algorithms

MDPI

Article

In-Depth Analysis of GAF-Net: Comparative Fusion Approaches in Video-Based Person Re-Identification

Moncef Boujou [1,*], Rabah Iguernaissi [1], Lionel Nicod [2], Djamal Merad [1,*] and Séverine Dubuisson [1]

[1] LIS, CNRS, Laboratoire d'Informatique et des Systèmes, Centre National de la Recherche Scientifique, Aix Marseille University, 13284 Marseille, France; rabah.iguernaissi@univ-amu.fr (R.I.); severine.dubuisson@univ-amu.fr (S.D.)
[2] CERGAM, CERGAM, Centre d'études et de recherche en gestion d'Aix Marseille, Aix Marseille University, 13284 Marseille, France; lionel.nicod@univ-amu.fr
* Correspondence: moncef.boujou@univ-amu.fr (M.B.); djamal.merad@univ-amu.fr (D.M.)

Abstract: This study provides an in-depth analysis of GAF-Net, a novel model for video-based person re-identification (Re-ID) that matches individuals across different video sequences. GAF-Net combines appearance-based features with gait-based features derived from skeletal data, offering a new approach that diverges from traditional silhouette-based methods. We thoroughly examine each module of GAF-Net and explore various fusion methods at the both score and feature levels, extending beyond initial simple concatenation. Comprehensive evaluations on the iLIDS-VID and MARS datasets demonstrate GAF-Net's effectiveness across scenarios. GAF-Net achieves state-of-the-art 93.2% rank-1 accuracy on iLIDS-VID's long sequences, while MARS results (86.09% mAP, 89.78% rank-1) reveal challenges with shorter, variable sequences in complex real-world settings. We demonstrate that integrating skeleton-based gait features consistently improves Re-ID performance, particularly with long, more informative sequences. This research provides crucial insights into multi-modal feature integration in Re-ID tasks, laying a foundation for the advancement of multi-modal biometric systems for diverse computer vision applications.

Keywords: person re-identification; gait recognition; feature fusion; deep learning; computer vision

Citation: Boujou, M.; Iguernaissi, R.; Nicod, L.; Merad, D.; Dubuisson, S. In-Depth Analysis of GAF-Net: Comparative Fusion Approaches in Video-Based Person Re-Identification. *Algorithms* **2024**, *17*, 352. https://doi.org/10.3390/a17080352

Academic Editor: Paolo Spagnolo

Received: 26 June 2024
Revised: 5 August 2024
Accepted: 7 August 2024
Published: 11 August 2024

1. Introduction

Person re-identification (Re-ID) is a crucial task in computer vision with a wide range of applications, from security and public safety [1,2] to retail analytics [3] and smart city implementations [4]. This task involves re-identifying individuals across different camera views by analyzing visual cues such as clothing, body shapes, and other attributes. Despite its significance, Re-ID faces substantial challenges, including variations in lighting, background clutter, and the presence of individuals with similar appearances. These factors reduce the effectiveness of Re-ID systems, particularly in dynamic environments where factors are uncontrollable.

The primary challenge in Re-ID is to develop robust architectures capable of accurately matching a query image ($\mathbf{q}$) with the correct identity in a gallery set ($\mathcal{G} = \{\mathbf{g}_i \mid i = 1, 2, \ldots, N\}$). Recent trends have shifted towards video-based methodologies, leveraging datasets such as iLIDS-VID [5] and MARS [6]. These video sequences provide richer spatial and temporal details compared to static images. Advancements in deep learning have significantly improved Re-ID through the use of methods such as hypergraphs [7], attention mechanisms [8], pyramidal spatial–temporal aggregation [9], and vision transformers [10,11]. However, these approaches primarily focus on appearance-based cues, leaving the challenge of similar appearances unresolved.

To address these limitations, our research extends foundational studies by merging appearance-based features with gait features—a biometric modality essential for capturing distinctive walking patterns for individual identification. Traditionally, gait recognition

has relied on silhouette-based methods [12] that focus on a person's shape and movement. However, recent advances in pose estimation have made skeletal-based gait analysis more feasible and effective. For instance, Teepe et al. [13] demonstrated highly accurate results using skeletal-based gait in controlled environments. Such advancements have now made skeletal-based approaches both viable and promising for Re-ID tasks.

Building on these developments, we present an improved version of GAF-Net [14], which, to the best of our knowledge, is the first method to fuse skeletal-gait information with appearance-based data for comprehensive individual representation. Our paper refines this model by thoroughly exploring the original architecture and introducing advanced fusion techniques at both the feature and score levels. We investigate various approaches, including conditional concatenation and PCA-enhanced methods, to optimize the integration of appearance and gait features.

We evaluate GAF-Net on both the iLIDS-VID and MARS datasets. The MARS dataset presents more challenging, real-world scenarios with shorter and more variable sequences, allowing us to assess GAF-Net's performance in diverse Re-ID contexts. This cross-dataset analysis provides insights into GAF-Net's effectiveness across different scenarios, ranging from long, informative sequences to challenging real-world conditions. While our results are promising, we also identify limitations, particularly in handling shorter, variable-length sequences. These findings guide our future research directions.

Our research advances the field of algorithms and machine learning by introducing novel computational methods for video-based person re-identification . GAF-Net's innovative approach to feature extraction, fusion, and re-identification enhances the efficiency of multimodal data processing algorithms. This study provides valuable insights into algorithmic design for multimodal systems, addressing key challenges in computer vision and machine learning. By combining theoretical analysis with practical application, we offer a comprehensive framework for implementing advanced algorithms in this domain. These innovations have potential applications in various areas of computer vision and pattern recognition, where multimodal data fusion is essential, contributing to the advancement of state-of-the-art approaches in image understanding and analysis.

The key contributions of this paper include the following:

- Advanced fusion: We introduce more fusion techniques at both the feature and score levels, further enhancing Re-ID accuracy by leveraging the complementary strengths of appearance and gait data and analyzing their effectiveness across different dataset characteristics.
- In-depth theoretical analysis: We provide a deeper exploration of the foundational concepts of our approach, offering a comprehensive analysis of the models and fusion strategies, as well as their impacts on the Re-ID process.
- Comprehensive dual-dataset evaluation: We provide an in-depth analysis of GAF-Net's performance on both the iLIDS-VID and MARS datasets, offering insights into its efficacy in various Re-ID scenarios.
- Enhanced performance: The improved GAF-Net achieves state-of-the-art results on the iLIDS-VID dataset and competitive performance on the MARS dataset.

This paper is organized as follows: Section 2 reviews related works, Section 3 introduces the proposed method, Sections 4 and 5 present experimental results and analyses, and Section 6 concludes the study.

2. Related Works

The primary challenge in person Re-ID is developing effective methods that accurately represent images or videos for individual identification. This section explores the follow three main approaches: those solely based on appearance data, those focusing on gait data, and those combining both types of information.

2.1. Appearance-Based Person Re-ID

Video-based person Re-ID is a complex task that requires the identification of individuals across varying camera views. This challenge involves the fusion of spatial and temporal features. Recent developments in deep learning architectures have enhanced performance substantially, significantly surpassing classical methods.

Convolutional Neural Networks (CNNs), as exemplified in studies by Suh et al. [15] and Zhou et al. [16], are mainly used to extract appearance features from video sequences. However, they may occasionally overlook crucial discriminative spatial and temporal details. To capture sequence-level representations, Recurrent Neural Networks (RNNs) were utilized by McLaughlin et al. [17] to preserve information from previous frames. Furthermore, 3D-CNNs were used by Li et al. [18] to derive a holistic capture of both spatial and temporal cues.

Attention-driven models, such as those presented by Fu et al. [19], integrate spatial and temporal attention mechanisms to derive latent individual representations. Addressing these limitations, in [20] Bai et al. proposed the innovative Salient-to-Broad Module (SBM) and the Integration-and-Distribution Module (IDM), which expand attention regions within video frames to enhance frame-level representations. Despite these advances, such methods still face challenges in real-world conditions.

Graph-based approaches have gained prominence in video-based person Re-ID, as highlighted in research by Yang et al. [21]. These techniques excel at uncovering complex interconnections within data, a crucial aspect of enhancing discriminative feature representation. A prime example of this approach is the MGH framework proposed by Yan et al. [7]. This multi-granular hypergraph learning framework uses spatial and temporal cues extracted from tracklets, employing a hypergraph neural network. The model incorporates an attention mechanism during the learning process to focus on node-level feature aggregation, which leads to more distinct graph representations.

Furthermore, Pyramid Spatial–Temporal Aggregation (PSTA) [9] identifies spatial correlations within frames and uses information on temporal consistency. This method highlights discriminative features while suppressing irrelevant ones.

Recently, Vision Transformers (ViTs), as proposed by Dosovitskiy et al. [22], represent a major advancement. Architectures like PiT, introduced by Zang et al. [10], are noteworthy for their innovative use of transformers with pyramidal structures that efficiently process information across various patches, offering a nuanced understanding of visual data.

Despite these technological advances, some practical challenges remain unresolved. A notable difficulty is the accurate differentiation of individuals who share similar clothes, an issue that frequently arises in real-world scenarios and adds complexity to the task of effective person Re-ID. This challenge necessitates that current methods go beyond simple pattern recognition and develop more sophisticated approaches to extract and analyze unique identifiers from a person's appearance that are not solely dependent on clothing.

2.2. Gait Recognition

As we shift focus from appearance-based Re-ID to biometric modalities, gait recognition emerges as a pivotal method in computer vision for identifying individuals. This approach relies on analyzing unique spatiotemporal patterns in a person's walking style, extracting distinct features that define an individual's gait.

Gait recognition methodologies are categorized into appearance-based and model-based methods. Appearance-based approaches predominantly use silhouettes or body shapes for identification, as highlighted in [12,23]. These methods can capture the visual elements of gait, providing a surface-level analysis.

On the other hand, model-based approaches examine the structural aspects of the body, constructing two-dimensional (2D) or three-dimensional (3D) skeletal models, as demonstrated by Liao et al. in [24]. Advances in pose estimation have significantly improved the accuracy of these methods.

In the domain of gait recognition, these strategies further branch into different temporal representations, which are essential for capturing the dynamic nature of an individual's walk, such as image-based, sequence-based, and set-based techniques.

In gait recognition, image-based methods, such as the Gait Entropy Images (GEI) approach proposed by Babaee et al. [25], aim to capture an entire gait cycle in a single image. While these methods are efficient, they can sometimes overlook important dynamic features of gait. To address this, sequence-based methods like the 3D-CNNs proposed by Thapar et al. [26] and the LSTMs proposed by Liao et al. [24] focus on capturing the temporal evolution of gait, yet they require more computational resources. Set-based methods, as demonstrated in the work of Chao et al. [27], blend various data inputs but do not explicitly model the temporal aspect, thereby achieving a balance between detail and computational efficiency. In this context, a study conducted by Fendri et al. [28] introduced an innovative approach for gait-based person Re-ID that combines signature extraction with matching techniques to analyze temporal gait elements. GaitPart, as presented by Fan et al. [12], is a significant contribution to gait recognition, introducing a temporal module specifically designed to capture short-term temporal features within the gait cycle. This approach concentrates on subdividing the gait sequence into smaller, more manageable segments, facilitating a thorough examination of the gait pattern. This method provides deeper insights into the subtle variations in gait that occur over short intervals, rendering it a valuable tool for conducting more refined gait analysis.

