GAF-Net: Video-Based Person Re-Identification via Appearance and Gait Recognitions

Moncef Boujou[1][a], Rabah Iguernaissi[1][b], Lionel Nicod[2][c], Djamal Merad[1][d]
and Séverine Dubuisson[1][e]
[1] *LIS, CNRS, Aix-Marseille University, Marseille, France*
[2] *CERGAM, Aix-Marseille University, Marseille, France*

Keywords: Deep Learning, Computer Vision, Person Re-Identification, Gait Recognition, Representation Learning.

Abstract: Video-based person re-identification (Re-ID) is a challenging task aiming to match individuals across various cameras based on video sequences. While most existing Re-ID techniques focus solely on appearance information, including gait information, could potentially improve person Re-ID systems. In this study, we propose, GAF-Net, a novel approach that integrates appearance with gait features for re-identifying individuals; the appearance features are extracted from RGB tracklets while the gait features are extracted from skeletal pose estimation. These features are then combined into a single feature allowing the re-identification of individuals. Our numerical experiments on the iLIDS-Vid dataset demonstrate the efficacy of skeletal gait features in enhancing the performance of person Re-ID systems. Moreover, by incorporating the state-of-the-art PiT network within the GAF-Net framework, we improve both rank-1 and rank-5 accuracy by 1 percentage point.

1 INTRODUCTION

Person re-identification is one of the most important keys for the automation of procedures related to various application domains such as surveillance & security (Kim et al., 2023; Iguernaissi et al., 2019) and retail applications (Merad et al., 2016), among others. It aims to recognize individuals across various camera views, leveraging appearance-based cues, including clothing, facial features, and physique. Yet, such cues can be affected by changes in lighting, pose, background noise, and similar-looking individuals. Conventionally, the Re-ID problem is defined as the ability of matching a set of query identities $\mathcal{Q}$ with a set of gallery identities $\mathcal{G}$. A gallery consists of a set $G = \{g_i \mid i = 1, 2, \ldots, N\}$ of N images $\mathbf{g}_i$ from N different identities. Given a query image represented as $\mathbf{q}$, its identity, denoted by i^*, can be determined using the following equation:

$$i^* = \arg\max_i \text{similarity}(\mathbf{q}, \mathbf{g}_i). \quad (1)$$

Here similarity$(\cdot,\cdot)$ describes a function that measures similarity between distinct images.

Recently, research in the field of Re-ID tend toward the video-based methodologies, supported by datasets such as in (Zheng et al., 2016; Wang et al., 2014). Instead of using single images, these techniques evaluate video sequences, underlining the importance of understanding temporal fluctuations and video data's dynamics. Such data offers enhanced spatial and temporal details by capturing multiple frames for each subject. These data coupled with the recent advances in Deep learning have achieved a significant Re-ID improvement, with methods employing hypergraphs (Yan et al., 2020) and strategies like Pyramidal Spatial-Temporal Aggregation (Wang et al., 2021) as well as Vision Transformers in (Zang et al., 2022). However, most of these methods focus primarily on appearance-based cues. Thus, challenges, particularly related to similar appearances, persist.

To address this issue, we propose a novel framework based on combining appearance-based features with skeleton-based gait recognition. Gait recognition has emerged as an important biometric modality in computer vision. It identifies individuals based on their distinct walking patterns. There are two main categories of methods: silhouette-based method (Fan

[a] https://orcid.org/0009-0005-8339-1485
[b] https://orcid.org/0000-0002-1728-3532
[c] https://orcid.org/0000-0002-8845-786X
[d] https://orcid.org/0000-0001-9233-0020
[e] https://orcid.org/0000-0001-7306-4134

Boujou, M., Iguernaissi, R., Nicod, L., Merad, D. and Dubuisson, S.
GAF-Net: Video-Based Person Re-Identification via Appearance and Gait Recognitions.
DOI: 10.5220/0012364200003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 493-500
ISBN: 978-989-758-679-8; ISSN: 2184-4321

et al., 2020), and the skeleton-based method (Teepe et al., 2021). The silhouette method looks at the shape and movement of a person, while the skeleton method focuses on how the joints move. Despite advances in both methods, various factors such as physical condition, emotional state, or footwear type still pose significant challenges to walking rhythm variations.

This approach combines gait and appearance using advanced pose estimation to extract gait features from skeletal movements, regardless of appearance changes. These are then merged with appearance features in a unified framework, creating a distinctive individual representation by integrating both gait and appearance characteristics.

