform the gait-only baseline.

Even lightweight HAR models such as X3D_S and C2D yield notable improvements, suggesting that HAR features consistently provide complementary motion and activity-level cues that enhance identity discrimination. Overall, these findings validate the effectiveness of the



(a) RP2_23 → RP3_23 (b) RP3_23 → RP2_23

Fig. 5. CMC curves for short-term evaluation between RP2_23 and RP3_23, comparing HAR-enhanced models. GaitGL_G is shown in black dotted, and AlignedReID (baseline) in red dotted. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 7
Short-term Re-ID results fusing gait and HAR. Performance evaluation results for same-year, cross-location Re-ID (probe → gallery). Performance is reported using mAP, Rank-1, and Rank-5 metrics. Baseline appearance-based methods are shown in gray, the gait-based backbone under consideration is shown in blue, and the best-performing HAR model is highlighted in green.

HAR Backbone	RP2_23 → RP3_23			RP3_23 → RP2_23		
	mAP↑	R1↑	R5↑	mAP↑	R1↑	R5↑
MobileNet V2 [33]	32.5 %	17.7 %	47.1 %	21.6 %	5.9 %	29.4 %
DenseNet 121 [32]	20.0 %	5.9 %	23.5 %	25.2 %	11.8 %	47.1 %
OSNet [31]	30.8 %	11.8 %	64.7 %	21.6 %	5.9 %	29.4 %
AlignedReID [34]	84.4 %	76.5 %	94.1 %	73.0 %	64.7 %	82.4 %
GaitGL_G [25]	73.6 %	64.7 %	94.1 %	**84.7 %**	76.5 %	100.0 %
C2D [21]	77.9 %	64.7 %	94.1 %	79.9 %	64.7 %	100.0 %
I3D [22]	80.4 %	70.6 %	94.1 %	75.7 %	58.8 %	100.0 %
$I3D_{NLN}$ [29]	76.7 %	64.7 %	100.0 %	81.9 %	70.6 %	94.1 %
SL_4x16 [28]	77.0 %	64.7 %	94.1 %	77.5 %	64.7 %	100.0 %
SL_4x16_{NLN} [29]	**85.2 %**	76.5 %	94.1 %	**84.7 %**	76.5 %	100.0 %
SL_8x8 [28]	83.6 %	76.5 %	94.1 %	80.4 %	64.7 %	100.0 %
SL_8x8_{NLN} [29]	80.6 %	70.6 %	94.1 %	80.4 %	64.7 %	100.0 %
SLF_4x16 [23]	71.1 %	58.8 %	88.2 %	**84.7 %**	76.5 %	100.0 %
SLF_4x16_{NLN} [29]	71.3 %	58.8 %	94.1 %	81.6 %	70.6 %	100.0 %
SLF_8x8 [23]	78.6 %	64.7 %	100.0 %	82.4 %	70.6 %	100.0 %
SLF_8x8_{NLN} [29]	76.7 %	64.7 %	94.1 %	77.5 %	58.8 %	100.0 %
$X3D_L$ [24]	81.0 %	70.6 %	94.1 %	81.9 %	70.6 %	100.0 %
$X3D_M$ [24]	65.1 %	47.1 %	94.1 %	77.0 %	58.8 %	100.0 %
$X3D_S$ [24]	80.0 %	70.6 %	94.1 %	81.4 %	70.6 %	100.0 %
$X3D_{XS}$ [24]	84.3 %	76.5 %	94.1 %	81.9 %	70.6 %	94.1 %

cross-attention fusion strategy and demonstrate that augmenting gait with HAR improves robustness in short-term Re-ID tasks.

When comparing these results to those in Tables 4 and 5 (long-term Re-ID with HAR), we observe that performance also drops across the board in the long-term setting, which is expected due to the three-year gap and environmental differences between the recordings. For example, the best mAP in the short-term scenario reaches 85.2 % (SL_4x16_{NLN}), while in the most challenging temporal case (RP1_20 → RP3_23), the best mAP is 63.8 % (X3D_L). However, the relative improvements introduced by HAR remain consistent in both scenarios, validating the generalization capacity of fused motion/activity-based representations. Moreover, HAR-enhanced models consistently outperform the GaitGL_G baseline (in blue) in all short-term and long-term cases, establishing HAR integration-especially with two models such as X3D_L and SLF_8x8_{NLN}, demonstrating an effective and resilient solution for long-term person Re-ID in real-world settings.

The CMC curves for the short-term Re-ID cover two settings: RP2_23 ↔ RP3_23. By highlighting the probability of correctly identifying the target at various ranks, these visualizations provide deeper insight into where HAR contributes most through early rank improvements (e.g., Rank-1 and Rank-5) or enhanced overall retrieval consistency.

Fig. 5 presents the CMC curves for the short-term Re-ID evaluations. Across both directions, integrating HAR features leads to consistent performance gains over the gait-only and appearance-only baselines.