In model-based techniques, a recent development is the work of Rao et al. [29], who proposed a novel self-supervised learning approach using 3D skeletal models for gait recognition. Traditional methods like CNNs, as shown in [30], and combinations of CNNs and LSTMs, as in [31], remain fundamental in the field. The use of graph convolutional networks, as explored by Teepe et al. in [13], represents an advanced technique for capturing human motion in gait analysis.

It is important to note that while these methods are effective in controlled environments, such as those present in the CASIA-B dataset [32], their performance can vary significantly in more complex, real-world settings.

2.3. Fusion of Appearance and Gait-Based Person Re-ID

The fusion of gait and appearance features in person Re-ID has shown promise, yet only a limited number of studies have explored this approach in depth.

A pioneering work was proposed by Bedagkar-Gala et al. in [33], where the authors used Gait Energy Images (GEI) to capture key characteristics from silhouettes, integrating those with color features to enhance long-term person Re-ID. Similarly, Liu et al. merged GEI with color and texture features in [34], while Frikha et al. [35] developed a bi-modal Re-ID method that encodes visual appearance characteristics based on color, texture, and semantic attributes, as well as gait information, using GEI and local color and texture characteristics. They employed a score-level fusion technique to combine these features for enhanced Re-ID accuracy.

Recent works relying on deep learning architectures have combined silhouette-derived gait attributes with sophisticated appearance feature extraction methods. For example, in [36], Lu et al. proposed a model that merges appearance features extracted using ResNet with gait features obtained from an Improved Sobel-Masking Active Energy Image (ISMAEI) model, providing a thorough representation of gait.

Addressing challenges in scenarios with long-term clothing changes, Jin et al. introduced GI-REID [37], a dual-stream framework that integrates a gait recognition stream based on silhouettes with an image-based Re-ID stream. This approach demonstrates the efficacy of combining different biometric features to enhance person Re-ID.

Building on this, the work by Tu et al. reported in [38] focused on combining gait recognition with RGB modalities, particularly in scenarios involving changes in clothing. This study leveraged silhouettes for gait recognition, reflecting the constant evolution of gait-based identification methods. The authors employed a unique fusion technique

that combines dynamic gait information extracted from silhouettes with static appearance features from RGB data, enabling effective identification despite clothing variations.

Furthermore, Soni et al. [39] integrated visual appearance and gait features using a novel approach that utilizes silhouettes for gait analysis. This method demonstrated its ability to handle complex indoor scenarios. This paper innovatively applied graph convolutional networks to integrate spatiotemporal gait features from silhouettes with high-level appearance characteristics, offering a more robust and context-aware approach to person Re-ID in diverse indoor environments.

In light of these advancements, we propose a novel approach that combines visual appearance features with skeleton-based gait features, moving beyond traditional silhouette-based methods to address the challenge of person Re-ID.

3. Proposed Method

Video-based person Re-ID focuses on establishing correlations between individuals appearing in video segments captured from different camera viewpoints. This task essentially revolves around metric learning, where the goal is to develop a function (ϕ) to effectively map tracklets into a metric space, bringing similar tracklets closer together. This section introduces the Gait–Appearance Fusion Network (GAF-Net), a novel framework made up of the following three main components: (1) a gait feature extraction module, (2) an appearance feature extraction module, and (3) a gait and appearance feature fusion module. GAF-Net uses GaitGraph for gait feature extraction. This method combines skeleton pose estimation with a Graph Convolutional Network (GCN) to efficiently extract gait attributes. Appearance feature extraction is performed with various backbone architectures to extract image-level features. Finally, the fusion module combines these appearance and gait features to derive a unified and normalized representation of the individual. Figure 1 graphically illustrates the global architecture of GAF-Net. Hereafter, we consider a person labeled with identifier i. This person is represented by a series of T RGB images of size $W \times H$.

This series can be seen as a four-dimensional (4D) tensor, denoted as $\mathbf{A}^i$, as defined by $\mathbb{R}^{W \times H \times 3 \times T}$.

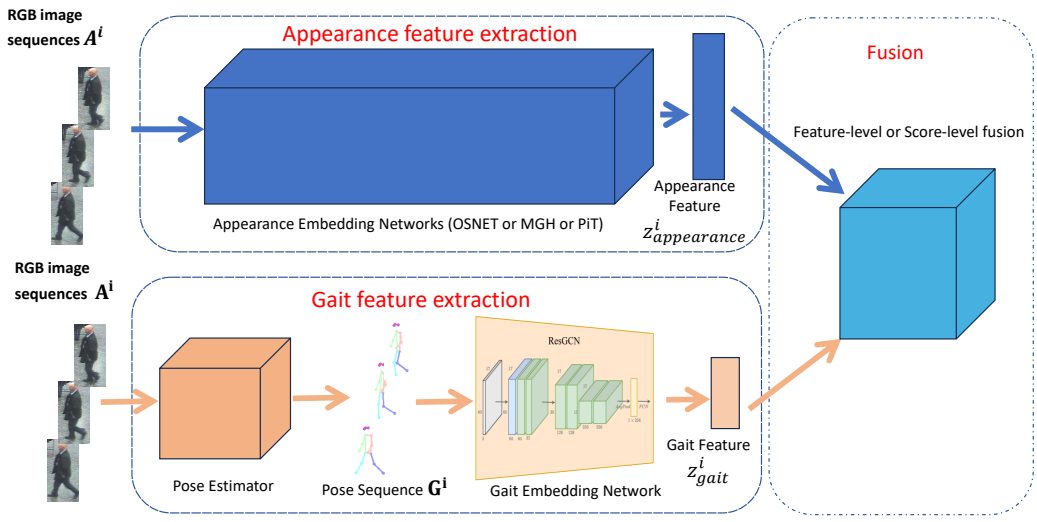

Figure 1. A schematic representation of Improved GAF-Net illustrating its three main modules, namely the appearance feature module (with various backbones), the gait feature module, and the fusion module.

3.1. Gait Feature Extraction Module

To obtain gait features, we used the approach proposed by Teepe et al. [13]. This approach first estimates the human pose, which is subsequently utilized as input for the ResGCN [40] model to construct a graph-based representation.

3.1.1. Skeleton Graph Representation of Gait

In our approach, we create a graph representation of a person's gait by estimating the 2D poses for each frame, which we then integrate over the sequence for a complete gait representation. In each frame, key points like the nose, eyes, and knees are detected using a 2D pose estimator. These key points collectively define the human skeleton, which is represented as a graph denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ corresponds to the nodes (e.g., joints) and $\mathcal{E}$ to the edges (e.g., bones).

For $2D$ human pose estimation, we utilize YOLO-pose [41] pre-trained on the COCO dataset [42] rather than the more computationally expensive HRNET [43]. YOLO-pose identifies $N = 17$ key points representing the human body's pose.

These 2D poses are then used to construct the graph. We represent each joint with a C-dimensional feature vector ($\mathbf{g}_{t,n}$) containing the position of the n-th joint in frame t, ($n \in [1, N]$ and $t \in [1, T]$) and an associated confidence value.

Thus, the gait for person i becomes a 3D tensor denoted as

$$\mathbf{G}^i \in \mathbb{R}^{T \times N \times C}, \text{ defined as: } \mathbf{G}^i = \left\{ \mathbf{g}^i_{t,n} \in \mathbb{R}^C \mid t, n \in \mathbb{Z}, 1 \leq t \leq T, 1 \leq n \leq N \right\}. \quad (1)$$

This tensor comprehensively captures the joint-specific feature vectors for each frame (t) and joint (n) and across the various dimensions (C), providing a detailed representation of the gait pattern for person i.

For a specific frame (t) and person (i), the corresponding pose can be extracted from the tensor as a matrix ($G^i_t \in \mathbb{R}^{N \times C}$). It is important to note that in our case, we are working with 2D positions, which implies $C = 3$ (position plus confidence coefficient) due to the inclusion of the confidence value.

3.1.2. Vector Representation of Gait

To construct a more compact and comprehensive vector representation of gait, we used the same pipeline as in GraphGait [13]. At the core of our model lies the ResGCN network, as introduced by Pei et al. in [40] and adapted for our specific context. The ResGCN network processes feature tensors ($\mathbf{G}^i$), which are the graph representation of gait information (see Section 3.1.1), to obtain a vector representation ($\mathbf{z}^i_{\text{gait}}$). ϕ^{gait} the function that transforms $\mathbf{G}^i$ into $\mathbf{z}^i_{\text{gait}}$.

The structural design of each ResGCN block blends advanced graph convolution layers with traditional 2D temporal convolution layers. This combination allows for the processing of both spatial and temporal features. In the architecture, a residual connection is integrated to maintain the fidelity of the input data as it progresses through the network layers. Additionally, there is an optional bottleneck structure designed to optimize the processing of features. This network comprises multiple ResGCN blocks sequenced in such a way that each block enhances the output of its predecessor.

After these ResGCN blocks are processed, the data undergo an average pooling operation. This step aggregates the features across the entire network, distilling them into a more manageable form. Following this pooling, a fully connected layer takes over, transforming the pooled features into a final, comprehensive feature vector representing gait.

The network is trained with the supervised contrastive (SupCon) loss function proposed by Khosla et al. in [44] to effectively differentiate between gait patterns.

For each individual (i), ResGCN converts the gait graph ($\mathbf{G}^i$) into a vector ($\mathbf{z}^i_{\text{gait}}$) that captures the person's gait characteristics over T frames in a compact and analyzable form, as expressed by the following equation:

$$\mathbf{z}^i_{\text{gait}} = \phi^{\text{gait}}(\mathbf{G}^i) \quad (2)$$

This equation represents the transformation of the gait graph data into a vector that captures an individual's gait patterns, enabling effective gait-based person Re-ID.

3.2. Appearance Feature Extraction Module

To extract appearance features, we utilize prominent backbone architectures, including ResNet-50 [45], the OSNet framework [16], precision-optimized MGH [7], and the high-dimensional PiT model [10]. These models yield an appearance vector ($\mathbf{z}^i_{\text{appearance}}$) encoding the visual traits of an individual (i) from tracklet $\mathbf{A}^i$:

$$\mathbf{z}^i_{\text{appearance}} = \phi^{\text{appearance}}(\mathbf{A}^i) \tag{3}$$

In developing GAF-Net, we chose to focus on a subset of architectures for appearance feature extraction that includes state-of-the-art models and specific architectures offering unique advantages for video-based person re-identification. This decision balances computational efficiency with high performance in real-world scenarios. We focused on the last three architectures—OSNet [16], MGH [7], and PiT [10]—due to their unique capabilities in addressing specific challenges in appearance-based person re-identification. The OSNet architecture is lightweight yet powerful, designed for person Re-ID with an emphasis on omni-scale feature learning, which efficiently captures and integrates multi-scale feature representations. By incorporating OSNet, we aim to enhance the scalability and speed of GAF-Net, making it particularly suitable for scenarios where resources are limited but high accuracy is required. Following OSNet, we explored the MGH architecture, noted for its ability to manage complex relationships within data. Hypergraphs in MGH extend traditional graph-based methods by enabling higher-order connections among data points, which improves the handling of variations in pose, scale, and occlusion in videos. Lastly, the PiT model achieved top performance on the iLIDS-VID dataset [5], utilizes transformers to achieve rich contextual understanding and fine-grained feature extraction. This is crucial for distinguishing subtle differences in appearance and setting new benchmarks in the field. Our aim is to demonstrate the impact of incorporating gait information and using different fusion schemes to optimize these appearance-based architectures for improved re-identification performance.