Our proposed GAF-Net achieves state-of-the-art performance on the challenging video-based benchmark, iLIDS-VID (Wang et al., 2014).

The main contributions of this paper are:

- **A Multimodal Framework:** Incorporating both appearance-centric gait information based on pose rather than silhouettes for enhanced person Re-ID.
- **A Performance Gain:** the proposed GAF-Net achieves state-of-the-art performance on the iLIDS-Vid dataset, and experiment studies of parameters are presented.

This paper is structured as follows: in Section 2, related works are reviewed and analyzed, then the proposed method is introduced in Section 3. Experimental results and analysis are presented in Section 4 and Section 5. Finally, Section 6 concludes the paper.

2 RELATED WORKS

The primary challenge in person re-identification (Person Re-ID) lies in finding the optimal representation of an image or video for Re-ID. The following subsections describe the three main approaches used for Person Re-ID : appearance-based, gait-based, and a fusion of both methods.

2.1 Appearance-Based Person Re-ID

Video-based Person Re-ID is a complex task, that requires identification of individuals across different camera views. The challenge relies on an adapted fusion of spatial and temporal features. A variety of deep-learning architectures achieved a significant increase of performance, largely outperforming classical methods.

Among these methods, Convolutional Neural Networks (CNNs) are the most commonly used to extract appearance features directly from video sequences as shown in (Suh et al., 2018) and (Zhou et al., 2019). Nonetheless, they sometimes miss some discriminative details both in spatial and temporal domains. Recurrent Neural Networks (RNNs) (McLaughlin et al., 2016) have been used to encapsulate sequence-level representations, preserving prior frame information. 3D-CNNs, (Li et al., 2019), have been used to holistically capture both spatial and temporal cues.

Attention-driven models, such as those presented in (Fu et al., 2019), aggregate spatial and temporal attention mechanisms to derive latent individual representations. However, these models may occasionally miss crucial information due to their context window's constraints.

Recently, there has been a surge in the adoption of graph-based approaches, like in (Yang et al., 2020). These models enhance the discriminative power of representations due to their inherent ability to reveal complex relationships between data. An application of this is MGH by (Yan et al., 2020), a multi-granular hypergraph learning framework that harnesses multi-granular spatial and temporal cues in tracklets through the utilization of a hypergraph neural network. During the learning process, it uses an attention mechanism to aggregate node-level features, generating more distinctive graph representations.

Furthermore, Pyramid Spatial-Temporal Aggregation (PSTA) (Wang et al., 2021) identifies spatial correlations within frames and uses temporal consistency information. This method highlights discriminative features while suppressing irrelevant ones.

Recently Vision Transformers (ViTs) (Dosovitskiy et al., 2020) have emerged in Person Re-ID. These architectures, such as, PiT (Zang et al., 2022), use transformers with diverse pyramidal structures to encapsulate information from different patches.

Nevertheless, challenges persist, notably in differentiating between individuals dressed in similar clothing styles, a common issue that complicates accurate person Re-ID.

2.2 Gait Recognition

Gait recognition has emerged as a biometric modality in computer vision for person recognition. It is based on the extraction of discriminative spatio-temporal features representing the walking pattern of a given individual.

Gait recognition methods can be categorized into appearance-based and model-based approaches. Most of the existing studies are appearance-based. They are mainly using either the entire silhouettes or their shapes, as in (Song et al., 2019), or focus on specific body segments, as in (Fan et al., 2020).

Model-based approaches, in contrast, use the physical structure of the body, crafting $2D$ or $3D$ skeletons, as in (Liao et al., 2017). With the recent advances in pose estimation, these strategies have become more reliable. They can be categorized into image-based, sequence-based, and set-based.

Image-based techniques, like the Gait Entropy Images (GEI) in (Babaee et al., 2018), aim to encapsulate an entire gait cycle in a single image. Although these are computationally efficient, they often miss some temporal features. While, sequence-based and set-based approaches try to capture temporal nuances. Sequence-based approaches, tools such as 3D-CNNs (Thapar et al., 2018) and LSTMs (Liao et al., 2017), stand out for capturing temporal nuances. However, their computational costs are high. On the other hand, set-based approaches (Chao et al., 2019) mix inputs without explicitly modeling temporal nuances, thereby reducing computational complexity. A novel temporal module introduced in GaitPart (Fan et al., 2020) focuses on capturing brief temporal characteristics.