In the RP2_23 → RP3_23 direction, $Slow\_4x16_{NLN}$ (SL_4x16_{NLN}) achieves the best early-rank performance with 76.5 % Rank-1 and 88.2 % Rank-2 accuracy, converging to 100.0 % by Rank-7. X3D_L and $Slow\_8x8_{NLN}$ (SL_8x8_{NLN}) show very competitive results, reaching 70.6 % Rank-1 and also converging to 100.0 % by Rank-9. Notably, GaitGL_G lags slightly behind at early ranks (64.7 % Rank-1, 70.6 % Rank-2) but rapidly improves after Rank-4. AlignedReID shows strong Rank-1 performance (76.5 %) but flattens slightly before reaching 100.0 %, indicating weaker retrieval consistency at higher ranks compared to HAR-enhanced models.

In the reverse direction (RP3_23 → RP2_23), the trend continues with all HAR-enhanced models surpassing both baselines. $Slow\_8x8_{NLN}$ achieves the fastest convergence, reaching 100.0 % by Rank-3. X3D_L and $Slow\_4x16_{NLN}$ follow closely, also achieving 100.0 % by Rank-4. While GaitGL_G maintains a strong baseline (76.5 % Rank-1), it converges more slowly than the HAR-fused models. AlignedReID again performs the weakest at early and mid ranks, achieving only 64.7 % Rank-1 and not reaching 100.0 % until Rank-13.

These results confirm that HAR integration improves early retrieval accuracy and convergence speed in short-term Re-ID tasks. Among the HAR-enhanced models, $Slow\_4x16_{NLN}$, X3D_L, and $Slow\_8x8_{NLN}$ consistently outperform the gait-only and appearance-only baselines, reinforcing the value of motion-aware representations in realistic, location-varying settings.

9. Ablation study on fusion strategies

To further analyze the role of the fusion mechanism in our framework, we provide an ablation study using the best-performing gait backbone (GaitGL trained on GREW) and the best-performing HAR backbone (X3D-L) identified in our previous experiments. We evaluate three alternative fusion strategies: cross-attention, cosine similarity-based fusion, and concatenation. Each represents a distinct way of integrating gait and HAR embeddings. Cross-attention adaptively emphasizes behavioral cues conditioned on gait, while cosine similarity enforces alignment by maximizing directional closeness between modalities. In contrast, concatenation stacks the embeddings without explicit interaction, serving as a baseline.

All variants were trained following the same protocol described in Section 3, using triplet loss with semi-hard negative mining and the modality-aware distance formulation to balance the contribution of gait and HAR streams. By fixing the underlying encoders to their

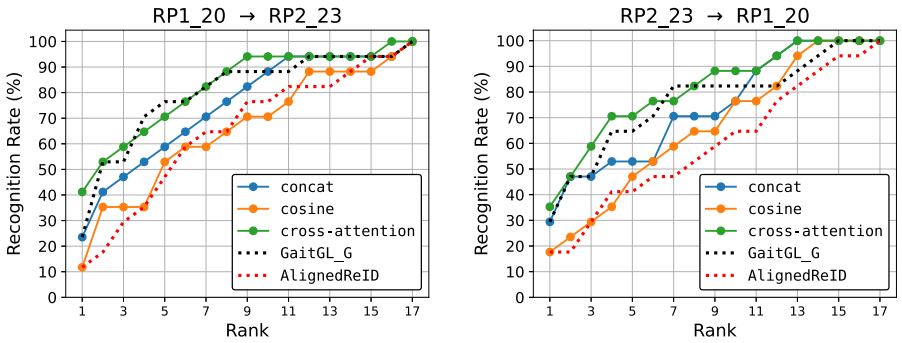


Fig. 6. CMC curves for long-term evaluation (RP1_20–RP2_23) with HAR model X3DL under different fusion strategies. Solid lines show the best model per fusion; dotted lines denote reference methods: GaitGL_G (black) and AlignedReID (red).. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

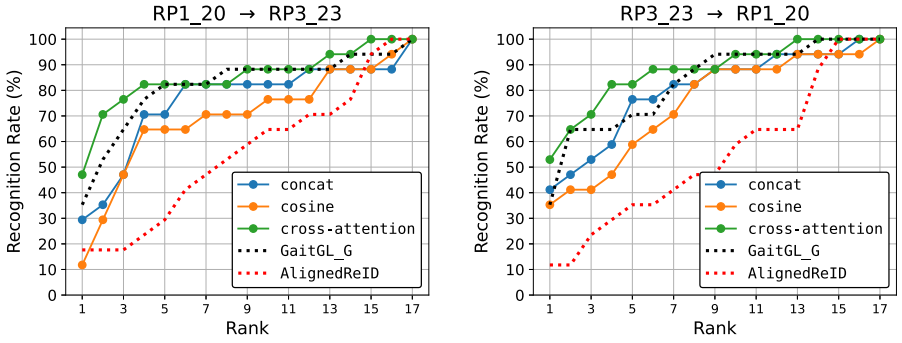


Fig. 7. CMC curves for long-term evaluation (RP1_20–RP3_23) with HAR model X3DL under different fusion strategies. Solid lines show the best model per fusion; dotted lines denote reference methods: GaitGL_G (black) and AlignedReID (red).. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