3.2.1. OSNet Framework

The Omni-Scale Network (OSNet), introduced by Zhou et al. in [16], revolutionized person Re-ID. It excels at processing features at various spatial scales to adeptly handle Re-ID challenges like pose variations and occlusions.

Depthwise separable convolutions are central to OSNet, which efficiently reduces the number of parameters and computational cost. These convolutions are divided into depthwise and pointwise operations, known as Lite 3×3 convolutions in OSNet. This approach enhances the network's ability to learn features at multiple scales, which is essential for discerning individual identities.

Another core element of OSNet is its residual bottlenecks, each of which incorporates a Lite 3×3 layer. These bottlenecks are characterized by an exponent (s) that denotes the scale of feature learning. By stacking multiple such layers ($s \geq 1$), the network expands its receptive field, thereby capturing a more extensive range of spatial scales. This design enables a comprehensive analysis of features at varying scales. Furthermore, one of the most innovative aspects of OSNet is the unified aggregation gate (AG), which dynamically mixes features from multiple scales based on the input data. This gate assigns weights to feature maps at various scales to adapt the feature learning to each specific input image.

This dynamic fusion makes OSNet very efficient for person Re-ID tasks.

Then, each frame of the sequence feeds OSNet to obtain feature vectors that are then averaged to derive a unified feature representation for the entire sequence so that the final feature vector combines features from all frames, which is crucial for accurate video-based person Re-ID. This final feature vector ($\mathbf{z}^i_{\text{OSNet}}$) for an individual identified by ID i is expressed as follows:

$$\mathbf{z}^i_{\text{OSNet}} = \phi^{\text{OSNet}}(\mathbf{A}^i), \tag{4}$$

where $\mathbf{A}^i$ is the input tensor containing appearance data from different camera views, and ϕ^{OSNet} is a function representing the transformation of these raw data into a refined feature representation.

3.2.2. MGH Framework

The Multi-Granular Hypergraph (MGH) framework, introduced by Yan et al. in [7], represents an advanced solution for video-based person Re-ID. MGH addresses common issues like occlusions and misalignment in videos by leveraging spatial and temporal features at multiple levels of granularity.

MGH constructs hypergraphs for different granularity levels, each representing a pair consisting of a set of node features at a specific granularity level and a set of hyperedges, capturing temporal relationships (correlations) among these features. These hyperedges are obtained by identifying the nearest neighbors of each node in the graph within temporal ranges based on the affinity of temporal features, e.g., neighbors that share similar temporal characteristics. These hyperedges represent short-term correlations between temporally adjacent features, as well as mid- and long-range correlations across different temporal lengths.

MGH employs a HyperGraph Neural Network (HGNN) for feature learning and propagation. Within this network, information is gathered across hyperedges, and node features are updated iteratively.

Within this framework, an attentive hypergraph feature aggregation mechanism is used to enhance the discriminative power of features. It combines weighted node features from all hypergraphs to create the final video representation.

For the purpose of appearance feature extraction, the MGH framework aggregates complex multi-granular features into a unified representation denoted as $\mathbf{z}^i_{\text{MGH}}$ for a person identified by ID i in a video, expressed as follows:

$$\mathbf{z}^i_{\text{MGH}} = \phi^{\text{MGH}}(\mathbf{A}^i) \tag{5}$$

where $\mathbf{A}^i$ is the tensor that contains individual appearance data.

3.2.3. Pyramid in Transformer (PiT)

The Pyramid in Transformer (PiT) model, introduced by Zang et al. in [10], marks a significant advancement in video-based pedestrian retrieval. Relying on transformer models, PiT adeptly handles complex visual data hierarchies. Its multi-directional and multi-scale approach generates comprehensive representations of pedestrians, which are crucial for accurate person Re-ID under a variety of environmental conditions and for various camera perspectives.

PiT first segments each pedestrian image into patches, which are then converted into embeddings. These embeddings are processed through multiple transformer layers, employing strategies like no division, as well as vertical, horizontal, and patch-based divisions. This enables PiT to capture different ranges of features, enabling easier recognition of individuals across different poses and environmental settings.

For each image, PiT constructs a pyramid by combining features from various layers (or divisions, such as vertical, horizontal, etc.). The final representation is derived by averaging these pyramids across all images.

The combination of multi-directional and multi-scale features yields the output vector ($\phi^{\text{PiT}}(\mathbf{A}^i)$) such that

$$\mathbf{z}^i_{\text{PiT}} = \phi^{\text{PiT}}(\mathbf{A}^i), \tag{6}$$

were $\mathbf{A}^i$ is the tensor that contains the appearance data. The feature vector ($\mathbf{z}^i_{\text{PiT}}$) offers a multi-faceted view of the individual's appearance, enabling more accurate and efficient person re-identification, even in complex and dynamic visual environments.

3.3. Feature Fusion Module

In the field of person Re-ID, feature fusion is an important process for combining features obtained from different modalities into a coherent and comprehensive representation (cross-modality features), which significantly enhances the accuracy and reliability of a Re-ID system.

3.3.1. Fusion at the Feature Level

Feature-level fusion is essential in person Re-ID, merging distinct feature sets like gait and appearance into a unified representation. Studies such as that conducted by Lu et al. [36] have shown the effectiveness of feature vector concatenation. For a person identity (i) in our dataset, we obtain two feature vectors, namely $\mathbf{z}^i_{\text{appearance}}$ and $\mathbf{z}^i_{\text{gait}}$. These are combined using weighted concatenation, as described by the following equation:

$$\mathbf{z}^i = \phi_{\text{concat}}\left(\mathbf{z}^i_{\text{appearance}}, \mathbf{z}^i_{\text{gait}}\right) = \left[\lambda \cdot \mathbf{z}^i_{\text{appearance}}, (1-\lambda) \cdot \mathbf{z}^i_{\text{gait}}\right], \tag{7}$$

where λ is a scalar weighting factor within the range of [0,1] used to balance the contributions of the gait and appearance features, with λ specifically adjusting the influence of the gait features.

We also explore an alternative approach to feature-level fusion by first applying Principal Component Analysis (PCA) to enhance the accuracy and reliability of our person re-identification(Re-ID) system. Specifically, PCA is used to reduce the dimensionality of the appearance features to 128, aligning with the gait features' dimensionality. This approach helps preserve the unique characteristics of each modality while ensuring harmonized integration within the fused feature vector. As a result, we achieve a more comprehensive representation of the person's identity, which is critical for improving the performance of the Re-ID system. After dimensionality reduction, we concatenate the appearance and gait features to maintain the full spectrum of information, enhancing the classification process. We posit that using PCA for feature fusion offers a significant advantage in person Re-ID by facilitating more effective integration of multiple modalities. The details of the concatenated fusion are represented in the following equation:

$$\mathbf{z}^i = \phi_{\text{PCA-concat}}\left(\mathbf{z}^i_{\text{PCA-appearance}}, \mathbf{z}^i_{\text{gait}}\right) = \left[\lambda \cdot \mathbf{z}^i_{\text{PCA-appearance}}, (1-\lambda) \cdot \mathbf{z}^i_{\text{gait}}\right]. \tag{8}$$

Here, the PCA-modified appearance features are combined with the gait features using a weighted sum, where λ is a term used to balance the contributions after PCA reduction. This method ensures the harmonized integration of dimensionality-normalized appearance features with gait features.

3.3.2. Fusion at the Score Level

Score-level fusion integrates the output scores or ranks from different models, focusing on the distance matrix that measures the similarity between the query and gallery sets in re-identification tests. This approach is particularly efficient when each model captures unique aspects. It is important to note that in our scoring system, a lower score represents a better match. This means that the Re-ID system is more confident in the identity match when the score is minimized. Consequently, during the evaluation phase, our aim is to achieve the lowest mean score across all identities, indicating the highest overall system accuracy.

In the weighted averaging method, similarity scores or ranks from different models are averaged, with specific weights assigned to each score. The final score for a person (i), denoted as S^i_{final}, is calculated using the following equation:

$$S^i_{\text{final}} = \alpha \cdot S^i_{\text{appearance}} + (1-\alpha) \cdot S^i_{\text{gait}} \tag{9}$$

where S^i_{gait} represents the gait score and $S^i_{\text{appearance}}$ stands for the appearance score. The variable α serves as a weight factor that balances the significance of the appearance and gait scores in the final score.

Another approach is threshold-based conditional fusion. In this technique, gait and appearance are always utilized, but appearance features are only considered if the gait score is above the threshold. The score for a person (i), expressed as $S^i_{\text{conditional}}$, is determined by the following conditions:

$$S^i_{\text{conditional}} = \begin{cases} w \cdot S^i_{\text{appearance}} + (1-w) \cdot S^i_{\text{gait}} & \text{if } S^i_{\text{gait}} < \text{threshold} \\ S^i_{\text{appearance}} & \text{otherwise} \end{cases} \quad (10)$$

In this case, w is the weight factor that influences the contribution of the appearance score when the fusion is conditional on the gait score being above the threshold.

This paper aims to conduct a comparative analysis of these feature-level and score-level fusion methods to assess their distinct and collective impacts on re-identification (Re-ID) performance.

4. Experimental Settings

We present the experimental setup for evaluating GAF-Net. The following subsections detail the used datasets, model implementations, hardware and software specifications, and performance evaluation metrics.

4.1. Datasets

To rigorously assess GAF-Net's performance, we use two datasets, namely iLIDS-VID [5] and MARS [6]. Our primary focus is on iLIDS-VID due to its suitability for comprehensive gait analysis, while MARS serves to test our method's limitations in more complex, real-world scenarios.

The iLIDS-VID dataset [5] is our primary evaluation benchmark, comprising 600 video sequences of 300 distinct pedestrians captured by two non-overlapping cameras in an airport setting. This dataset is particularly valuable for our research due to its long sequence lengths, ranging from 23 to 192 frames, with an average of 73 frames per sequence. These long sequences are crucial for capturing complete gait cycles, enabling a thorough evaluation of our gait-based features. iLIDS-VID presents significant challenges for person re-identification (Re-ID), including complex background clutter, lighting variations, and instances of self-occlusions and mutual occlusions. These characteristics make iLIDS-VID an ideal testbed for evaluating GAF-Net's ability to integrate appearance and gait features effectively in realistic scenarios.

To further test GAF-Net's performance and find its potential weaknesses, we evaluate our method on the MARS (Motion Analysis and Re-identification Set) dataset [6]. MARS comprises over 20,000 video sequences collected from 1261 different individuals across 6 cameras installed on a university campus. Compared to iLIDS-VID, MARS is more diverse and complex. While some traklets contain between 25 and 50 frames, the majority of identities are represented by sequences with only 5 to 20 frames. This variation in sequence length adds to the dataset's complexity and poses a particular challenge for gait-based Re-ID methods, as many sequences may not contain a complete gait cycle. Additionally, MARS deliberately includes erroneous detection or tracking results as distractors. While these short tracklets are less ideal for comprehensive gait analysis, they reflect real challenges in practical video surveillance applications. The unique structure of MARS makes it a particularly demanding and representative benchmark for real-world re-identification scenarios, allowing us to test GAF-Net's performance under complex, real-world surveillance conditions and offering a rigorous evaluation framework for person re-identification algorithms.

When considering datasets for person re-identification research, it is important to note that the PRID2011 dataset [46] offers limited challenges, with state-of-the-art methods achieving over 95% rank-1 accuracy [9]. Additionally, the DukeMTMC-VideoReID

dataset [47] was retracted due to ethical concerns. These factors underscore the importance of selecting datasets that are both technically challenging and ethically sourced for the evaluation of Re-ID algorithms.

Our dual-dataset evaluation approach ensures a comprehensive evaluation of GAF-Net, leveraging iLIDS-VID's strengths for gait analysis while using MARS to assess robustness under challenging conditions. This strategy allows us to validate GAF-Net's capabilities across different real-world scenarios and provide insights into its potential and limitations in diverse Re-ID contexts.

4.2. Models and Implementation

Our Gait–Appearance Fusion Network (GAF-Net) integrates both gait and appearance features for video-based person Re-ID. This is crucial for efficient person Re-ID in multi-camera scenarios. GAF-Net consists of the following three modules: gait feature extraction and appearance feature extraction, followed by a fusion module that combines the extracted features.

4.2.1. Gait Feature Extraction

For gait extraction, we use the approach proposed by Teepe et al. in [13]. The GaitGraph model is fine-tuned on the CASIA-B dataset, focusing on sequences of 60 frames per tracklet. The training process incorporates a one-cycle learning-rate strategy paired with a weight decay penalty set to 10^{-5}. Initially, the learning rate peaks at 10^{-2} over a span of 1000 epochs, using a loss function temperature also set to 10^{-2}. For the training, the learning rate is reduced to 10^{-5}, and the model is further refined over 300 epochs, using a batch size of 128. This ensures that gait feature extraction is both robust and accurate.

4.2.2. Appearance Feature Extraction

We evaluate the following three backbone architectures for appearance feature extraction: OSNet [16], MGH [7], and PiT [10]—each selected for its distinctive strengths. For example, OSNet is known for its lightweight design, MGH utilizes hypergraphs to capture complex relationships, and PiT employs a transformer-based approach to leverage global context. Detailed specifics for each architecture are provided in Section 3.2.

- **OSNet Implementation.** OSNet is pretrained on ImageNet, with input frames resized to 256×128 pixels. Training extends over 60 epochs, with a batch size of 6 and using the Adam optimizer and cross-entropy loss. The learning rate is set at $10^{-3.5}$.
- **MGH Implementation.** In the MGH architecture, images are resized to 256×128 pixels. Training batches consist of 8 random sub-sequences from 8 individuals. We use a hypergraph with $L = 2$ layers and $K = 3$ neighbors; spatial partitions of 1, 2, 4, and 8; and temporal thresholds set to 1, 3, and 5. The Adam optimizer uses a weight decay of 5×10^{-4}, and the learning rate, starting from $10^{-3.5}$, reduces tenfold every 100 epochs up to 300 epochs. Graph-level features are concatenated after training, using cosine similarity as the matching metric [48].
- **PiT Implementation.** The PiT model, based on the pre-trained ViT-B16 transformer [22], processes images resized to 256×128 pixels. The batch size and m parameter are set to 16 and 11, respectively. The convolution layer, with a kernel size of 16 and stride of 12, results in feature embeddings with dimensions of $21 \times 10 \times 768$. Trade-off parameters λ_1 and λ_2 are fine-tuned to 1.0 and 1.5, respectively. Division parameters are set to $D_v = 2$, $D_h = 3$, and $D_p = 6$ (see [10] for more details). Training uses Stochastic Gradient Descent (SGD) with momentum, starting from a learning rate of 10^{-2} for 120 epochs, which is adjusted using cosine annealing. The convolution and transformer layers are initially frozen for the first five epochs to optimize the refinement of the classifier.

4.3. Hardware and Software

Our experiments are conducted on a QUADRO RTX 4000 with PyTorch, chosen for its flexibility in neural network manipulation. We ensured that all software and libraries are up-to-date, guaranteeing the reproducibility of our results.

4.4. Evaluation Metrics

We employ dataset-specific protocols and metrics to ensure rigorous and comparable evaluations of GAF-Net.

For iLIDS-VID [5], we use the ten-split protocol, where each split randomly divides the dataset into two groups of 150 individuals for training and testing. We report the average performance across all splits.

We primarily use Cumulative Matching Characteristic (CMC) curves, reporting results from rank 1 (R-1) to rank 20 (R-20), with an emphasis on accuracies at R-1, R-5, and R-20.

For the MARS dataset [6], we follow the standard train/test split. In addition to CMC, we incorporate the mean Average Precision (mAP) metric. The mAP is particularly relevant for multi-camera setup of MARS, providing a holistic measure of retrieval quality.

This combination of metrics offers a comprehensive view of our model's performance across different datasets and scenarios, facilitating fair comparisons with other state-of-the-art methodologies in person Re-identification.

5. Experimental Results

Our comprehensive evaluation of GAF-Net covers key aspects of person re-Identification using two distinct datasets, namely iLIDS-VID and MARS. This study focuses on the integration of skeleton gait data into appearance features, exploring both feature-level and score-level fusion techniques. We evaluate our approach on these datasets to demonstrate its effectiveness in different scenarios, comparing it with existing state-of-the-art methods. Additionally, we investigate the influence of fusion factors λ and α on various deep learning models, providing insights into optimal settings for the enhancement of Re-ID accuracy.

We begin by presenting our results obtained on the iLIDS-VID dataset, followed by an analysis of performance on the more complex MARS dataset. This dual-dataset approach allows us to assess GAF-Net's robustness and adaptability across varying conditions.

5.1. Results on iLIDS-VID Dataset

The iLIDS-VID dataset, with its long video sequences and controlled environment, provides an ideal test to evaluate GAF-Net's ability to integrate skeleton-based gait features with appearance information. Our analysis of this dataset encompasses the performance of different fusion techniques, comparisons with state-of-the-art methods, and an examination of optimal fusion factor settings. These results serve as a basis for understanding GAF-Net's effectiveness in scenarios where gait information can be reliably extracted and utilized.

5.1.1. Feature-Level Fusion

In this experiment, we explored the impact of feature-level fusion techniques on the performance of various backbone architectures. We selected three distinct architectures, each with unique characteristics, to generalize our findings effectively. This involved two primary approaches, which are tested in the following sub-sections.

To this end, we employed two primary fusion approaches, namely direct concatenation and PCA-concatenation. These were chosen for their potential to integrate complementary information effectively, thereby improving the robustness and accuracy of the identification process.

Concatenation:

Table 1 illustrates the enhancements achieved by concatenating gait features, with a dimensionality of 128, into various appearance feature extraction architectures. Each model was optimized with a specific concatenation factor (λ), as detailed in Equation (7).

Feeding the tested models the 128-dimensional gait feature vector output by the GaitGraph architecture resulted in the following noticeable improvements for all models:

- OSNet: Initially achieving a rank-1 accuracy of 59.20% with its compact 512-dimensional architecture, the accuracy of OSNet increased by 11.73% following the integration of this 128-dimensional gait feature vector.
- MGH: For the MGH model, which produces 5120-dimensional vectors, the integration of the 128-dimensional gait features increased the rank-1 accuracy by 4.7%.
- PiT: The PiT model, known for its extensive 9216-dimensional output vector, showed an initial rank-1 accuracy of 92.07%. This accuracy was increased by 1% following the integration of the gait feature vector.

These results demonstrate the value of integrating gait features, especially the 128-dimensional representation output by the GaitGraph architecture, in enhancing the performance of various appearance feature backbones in person re-identification (Re-ID) tasks. This integration not only improves upon existing methods but also pushes the state of the art.

Table 1. Person Re-ID performances (rank-1/rank-5/rank-10/rank-20) on the iLIDS-VID dataset depending according to the model with or without concatenated gait information.

Backbone	Appearance				Appearance + Gait (Concatenation)			
	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20
OSNet [16]	59.20	82.60	89.34	94.80	70.93 (+11.73)	89.47 (+6.87)	93.80 (+4.46)	97.27 (+2.47)
MGH [7]	85.60	97.10	99.00	99.30	90.40 (+4.80)	98.67 (+1.57)	98.94 (−0.06)	99.67 (+0.37)
PiT [10]	92.07	98.93	99.80	100	93.07 (+1.00)	99.34 (+0.41)	99.94 (+0.14)	100.00 (+0.00)

PCA Concatenation:

Our experimental results, as presented in Table 2, clearly demonstrate the enhancement achieved by applying PCA with concatenation as a fusion technique in person Re-ID. This approach was employed after normalizing the dimensionality of appearance features to match the 128 gait features using principal component analysis (PCA). PCA played a crucial role in transforming the appearance features derived from various backbone architectures into a uniform 128-dimensional format, which is essential for ensuring a balanced and efficient fusion process. For each studied model, we found the optimal value for the fusion process. The results obtained with the different tested models are described below.

- OSNet: The OSNet architecture, with an original rank-1 accuracy of 59.20%, achieved a score of 74.00% when fused using PCA with concatenation, corresponding to an important increase of 14% attributed to the effective integration of the normalized gait and appearance features. This significant improvement demonstrates the potential of this real-time method.
- MGH: The MGH model, initially achieving a rank-1 accuracy of 85.60%, reached an accuracy score of 90.07% post fusion. This enhancement underscores the benefit of applying PCA for dimensionality normalization before the fusion process.
- PiT: PiT, starting with a high rank-1 accuracy of 92.07%, showed a slight increase to reach 92.87%. This demonstrates the impact of feature fusion on already high-performing architectures.

These results highlight the positive impact of balancing dimensionality with concatenation in integrating gait and appearance features for person re-identification.

Table 2. Person Re-ID performances (rank-1/rank-5/rank-10/rank-20) in the iLIDS-VID dataset according to the model with or without the integration of PCA concatenation.