strongest configurations, this ablation isolates the impact of the fusion strategy itself. The comparative results quantify the impact of each method on retrieval performance in both short-term and long-term settings, highlighting the advantages of attention-based fusion over simpler alternatives.

In the long-term ReID evaluation, cross-attention consistently outperformed the other fusion strategies across all transfer scenarios. For RP2_23→RP1_20 and the reverse direction (Fig. 6), cross-attention achieved higher mAP values (51.4 % and 55.2 %, respectively) compared to concat (44.8 % and 41.6 %) and cosine (31.9 % and 31.7 %). Similar improvements were observed in transfers involving RP3_23 (Fig. 7), where the cross-attention fusion reached 63.8 % mAP for RP1_20→RP3_23 and 65.7 % for RP3_23→RP1_20, clearly surpassing concat (45.3 % and 53.4 %) and cosine (33.9 % and 46.8 %). In all cases, cross-attention also outperformed the baselines GaitGL_G and AlignedReID, demonstrating superior generalization capability in long-term cross-environment ReID.

10. Conclusion

In this work, we propose a two-stream architecture for person Re-ID that jointly models gait and HAR features. Our framework leverages existing components (BiLSTMs, pre-trained HAR extractors, attention) but is original in how they are integrated and adapted to the long-term Re-ID challenge. The model learns discriminative embeddings by treating gait as the primary identity signal and enriching it with activity cues through cross-attention, thereby improving both short-term and long-term performance.

We also introduce a real-world long-term Re-ID dataset from two editions (2020, 2023) of an ultra-distance sporting event. It captures natural variations across years and locations, enabling realistic longitudinal evaluation.

Our fused gait-HAR model outperformed appearance-based and gait-only baselines, with +12 % mAP in long-term and +11.6 % in short-term setups. HAR-enhanced models also improved early-rank retrieval, confirming the value of multimodal fusion in unconstrained conditions. Limitations include reliance on reliable silhouettes and a HAR branch based on pre-trained backbones. The test set comprises 17 identities and 51 videos, resulting in 1734 comparisons under a subject-disjoint protocol, which ensures a meaningful and comprehensive assessment.

This two-stream paradigm opens avenues for broader multimodal identity modeling. Future work may expand datasets to more diverse contexts, validating generalization. Overall, we present a robust framework that bridges motion understanding and identity recognition in real-world Re-ID.

Declarations

'Funding' and/or 'Competing interests' This work is partially funded funded by project PID2021-122402OB-C22/MICIU/AEI /10.13039/501100011033 FEDER, UE and by the ACIISI-Gobierno de Canarias and European FEDER funds under project ULPGC Facilities Net and Grant EIS 2021 04.

CRediT authorship contribution statement

David Freire-Obregón: Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Conceptualization; **Oliverio J. Santana :** Writing – review & editing, Writing – original draft, Validation, Supervision, Data curation, Conceptualization; **Javier Lorenzo-Navarro:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Conceptualization; **Daniel Hernández-Sosa:** Writing – review & editing, Visualization, Validation, Supervision, In-

vestigation, Conceptualization; **Modesto Castrillón-Santana:** Writing – review & editing, Validation, Supervision, Investigation, Formal analysis, Conceptualization.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A.K. Jain, A. Ross, Introduction to Biometrics, Springer US, Boston, MA, Boston, MA, 2008.

[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, S.C.H. Hoi, Deep learning for person re-identification: a survey and outlook, IEEE Trans. Pattern Anal. Mach. Intell. 44 (6) (2022) 2872–2893.

[3] P. Xu, X. Zhu, DeepChange: a long-term person re-identification benchmark with clothes change, in: Proceedings of the IEEE international conference on computer vision (ICCV), p. 2023.

[4] G. Zhang, J. Zhou, Y. Zheng, G. Martin, R. Wang, Adaptive transformer with pyramid fusion for cloth-changing person re-identification, Pattern Recognit. 163 (2025) 111443.

[5] Y. Makihara, D. Muramatsu, Y. Yagi, Gait Recognition: Databases, Representations, and Applications, Wiley Encyclopedia of Electrical and Electronics Engineering, 2020.

[6] M. Kim, M. Cho, H. Lee, S. Lee, Spatio-temporal feature-level augmentation vision transformer for video-based person re-identification, Pattern Recognit. 168 (2025) 111813.