Backbone	Appearance				Appearance + Gait (PCA-Concatenation)			
	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20
OSNet [16]	59.20	82.60	89.34	94.80	74.00 (+14.20)	91.94 (+9.34)	95.67 (+6.33)	98.47 (+3.67)
MGH [7]	85.60	97.10	99.00	99.30	90.07 (+4.47)	98.54 (+1.44)	99.34 (+0.34)	99.80 (+0.50)
PiT [10]	92.07	98.93	99.80	100	92.87 (+0.80)	99.40 (+0.47)	99.94 (+0.14)	100 (+0.00)

5.1.2. Score-Level Fusion

In addition to feature-level fusion, we also explored score-level fusion techniques, which focus on combining the scores or ranks obtained by different models. Each model excels in capturing unique aspects of the data for Person Re-ID. We investigated two main score-level fusion strategies, namely weighted averaging and threshold-based conditional fusion.

Weighted Averaging

The effectiveness of weighted averaging as a score-level fusion strategy is underscored by the notable performance improvements in person Re-ID, as detailed in Table 3. This approach strategically combines the scores from different models, assigning weights that reflect their relative confidence, leading to a cumulative enhancement of the overall identification accuracy.

- OSNet: The integration of gait data via weighted averaging boosted OSNet's performance by 11.00%, reaching a rank-1 accuracy score of 70.66%.
- MGH: The performance of the MGH model, which achieved a rank-1 accuracy of 85.60% using only appearance features, reached 90.20% using weighted averaging.
- PiT: For the PiT model, known for its high initial rank-1 accuracy of 92.07%, the application of score-level fusion with weighted averaging further boosted its performance, achieving a rank-1 accuracy of 93.20%. Although PiT achieves high scores by using only appearance features, we can see that the addition of gait features can also increase its performance.

These results show the efficiency of weighted averaging in score-level fusion for person Re-ID. This approach, which integrates both appearance and gait features, increased the accuracy scores of all tested models.

Table 3. Person Re-ID performances (rank-1/rank-5/rank-10/rank-20) on the iLIDS-VID dataset according to the model with or without the integration of score-level fusion with weighted averaging.

Backbone	Appearance				Appearance + Gait (Weighted Averaging)			
	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20
OSNet [16]	59.20	82.60	89.34	94.80	70.66 (+11.46)	89.93 (+7.33)	94.40 (+5.06)	97.53 (+2.73)
MGH [7]	85.60	97.10	99.00	99.30	90.20 (+4.60)	98.53 (+1.43)	99.00 (+0.00)	99.60 (+0.30)
PiT [10]	92.07	98.93	99.80	100	93.20 (+1.13)	99.26 (+0.33)	99.86 (+0.06)	100.00 (+0.00)

Threshold-Based Conditional Fusion

The impact of implementing threshold-based conditional fusion on person Re-ID is documented in Table 4. This advanced approach combines gait information with appearance features in a dual-case strategy; it either fuses the two data types or selects the superior model based on appearance according to a predetermined threshold. This methodology consistently enhanced the accuracy of person Re-ID across all tested architectures.

- OSNet: The OSNet model, originally achieving a rank-1 accuracy of 59.20%, reached a score of 70.26% using threshold-based conditional fusion.
- MGH: The MGH model, which recorded an initial rank-1 accuracy of 85.60%, saw its performance boosted to 89.66%. This demonstrates the substantial impact of integrating gait data to improve the accuracy of high-dimensional feature vectors.
- PiT: For the PiT model, known for its extensive output vector, the initial rank-1 accuracy of 92.07% increased to 93% after fusion. This increase, albeit modest, underscores the value of fusion even in models with high baseline performance.

These results show the benefits of threshold-based conditional fusion in person Re-ID tasks. The integration of gait features with appearance data proved to be a key factor in improving the performance of diverse backbone architectures, thereby enhancing their effectiveness in complex re-identification scenarios.

Table 4. Person Re-ID performances (rank-1/rank-5/rank-10/rank-20) on the iLIDS-VID dataset according to the model and the integration of gait information.

Backbone	Appearance				Appearance + Gait (Threshold Conditional Fusion)			
	Rank-1	Rank-5	Rank-10	Rank-20	Rank-1	Rank-5	Rank-10	Rank-20
OSNet [16]	59.20	82.60	89.34	94.80	70.26 (+11.06)	89.73 (+7.13)	93.86 (+4.52)	97.66 (+2.86)
MGH [7]	85.60	97.10	99.00	99.30	89.66 (+4.06)	98.40 (+1.30)	99.06 (+0.06)	99.40 (+0.10)
PiT [10]	92.07	98.93	99.80	100	93.00 (+2.67)	99.00 (+0.00)	99.60 (−0.36)	99.73 (−0.27)

5.1.3. Comparison with the SOTA Methods

In this section, we benchmark the performance of our improved GAF-Net, which utilizes the PiT architecture as the backbone for the appearance module and incorporates gait features, with other SOTA methods in the domain of person Re-ID. Our study emphasizes the significant improvements achieved by both fusion-level (concatenation) and score-level (weighted averaging) approaches, as discussed in the previous section.

The evaluation was performed on the iLIDS-VID dataset, where our GAF-Net achieved the best performance in terms of accuracy, as shown in Table 5. Several conclusions can be drawn from our study, which are discussed below.

Table 5. Comparison with state-of-the-art backbones on iLIDS-VID.

Backbone	Results			
	Rank-1	Rank-5	Rank-10	Rank-20
PiT [10]	92.07	98.93	99.80	**00**
DenseIL [8]	92.00	98.00	-	-
DCCT [11]	91.70	98.60	-	-
PSTA [9]	91.50	98.10	-	-
AA-RGTCN [49]	90.60	-	-	-
STRF [50]	89.30	-	-	-
MGH [7]	85.6	97.10	-	99.50
GAF-Net (ours—concatenation)	93.07	**99.34**	**99.94**	**100**
GAF-Net (ours—score level)	**93.20**	99.27	99.86	**100**

Note: values indicate the best performance for each metric.

By using PiT as the backbone, our GAF-Net increased rank-1 accuracy by 1.13% and rank-5 accuracy by 0.34% over the baseline PiT model with weighted sum fusion. Additionally, with concatenation, it improved by 1.00% and 0.41% for rank-1 and rank-5, respectively. These increases are significant, given the already high accuracy of the baseline model.

Furthermore, the DenseIL method [8], which utilizes a hybrid framework of CNN-based and attention-based architectures, and the DCCT approach [11], which uses a coupled

CNN with transformers, both demonstrate lower performance compared to GAF-Net. Our method surpassed DCCT by 1.5% and showed a slight advantage over DenseIL of 0.20% in rank-1 accuracy, illustrating the effectiveness of our dual fusion strategy.

In comparison to Pyramid Spatial–Temporal Aggregation (PSTA) [9], which combines frame-level features with hierarchical temporal elements for detailed video-level representation, our GAF-Net achieved higher scores, outperforming PSTA by 1.7% in rank-1 while maintaining similar rank-5 performance.

Furthermore, AA-RGTCN [49], which employs an adaptive alignment that effectively addresses frame misalignment and robustly captures temporal features, showed a rank-1 accuracy 2.6% lower than that of GAF-Net. This highlights GAF-Net's superior ability to effectively synthesize appearance and motion cues.

Compared to the STRF approach [50], which integrates spatiotemporal representations through the use of 3D convolutional neural networks, our GAF-Net provided a significant performance boost, increasing rank-1 accuracy by 3.9%.

Furthermore, compared to the multi-granular hypergraph (MGH) method [7], our GAF-Net achieved a substantial increase in accuracy, increasing rank-1 accuracy by 7.6% and rank-5 accuracy by 2.17%.

These comparative evaluations demonstrate the robustness and superiority of GAF-Net in harnessing gait and appearance features for the purpose of person Re-ID. The results confirm that our approach not only competes well with existing methods but also sets a new bar for accuracy in this challenging domain. Additionally, GAF-Net improves lightweight architectures significantly by integrating skeletal gait information.

5.1.4. Analysis of the Concatenation Fusion Factor (λ)

In this section, we analyze the fusion factor (λ) of Equation (7), which determines the fusion of appearance and gait information through concatenation. Figure 2 demonstrates the impact of varying λ values on rank-1 accuracy for the three tested architectures, namely OSNet, PiT, and MGH.

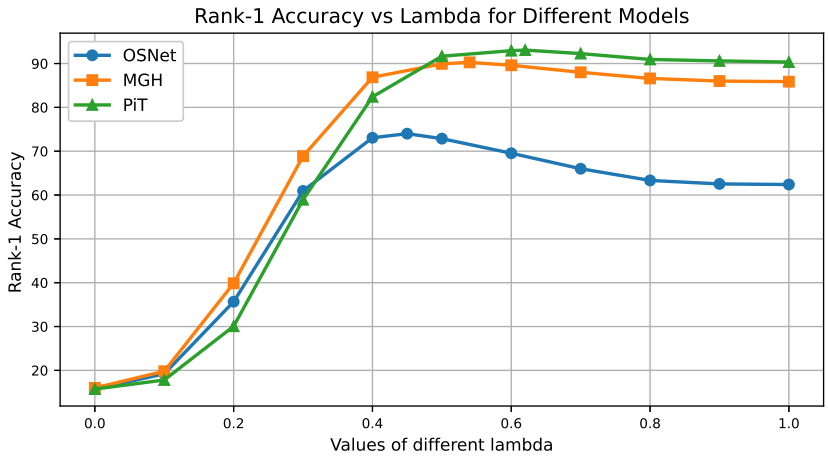

Figure 2. Impact of the fusion factor (λ; varying from 0 to 1) on the rank-1 accuracy.

For the OSNet architecture, we can see its rank-1 accuracy increases with λ until it reaches $\lambda = 0.45$, with a rank-1 accuracy of 70.93%. This accuracy remains relatively stable, then decreases with higher values of λ.

For the MGH architecture, the rank-1 accuracy also increases with λ—but more slowly—until $\lambda = 0.54$, reaching an accuracy of 90%, then remaining stable with higher values of λ.

The PiT model provides high rank-1 accuracy without adding any gait information ($\lambda = 0$). This accuracy slowly increases with increasing λ values, reaching 93% at $\lambda = 0.6$, beyond which it decreases.

In summary, the accuracy of all three architectures is increased by up to 10% by adding gait information. Our tests show that the best values of λ are in the $[0.4, 0.6]$ interval.

5.1.5. Evaluation of the Weighted Averaging Fusion Factor (α)

The experimental results reported in Figure 3 show the contribution of the fusion factor value (α) to the rank-1 accuracy in score-level fusion in the OSNet, MGH, and PiT architectures, as described by Equation (9).

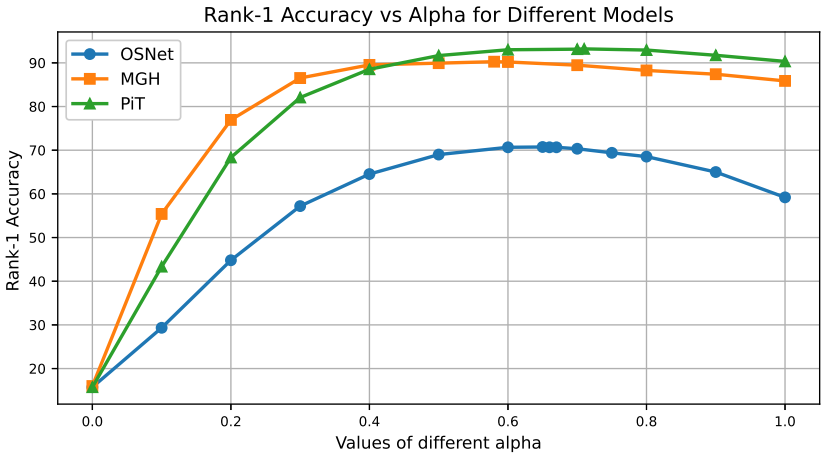

Figure 3. Impact of the fusion factor value (α; varying from 0 to 1) on the rank-1 accuracy.

For OSNet, the rank-1 accuracy first increases slightly with α until reaching a rank-1 accuracy of approximately 70% for $\alpha = 0.66$. For the highest values of α, the accuracy decreases.

For MGH, the rank-1 accuracy (starting from 86%) also increases with the value of α and reaches an accuracy of 90% for $\alpha = 0.6$, then decreases.

The PiT model starts with a high rank-1 accuracy (90%), showing high performance without the integration of additional information. We can see that this accuracy increases slightly with α until reaching 93% for $\alpha = 0.7$. For the highest values of α, the accuracy strongly decreases, suggesting potential overfitting or information saturation.

These results suggest that adding additional information through the use of a fusion factor can increase the performance of the models. The choice of α is, thus, important in increasing the predictive power of these models. For optimal results, it seems we can choose a fusion factor of $\alpha \in [0.6, 0.75]$ for all tested architectures.

5.2. Results on MARS Dataset

Following our analysis of iLIDS-VID, we extend our evaluation to the MARS dataset [6], one of the largest and most complex video-based person re-identification databases. MARS presents unique challenges, with its brief video sequences, numerous distractors, and significant variations in capture conditions. These characteristics rigorously test GAF-Net, pushing its limits in scenarios that closely resemble real-world conditions. Unlike iLIDS-VID, MARS revealed significant limitations in gait information extraction. Our analysis shows that while appearance features were obtainable for all tracklets, gait features were extractable for only 85.6% of test-set tracklets, with a notable disparity between the gallery (82.7%) and query (99%) of the tracklets. This disparity highlights the challenges inherent in real-world scenarios, such as occlusions, suboptimal camera angles, and brief video sequences. Considering these constraints and the significantly larger scale of MARS compared to iLIDS-VID, we focus our evaluation on two backbone architectures, namely OSNet [16] and MGH [7]. These architectures offer a balance between performance and efficiency more suitable for the extensive MARS dataset. We exclude PiT from this evaluation due to its higher computational demands, which are less practical for the scale of MARS. Both feature-level and score-level fusion approaches are explored, with particular attention to

conditional concatenation, which is especially relevant, given the inconsistent availability of skeleton-based gait information in this dataset.

This study demonstrates GAF-Net's adaptability and effectiveness in scenarios where gait feature extraction is not always feasible, providing insights into the robustness of our approach in complex, real-world Re-ID tasks, specifically through the lens of the OSNet and MGH architectures.

5.2.1. Feature-Level Fusion

To tackle MARS's unique challenges, we propose two adaptive fusion methods, namely Conditional concatenation and PCA conditional concatenation. These methods aim to optimize gait information use, maintain performance when gait data are unreliable or missing, and balance appearance and gait features. Conditional concatenation dynamically combines or separates features, while PCA conditional concatenation incorporates dimensionality reduction for enhanced efficiency. By implementing these methods with the OSNet and MGH architectures, we assess their effectiveness in improving Re-ID performance on MARS. This approach allows us to evaluate the impact of gait information integration and dimensionality reduction, demonstrating GAF-Net's adaptability in scenarios where gait feature extraction is challenging.

Conditional Concatenation

Based on the concatenation defined previously by Equation (7), we introduce a conditional concatenation for MARS to handle cases where gait features are unavailable. For a person's identity (i), conditional concatenation ($\mathbf{z}_c^i$) is defined as follows:

$$\mathbf{z}_c^i = \begin{cases} \phi_{\text{concat}}\left(\mathbf{z}^i_{\text{appearance}}, \mathbf{z}^i_{\text{gait}}\right) & \text{if gait available} \\ \left[\mathbf{z}^i_{\text{appearance}}, \mathbf{0}_{d_{\text{gait}}}\right] & \text{otherwise} \end{cases} \tag{11}$$

where ϕ_{concat} is the weighted concatenation function defined in Equation (7), including the weighting factor (λ), and d_{gait} is the dimension of the gait feature vector. In our implementation, $d_{\text{gait}} = 128$.

This approach allows for flexible integration of both types of features when available, using the λ factor to adjust their relative importance. When gait is not observable, the system can still function by relying only on appearance features, although this may reduce identification accuracy compared to using both types of features. Table 6 presents the results of applying conditional concatenation to the MARS dataset for the two architectures. This approach was used to handle cases where gait information was not available. The results show modest but consistent improvements for the tested models.

- OSNet: The OSNet architecture initially achieved an mAP of 81.64% and a rank-1 accuracy of 86.20%. After conditional integration of gait features, we observe a slight improvement in mAP to 81.67% (+0.03%), while rank-1 accuracy showed minimal variation, reaching 86.14% (−0.06%). These results suggest that the impact of integrating gait features is marginal for this model on the MARS dataset.
- MGH: The MGH model demonstrated more noticeable improvements with conditional concatenation. Its initial mAP of 85.52% increased to 85.56% (+0.04%), while its rank-1 accuracy improved from 89.40% to 89.51% (+0.11%). Although modest, these improvements are consistent across metrics, indicating a potential benefit of integrating gait features for this model.

Conditional concatenation of gait features enhances Re-ID performance on the MARS dataset, with the MGH model showing more notable improvements. Although these gains are smaller than those observed on iLIDS-VID, likely due to dataset differences and the use of the conditional fusion approach, our method consistently improves performance across architectures. This shows its effectiveness even when gait information is intermittently available, adapting to the challenges presented by the MARS dataset.

Table 6. Results of feature-level fusion on the MARS dataset (concatenation).

Backbone	Appearance					Appearance + Gait (Conditional Concatenation)				
	mAP	**Rank-1**	**Rank-5**	**Rank-10**	**Rank-20**	**mAP**	**Rank-1**	**Rank-5**	**Rank-10**	**Rank-20**
OSNet [16]	81.64	86.20	95.71	96.96	97.99	81.67	86.14	95.65	96.96	98.04
MGH [7]	85.52	89.40	96.63	97.72	98.42	85.56	89.51	96.68	97.72	98.42

PCA Conditional Concatenation

Building upon the PCA concatenation introduced in Equation (8), we propose a conditional version to handle cases where gait features are not available. This approach retains the advantages of PCA in terms of appearance features while adding flexibility to the model.

For a person's identity (i), PCA conditional concatenation ($\mathbf{z}^i_{ac}$) is defined as follows:

$$\mathbf{z}^i_{ac} = \begin{cases} \phi_{\text{PCA-concat}}\left(\mathbf{z}^i_{\text{PCA-appearance}}, \mathbf{z}^i_{\text{gait}}\right) & \text{if gait available} \\ \left[\mathbf{z}^i_{\text{PCA-appearance}}, \mathbf{0}_{d_{\text{gait}}}\right] & \text{otherwise} \end{cases} \tag{12}$$

where $\phi_{\text{PCA-concat}}$ is the weighted concatenation function defined in Equation (8), including the λ factor.

$$\phi_{\text{PCA-concat}}\left(\mathbf{z}^i_{\text{PCA-appearance}}, \mathbf{z}^i_{\text{gait}}\right) = \left[\lambda \cdot \mathbf{z}^i_{\text{PCA-appearance}}, (1-\lambda) \cdot \mathbf{z}^i_{\text{gait}}\right] \tag{13}$$

This conditional approach maintains system performance even when gait features are not observable while preserving λ and balancing when they are available. In cases where gait is not observable, the gait feature vector is replaced by a zero vector of the same dimension (d_{gait}, where d_{gait} is the dimension of the gait feature vector). In our implementation, $d_{\text{gait}} = 128$. This allows the system to function solely with PCA-optimized appearance features. This ensures model robustness in the face of missing data while preserving the advantages of dimensionality reduction, alignment of features obtained by PCA, and flexible weighting offered by λ. Table 7 presents the results of applying PCA conditional concatenation to the MARS dataset for different appearance feature extraction architectures. This approach demonstrates more significant improvements than simple concatenation while effectively handling cases where gait information is not available.

- OSNet: The OSNet architecture initially achieved an mAP of 81.64% and a rank-1 accuracy of 86.20%. After applying PCA conditional concatenation, we observed a considerable improvement in mAP to 81.89% (+0.25%) and an improvement in rank-1 accuracy to 86.52% (+0.32%). These results demonstrate the effectiveness of this method in improving OSNet's performance on the MARS dataset.
- MGH: The MGH model showed even more marked improvements. Its initial mAP of 85.52% increased to 86.09% (+0.57%), while its rank-1 accuracy improved from 89.40% to 89.78% (+0.38%). These larger gains suggest that the MGH model particularly benefits from the integration of gait features via PCA conditional concatenation.

Comparing the two fusion methods, PCA conditional concatenation outperforms conditional concatenation for both architectures. For OSNet, PCA conditional concatenation improved mAP by 0.25%, versus 0.03% with conditional concatenation. For MGH, the improvement was 0.57% versus 0.04% in mAP. These results suggest that PCA's dimensionality reduction enables more effective integration of appearance and gait features, especially benefiting complex architectures like MGH. The performance gains achieved through the use of PCA conditional concatenation, coupled with empirically determined λ values for each method and architecture, highlight its potential to enhance Re-ID systems. This approach shows promise for real-world scenarios involving large-scale datasets and varying availability of gait information.

Table 7. Results of feature-level fusion on the MARS dataset(PCA conditional concatenation).

Backbone	Appearance					Appearance + Gait (PCA-Conditional Concatenation)				
	mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20
OSNet [16]	81.64	86.20	95.71	96.96	97.99	81.89	86.52	95.87	97.01	98.10
MGH [7]	85.52	89.40	96.63	97.72	98.42	86.09	89.78	96.90	98.04	98.59

5.2.2. Score-Level Fusion

We now explore score-level fusion methods, following our analysis of feature-level fusion techniques and their impact on GAF-Net's performance on the MARS dataset. These approaches, operating later in the Re-ID pipeline, offer enhanced flexibility in handling variable gait information quality and availability. Score-level fusion enables adaptive decision making, robust handling of missing data, and potential performance optimization beyond feature-level techniques. We investigate the following two tailored methods: conditional weighted averaging and threshold-based conditional fusion. These approaches leverage both appearance and gait modalities while addressing the challenges of the MARS dataset, particularly when gait features are unreliable or absent.