[7] A. Zahra, N. Perwaiz, M. Shahzad, M.M. Fraz, Person re-identification: a retrospective on domain specific open challenges and future trends, Pattern Recognit. 142 (2023) 109669.

[8] Y. Wu, Y. Lin, X. Dong, et al., Tracklet-Based Graph Structure Learning for Robust Video-Based Person Re-Identification, CVPR, 2020.

[9] M.S. Nixon, T. Tan, R. Chellappa, Human Identification Based on Gait, Springer Science & Business Media, 2010.

[10] H. Chao, Y. He, J. Zhang, J. Feng, GaitSet: regarding gait as a set for cross-view gait recognition, in: AAAI Conference on Artificial Intelligence, 2018.

[11] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, Z. He, GaitPart: temporal part-based model for gait recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR, 2020, pp. 14213–14221.

[12] D. Freire-Obregón, P. Barra, M. Castrillón-Santana, M.D. Marsico, Exploring Biometric Domain Adaptation in Human Action Recognition Models for Unconstrained Environments, Multimedia Tools and Applications, 2024.

[13] D. Freire-Obregón, J. Lorenzo-Navarro, O.J. Santana, D. Hernández-Sosa, M. Castrillón-Santana, A large-scale re-identification analysis in sporting scenarios: the betrayal of reaching a critical point, in: 2023 IEEE International Joint Conference on Biometrics (IJCB), 2023, pp. 1–9.

[14] S. Yu, D. Tan, T. Tan, A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, 18th Int. Conf. Pattern Recognit. (ICPR'06) 4 (2006) 441–444.

[15] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition, IPSJ Trans. Comput. Vis. Appl. 10 (2018) 1–14.

[16] Z. Zhu, X. Guo, T. Yang, J. Huang, J. Deng, G. Huang, D. Du, J. Lu, J. Zhou, Gait recognition in the wild: a benchmark, in 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, 14769–14779, https://doi.org/10.1109/ICCV48922.2021.01452.

[17] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, N. Wang, Gait recognition via disentangled representation learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4705–4714.

[18] G. Jocher, A. Chaurasia, J. Qiu, Ultralytics YoLov 8 (2023). https://github.com/ultralytics/ultralytics.

[19] N. Aharon, R. Orfaig, B.-Z. Bobrovsky, Bot-Sort: Robust associations multi-pedestrian tracking, Technical Report, arXiv preprint, 2022.

[20] C.-Y. Yang, H.-W. Huang, W. Chai, Z. Jiang, J.-N. Hwang, SAMURAI: Adapting Segment Anything Model for Zero-Shot Visual Tracking with Motion-Aware Memory, 2024. https://arxiv.org/abs/2411.11922.

[21] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Proc. Int. Conf. Neural Inf. Process. Syst. 1 (2014) 568–576.

[22] J. Carreira, A. Zisserman, Q. Vadis, Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, in 2017 IEEE conference on computer vision and pattern recognition (CVPR). Action Recognition? A New Model and the Kinetics Dataset, 2017, 4724–4733, https://doi.org/10.1109/CVPR.2017.502.

[23] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: IEEE/CVF International Conference on Computer Vision (ICCV, 2018, pp. 6201–6210.

[24] C. Feichtenhofer, X3D: expanding architectures for efficient video recognition, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 200–210.

[25] B. Lin, S. Zhang, X. Yu, Gait recognition via effective global-local feature representation and local temporal aggregation, in: IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14628–14636.

[26] S. Hou, C. Cao, X. Liu, Y. Huang, Gait lateral network: learning discriminative and compact representations for gait recognition, in European Conference on Computer Vision, (2020), p. 382–398.

[27] C. Fan, J. Liang, C. Shen, S. Hou, Y. Huang, S. Yu, OpenGait: revisiting gait recognition toward better practicality, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9707–9716.

[28] C. Feichtenhofer, H. Fan, B. Xiong, R.B. Girshick, K. He, A large-scale study on unsupervised spatiotemporal representation learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3298–3308.

[29] X. Wang, R.B. Girshick, A.K. Gupta, K. He, Non-local neural networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 7794–7803.

[30] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, CoRR abs/1705.06950, 2017.

[31] K. Zhou, Y. Yang, A. Cavallaro, T. Xiang, Learning generalisable omni-scale representations for person re-identification, IEEE Trans. Pattern Anal. Mach. Intell. 44 (2019) 5056–5069.

[32] G. Huang, S. Liu, L.V.D. Maaten, K.Q. Weinberger, CondenseNet: an efficient densenet using learned group convolutions, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 2752–2761.

[33] M. Sandler, A.G. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 2 (2018) 4510–4520.

[34] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, C. Zhang, Alignedreid : dynamically matching local information for person re-identification, Pattern Recognit. 94 (2019) 53–61.