Conditional Weighted Averaging

Building upon the weighted averaging method introduced previously (Equation (9)), we propose a conditional version that accounts for gait availability. For a person's identity (i), the final score (S^i_{final}) is defined as follows:

$$S^i_{\text{final}} = \begin{cases} \alpha \cdot S^i_{\text{gait}} + (1-\alpha) \cdot S^i_{\text{appearance}} & \text{if gait available} \\ S^i_{\text{appearance}} & \text{otherwise} \end{cases} \quad (14)$$

This approach extends the previous weighted averaging formula with a condition for cases lacking gait features. It increases flexibility, allowing for dynamic adaptation to data availability while maintaining effective score fusion when all information is present

The efficacy of this method is demonstrated by performance improvements in person re-identification on the MARS dataset, as detailed in Table 8.

- OSNet: Gait data integration via conditional weighted averaging marginally enhanced OSNet's performance. mAP increased by 0.08% (81.64% to 81.72%). Rank-1 accuracy improved by 0.21% (86.20% to 86.41%). Rank-5 showed a slight decrease (−0.17%), while rank-10 and rank-20 improved by 0.11% each.
- MGH: The conditional fusion slightly improved MGH's performance. mAP increased by 0.08% (85.52% to 85.60%). Rank-1 accuracy improved by 0.27% (89.40% to 89.67%). Rank-5 showed a minor increase (+0.05%), while rank-10 (97.72%) and rank-20 (98.42%) remained unchanged.

These improvements demonstrate the positive contribution of gait information to the MARS dataset. This approach offers an adaptive solution capable of leveraging gait information when available while maintaining performance based solely on appearance in other cases.

Table 8. Results of score-level fusion on the MARS dataset(Conditional Weighted Averaging).

Backbone	Appearance					Appearance + Gait (Conditional Weighted Averaging)				
	mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20
OSNet [16]	81.64	86.20	95.71	96.96	97.99	81.72	86.41	95.54	97.07	98.10
MGH [7]	85.52	89.40	96.63	97.72	98.42	85.60	89.67	96.68	97.72	98.42

Threshold-Based Conditional Fusion

We propose an extension of threshold-based conditional fusion (Equation (10)). This improved method explicitly incorporates a criterion for gait data availability, offering increased adaptability. The score ($S^i_{\text{conditional}}$) for a person (i) is determined as follows:

$$S^i_{\text{conditional}} = \begin{cases} w \cdot S^i_{\text{appearance}} + (1-w) \cdot S^i_{\text{gait}} & \text{if gait available and } S^i_{\text{gait}} < \text{threshold} \\ S^i_{\text{appearance}} & \text{otherwise} \end{cases} \quad (15)$$

where $w \in [0,1]$ is a weight factor for the appearance score, and "threshold" determines the reliability of the gait score.

This formulation extends our previous approach by considering both gait data availability and quality. It enhances robustness for sequences with absent or unreliable gait information while allowing for fine tuning of each modality's contribution. The method adapts dynamically to each sequence's characteristics, optimizing the use of available information in the MARS dataset.

Table 9 presents the results of threshold-based conditional fusion on the MARS dataset, with the best performance achieved after optimizing w and threshold values. This approach adaptively combines gait and appearance data or relies solely on appearance when necessary. Our methodology yields modest but consistent improvements in re-identification accuracy across the tested architectures.

- OSNet: The OSNet model, with an initial mAP of 81.64% and rank-1 accuracy of 86.20%, showed a slight improvement when using threshold-based conditional fusion. The mAP increased to 81.71% (+0.07%). While the rank-1 accuracy remained unchanged at 86.20%, the rank-5 accuracy improved from 95.71% to 95.76% (+0.05%).
- MGH: The MGH model, initially recording a mAP of 85.52% and a rank-1 accuracy of 89.40%, saw its performance increase to a mAP of 85.57% (+0.05%) and a rank-1 accuracy of 89.46% (+0.06%) through the fusion process. A slight improvement was also observed at rank-5, increasing from 96.63% to 96.68% (+0.05%). These results indicate that gait information can contribute to performance enhancement, even marginally, under the complex conditions of the MARS dataset.

These results demonstrate the potential of threshold-based conditional fusion in person re-identification tasks on the MARS dataset. The integration of gait characteristics with appearance data contributes to the improvement of the performance of the tested architectures, adapting well to the dataset's specific challenges.

Table 9. Results of score-level fusion on the MARS dataset(Threshold-Based Conditional Fusion).

Backbone	Appearance					Appearance + Gait (Threshold-Based Conditional Fusion)				
	mAP	Rank-1	Rank-5	Rank-10	Rank-20	mAP	Rank-1	Rank-5	Rank-10	Rank-20
OSNet [16]	81.64	86.20	95.71	96.96	97.99	81.71	86.20	95.76	97.07	97.99
MGH [7]	85.52	89.40	96.63	97.72	98.42	85.57	89.46	96.68	97.72	98.42

5.2.3. Comparison with State-of-the-Art Methods

This section evaluates GAF-Net's performance against current state-of-the-art methods on the MARS dataset. MARS presents significant challenges for person re-identification systems due to its scale, complexity, and variability. Our approach employs the MGH architecture [7] as the appearance module backbone, incorporating gait characteristics through the use of PCA conditional concatenation. Table 10 provides a comparison of GAF-Net with recent Re-ID methodologies.

Table 10. Comparison with state-of-the-art methods on the MARS dataset [6].

Backbone	Results				
	mAP (%)	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)	Rank-20 (%)
OSNet [16]	81.64	86.20	95.71	96.96	97.99
MGH [7]	85.52	89.40	96.63	97.72	98.42
PSTA [9]	85.80	91.50	-	-	-
AA-RGTCN [49]	85.90	91.00	-	-	-
STRF [50]	86.10	90.30	-	-	-
PiT [10]	86.80	90.22	**97.23**	**98.04**	-
DenseIL [8]	87.00	90.80	97.10	-	**98.80**
DCCT [11]	**87.50**	**92.30**	-	-	-
GAF-Net (MGH-PCA concatenation)	86.09	89.78	96.90	**98.04**	98.59

Note: Bold values indicate the best performance for each metric.

Several important observations can be drawn from these results. Our GAF-Net, using MGH as the backbone with PCA conditional concatenation, achieved competitive performance with, an mAP of 86.09% and a rank-1 accuracy of 89.78%. These results, while solid, are slightly behind those of some of the most recent and best-performing methods. Among the recent top-performing approaches, DenseIL [8] achieves an mAP of 87% and a rank-1 accuracy of 90.8%, outperforming our method by 0.91% in mAP and 1.02% in rank-1 accuracy. Similarly, PiT [10] obtains comparable results, with a mAP of 86.8% and a rank-1 accuracy of 90.22%. DCCT [11], another recent method, shows even higher performance, with a mAP of 87.5% and a rank-1 accuracy of 92.3%, surpassing our GAF-Net by 1.41% in mAP and 2.52% in rank-1 accuracy. These results indicate that approaches based on transformers and hybrid CNN–transformer architectures can offer significant advantages for person re-identification. Nevertheless, our GAF-Net compares favorably to other recent methods. For example, our method slightly outperform STRF [50] in terms of mAP (86.09% versus 86.10%) while remaining close in rank-1 accuracy (89.78% versus 90.30%). Similarly, we obtained comparable results to those of PSTA [9] and AA-RGTCN [49], which achieve mAPs of 85.8% and 85.9%, respectively.

The results indicate that GAF-Net maintains robust performance for higher ranks, with a rank-5 accuracy of 96.90%, rank-10 accuracy of 98.04%, and rank-20 accuracy of 98.59%. These results are competitive or superior to those reported for other methods, demonstrating the robustness of our approach.

In conclusion, although our GAF-Net does not achieve the best performance among the most recent methods on the MARS dataset, it remains competitive, demonstrating the effectiveness of our approach fusing appearance and gait characteristics. These results suggest avenues for future improvement, potentially by incorporating elements from the best-performing methods, such as PiT [10] or DenseIL [8], into GAF-Net.

5.2.4. Comparative Analysis and Discussion

Our evaluation of GAF-Net on the iLIDS-VID and MARS datasets provides comprehensive insights into its performance and adaptability in diverse person re-identification (Re-ID) scenarios.

Performance on iLIDS-VID

On the iLIDS-VID dataset, GAF-Net, building upon the PiT architecture for appearance features, achieved state-of-the-art performance, with a rank-1 accuracy of 93.20% and a rank-5 accuracy of 99.34%, surpassing contemporary methods such as DenseIL [8], DCCT [11], and PSTA [9] by margins of 1.20%, 1.50% and 1.70%, respectively, in rank-1 accuracy. This performance demonstrates GAF-Net's effectiveness in using full gait cycles, which are more consistently available in iLIDS-VID's long sequences.

A key innovation of GAF-Net is its use of skeleton-based gait information, which differs from traditional silhouette-based approaches. This method improves the accuracy of re-identification by capturing more precise movement patterns over time. Skeleton-based approaches work better than silhouette-based methods because they are less affected by changes in clothing and can detect small details in movement. This allows for better identification of individuals, especially in situations where people look similar or backgrounds are complex, which can make silhouette extraction difficult.

Performance on MARS

On the MARS dataset, which presents more challenging real-world conditions, GAF-Net, utilizing MGH for appearance features, showed competitive performance, with a mean average precision (mAP) of 86.09% and a rank-1 accuracy of 89.78%. However, it fell slightly behind recent approaches such as DCCT (mAP: 87.5%; rank-1: 92.3%) and DenseIL (mAP: 87.0%; rank-1: 90.8%). This minor performance gap can be attributed to MARS's diverse conditions, including shorter and more variable sequence lengths, which can impede reliable gait feature extraction. Furthermore, the presence of distractors and false detections in the MARS dataset increases the complexity of the re-identification task, potentially affecting GAF-Net's performance.

Comparative Dataset Analysis

The performance difference between iLIDS-VID and MARS highlights GAF-Net's strengths and limitations under varied conditions. iLIDS-VID's longer sequences allow GAF-Net to fully utilize its skeleton-based gait analysis. In contrast, MARS presents the following three key challenges: shorter sequences, a much larger number of tracklets ($20,000$ vs. 600 in iLIDS-VID), and the presence of distractors and false detections.

MARS's shorter sequences often hinder complete gait cycle extraction, while its larger dataset size increases scenario diversity and the likelihood of encountering distractors or false detections. These factors explain GAF-Net's smaller performance gains on MARS compared to iLIDS-VID while still demonstrating its adaptability to diverse conditions.

Efficacy of Fusion Strategy

Both feature-level and score-level fusion strategies are effective, with greater benefits on the iLIDS-VID dataset. The conditional concatenation method, especially when combined with principal component analysis (PCA), improved performance on both datasets. Specifically, the PCA-enhanced method increased rank-1 accuracy by 0.32% on iLIDS-VID and 0.38% on MARS compared to simple concatenation. These results indicate that reducing dimensionality and aligning features significantly improve the integration of appearance and gait information.

Strengths and Limitations

GAF-Net demonstrates superior performance in synthesizing appearance and motion cues, outperforming existing methods across various accuracy metrics. Its integration of skeletal gait information enhances lightweight architectures, indicating potential for efficient implementation. However, the method's effectiveness is heavily dependent on reliable gait feature extraction, leading to performance discrepancies between controlled and real-world environments. GAF-Net shows sensitivity to video sequence length, performing optimally with longer, uniform sequences but struggling with shorter, variable ones. This characteristic potentially limits its applicability in scenarios requiring rapid identification from brief video segments (up to 10 frames). These limitations highlight areas for future improvement, particularly in developing more robust gait feature extraction techniques capable of handling the variability inherent in real-world surveillance footage and in optimizing the method for shorter video sequences and reduced computational load.

6. Conclusions

This study introduces GAF-Net, a novel method for video-based person re-identification (Re-ID) that uniquely integrates skeleton-based gait features with appearance information. Unlike traditional approaches, GAF-Net leverages precise movement patterns captured by skeletal data, offering robustness to clothing changes and complex backgrounds. We evaluated GAF-Net on the iLIDS-VID and MARS datasets, demonstrating its versatility across different video lengths and environments. On iLIDS-VID, GAF-Net achieved state-of-the-art performance, with 93.20% rank-1 and 99.34% rank-5 accuracies, surpassing existing methods. For MARS, despite challenges with shorter sequences, it showed competitive results, with 86.09% mAP and 89.78% rank-1 accuracy. Notably, GAF-Net consistently improved upon appearance-only architectures, enhancing PiT on iLIDS-VID and MGH on MARS.

Our findings highlight GAF-Net's potential and areas for future work. Despite excelling with long, uniform sequences, it faced limitations in diverse environments and with shorter videos. Future research should focus on improving gait feature extraction for varied conditions, developing adaptive fusion strategies, and optimizing computational efficiency. By addressing these challenges, GAF-Net can evolve into a more robust and practical Re-ID system, advancing real-world applications in security and surveillance. As the field progresses, we anticipate Re-ID methods that perform well across diverse scenarios, significantly impacting smart city technologies.

Author Contributions: Conceptualization, M.B., R.I., L.N., D.M. and S.D.; methodology, M.B., R.I., D.M. and S.D.; software, M.B.; validation, M.B., R.I., D.M. and S.D.; formal analysis, M.B., R.I., D.M. and S.D.; investigation, M.B.; writing—original draft preparation, M.B.; writing—review and editing, M.B., R.I., L.N., D.M. and S.D.; visualization, M.B.; supervision, R.I., L.N., D.M. and S.D.; project administration, R.I., D.M. and S.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors..

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kim, J.; Shin, W.; Park, H.; Baek, J. Addressing the occlusion problem in multi-camera people tracking with human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 5462–5468.
2. Iguernaissi, R.; Merad, D.; Aziz, K.; Drap, P. People tracking in multi-camera systems: A review. *Multimed. Tools Appl.* **2019**, *78*, 10773–10793. [CrossRef]
3. Merad, D.; Aziz, K.E.; Iguernaissi, R.; Fertil, B.; Drap, P. Tracking multiple persons under partial and global occlusions: Application to customers' behavior analysis. *Pattern Recognit. Lett.* **2016**, *81*, 11–20. [CrossRef]
4. Khan, S.U.; Hussain, T.; Ullah, A.; Baik, S.W. Deep-ReID: Deep features and autoencoder assisted image patching strategy for person re-identification in smart cities surveillance. *Multimed. Tools Appl.* **2024**, *83*, 15079–15100. [CrossRef]
5. Wang, T.; Gong, S.; Zhu, X.; Wang, S. Person re-identification by video ranking. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–14 September 2014; pp. 688–703.
6. Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q. Mars: A video benchmark for large-scale person re-identification. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 868–884.
7. Yan, Y.; Qin, J.; Chen, J.; Liu, L.; Zhu, F.; Tai, Y.; Shao, L. Learning multi-granular hypergraphs for video-based person re-identification. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
8. He, T.; Jin, X.; Shen, X.; Huang, J.; Chen, Z.; Hua, X.S. Dense interaction learning for video-based person re-identification. In Proceedings of the International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 1490–1501.
9. Wang, Y.; Zhang, P.; Gao, S.; Geng, X.; Lu, H.; Wang, D. Pyramid spatial-temporal aggregation for video-based person re-identification. In Proceedings of the International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 12026–12035.
10. Zang, X.; Li, G.; Gao, W. Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Trans. Ind. Inform.* **2022**, *18*, 8776–8785. [CrossRef]

11. Liu, X.; Yu, C.; Zhang, P.; Lu, H. Deeply coupled convolution–transformer with spatial–temporal complementary learning for video-based person re-identification. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–11. [CrossRef] [PubMed]
12. Fan, C.; Peng, Y.; Cao, C.; Liu, X.; Hou, S.; Chi, J.; Huang, Y.; Li, Q.; He, Z. Gaitpart: Temporal part-based model for gait recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14225–14233.
13. Teepe, T.; Khan, A.; Gilg, J.; Herzog, F.; Hörmann, S.; Rigoll, G. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In Proceedings of the IEEE International Conference on Image Processing, Anchorage, AK, USA, 19–22 September 2021; pp. 2314–2318.
14. Boujou, M.; Iguernaissi, R.; Nicod, L.; Merad, D.; Dubuisson, S. GAF-Net: Video-Based Person Re-Identification via Appearance and Gait Recognitions. In Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Rome, Italy, 27–29 February 2024.
15. Suh, Y.; Wang, J.; Tang, S.; Mei, T.; Lee, K.M. Part-aligned bilinear representations for person re-identification. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 402–419.
16. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-scale feature learning for person re-identification. In Proceedings of the International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3702–3712.
17. McLaughlin, N.; Del Rincon, J.M.; Miller, P. Recurrent convolutional network for video-based person re-identification. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1325–1334.
18. Li, J.; Zhang, S.; Huang, T. Multi-scale 3d convolution network for video based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8618–8625.
19. Fu, Y.; Wang, X.; Wei, Y.; Huang, T. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8287–8294.
20. Bai, S.; Ma, B.; Chang, H.; Huang, R.; Chen, X. Salient-to-broad transition for video person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7339–7348.
21. Yang, J.; Zheng, W.S.; Yang, Q.; Chen, Y.C.; Tian, Q. Spatial-temporal graph convolutional network for video-based person re-identification. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2020; pp. 3289–3299.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
23. Song, C.; Huang, Y.; Huang, Y.; Jia, N.; Wang, L. Gaitnet: An end-to-end network for gait based human identification. *Pattern Recognit.* **2019**, *96*, 106988. [CrossRef]
24. Liao, R.; Cao, C.; Garcia, E.B.; Yu, S.; Huang, Y. Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations. In Proceedings of the Chinese Conference on Biometric Recognition, Shenzhen, China, 28–29 October 2017; pp. 474–483.
25. Babaee, M.; Li, L.; Rigoll, G. Gait energy image reconstruction from degraded gait cycle using deep learning. In Proceedings of the European Conference on Computer Vision Workshops, Munich, Germany, 8–14 September 2018.
26. Thapar, D.; Nigam, A.; Aggarwal, D.; Agarwal, P. VGR-net: A view invariant gait recognition network. In Proceedings of the International Conference on Identity, Security, and Behavior Analysis (ISBA), Singapore, 11–12 January 2018; pp. 1–8.
27. Chao, H.; He, Y.; Zhang, J.; Feng, J. Gaitset: Regarding gait as a set for cross-view gait recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Hilton, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8126–8133.
28. Fendri, E.; Chtourou, I.; Hammami, M. Gait-based person re-identification under covariate factors. *Pattern Anal. Appl.* **2019**, *22*, 1629–1642. [CrossRef]
29. Rao, H.; Wang, S.; Hu, X.; Tan, M.; Guo, Y.; Cheng, J.; Liu, X.; Hu, B. A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 6649–6666. [CrossRef] [PubMed]
30. Liao, R.; Yu, S.; An, W.; Huang, Y. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognit.* **2020**, *98*, 107069. [CrossRef]
31. An, W.; Yu, S.; Makihara, Y.; Wu, X.; Xu, C.; Yu, Y.; Liao, R.; Yagi, Y. Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE Trans. Biom. Behav. Identity Sci.* **2020**, *2*, 421–430. [CrossRef]
32. Yu, S.; Tan, D.; Tan, T. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In Proceedings of the International Conference on Pattern Recognition, Hong Kong, 22–24 August 2006; Volume 4, pp. 441–444.
33. Bedagkar-Gala, A.; Shah, S.K. Gait-assisted person re-identification in wide area surveillance. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014; pp. 633–649.
34. Liu, Z.; Zhang, Z.; Wu, Q.; Wang, Y. Enhancing person re-identification by integrating gait biometric. *Neurocomputing* **2015**, *168*, 1144–1156. [CrossRef]
35. Frikha, M.; Chtourou, I.; Fendri, E.; Hammami, M. BiMPeR: A Novel Bi-Model Person Re-identification Method based on the Appearance and the Gait Features. *Procedia Comput. Sci.* **2021**, *192*, 913–922. [CrossRef]
36. Lu, X.; Li, X.; Sheng, W.; Ge, S.S. Long-Term Person Re-Identification Based on Appearance and Gait Feature Fusion under Covariate Changes. *Processes* **2022**, *10*, 770. [CrossRef]

37. Jin, X.; He, T.; Zheng, K.; Yin, Z.; Shen, X.; Huang, Z.; Feng, R.; Huang, J.; Chen, Z.; Hua, X.S. Cloth-changing person re-identification from a single image with gait prediction and regularization. In Proceedings of the Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
38. Tu, H.; Liu, C.; Peng, Y.; Xiong, H.; Wang, H. Clothing-change person re-identification based on fusion of RGB modality and gait features. In *Signal, Image and Video Processing*; Springer: Cham, Switzerland, 2023; pp. 1–10.
39. Soni, K.; Dogra, D.P.; Sekh, A.A.; Kar, S.; Choi, H.; Kim, I.J. Person re-identification in indoor videos by information fusion using Graph Convolutional Networks. *Expert Syst. Appl.* **2022**, *210*, 118363. [CrossRef]
40. Pei, Y.; Huang, T.; van Ipenburg, W.; Pechenizkiy, M. ResGCN: Attention-based deep residual modeling for anomaly detection on attributed networks. In Proceedings of the International Conference on Data Science and Advanced Analytics, Porto, Portugal, 6–9 October 2021; pp. 1–2.
41. Maji, D.; Nagori, S.; Mathew, M.; Poddar, D. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In Proceedings of the Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
42. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
43. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5386–5395.
44. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
46. Hirzer, M.; Beleznai, C.; Roth, P.M.; Bischof, H. Person re-identification by descriptive and discriminative classification. In Proceedings of the Scandinavian Conference on Image Analysis, Ystad, Sweden, 1 May 2011; pp. 91–102.
47. Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; Ouyang, W.; Yang, Y. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5177–5186.
48. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
49. Zhang, Y.; Lin, Y.; Yang, X. AA-RGTCN: Reciprocal global temporal convolution network with adaptive alignment for video-based person re-identification. *Front. Neurosci.* **2024**, *18*, 1329884. [CrossRef] [PubMed]
50. Aich, A.; Zheng, M.; Karanam, S.; Chen, T.; Roy-Chowdhury, A.K.; Wu, Z. Spatio-temporal representation factorization for video-based person re-identification. In Proceedings of the International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 152–162.