The model-based techniques are, usually, using pose estimation to extract skeleton information for gait recognition. These techniques often employ conventional CNNs, as demonstrated in (Liao et al., 2020), or they might integrate CNNs with LSTMs, as shown in (An et al., 2020). Another approach proposes the use of Graph Convolutional Networks to robustly represent human body motion through the extracted skeletal structure. as in (Teepe et al., 2021).

Yet, caution is necessary when generalizing the efficacy of these methods. While these approaches excel on controlled environment datasets like CASIA-B (Yu et al., 2006), their performance may drastically diminish in real-world scenarios.

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

Even though fusing both gait and appearance features gave promising results in Person Re-Id, only a few works have studied this fusion. For instance, (Bedagkar-Gala and Shah, 2014) combined GEI, a spatial representation derived from motion energy used to capture the most discriminative gait features of individuals from their silhouettes, with color attributes for long-term person re-identification. Another work (Liu et al., 2015) fused GEI with color and texture at the feature level and tested it on gait dataset, similarly (Frikha et al., 2021)introduces a bimodal person Re-ID method combining appearance (modeled by MLSAR) and gait features (modeled by GAIDS), fused using a score fusion method.

In the same context, some deep learning architectures combine gait attributes, especially silhouette-based ones, with high-level appearance feature extractors. An example is the approach proposed in (Lu et al., 2022), which combines ResNet-extracted appearance features with gait features derived from an Improved-Sobel-Masking Active Energy Image (IS-MAEI), thus facilitating comprehensive gait representation.

For long-term Person Re-ID scenarios involving cloth changing, GI-REID (Jin et al., 2022) introduces a dual-stream framework comprising an auxiliary gait recognition stream (Gait-Stream) and an image Re-ID Stream (Appearance). GI-ReID's gait recognition (silhouette-based) stream enhances cloth-independent identity learning and ensures better performance across various benchmarks through a dual-stream framework, thus demonstrating the growing potential of hybrid methods.

With significant advances in pose-based gait recognition, we propose a new method combining visual appearance features and skeleton-based gait recognition, rather than silhouettes as previously done, to address the problem of Person Re-ID.

3 PROPOSED METHOD

Video-based person Re-ID focuses on matching individuals across video segments acquired from different cameras. Typically, the Re-ID task is a metric learning challenge, whose aim is to find a function ϕ that maps tracklets into a metric subspace in which similar tracklets are close to each other.

In this section, we introduce Gait-Appearance Fusion Network (GAF-Net), which comprises three key components: a gait feature extraction module, an appearance feature extraction module, and a feature fusion module. The first module is based on GaitGraph which combines skeletal poses with a Graph Convolutional Network (GCN) to extract gait features. The second module uses various backbone architectures to extract image-level appearance features. Then, both the appearance features and the feature-level gait representation are normalized and then merged in the last module to obtain final feature representation of the individual. The entire process is depicted in Figure 1.

For our Re-ID problem, let's consider a person identified by Id i, associated with a sequence of T color images of size $W \times H$ (with H the height and W the width of images). This can also be seen as a $4D$ tensor $\mathbf{A}^i \in \mathbb{R}^{W \times H \times 3 \times T}$.

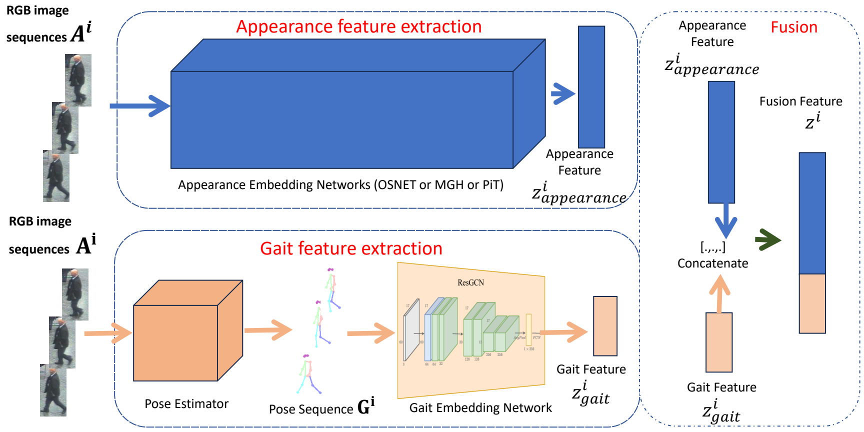
RGB image sequences A^i
Appearance feature extraction
Appearance Embedding Networks (OSNET or MGH or PiT)
Appearance Feature $z^i_{appearance}$
RGB image sequences $\mathbf{A}^\mathbf{i}$
Gait feature extraction
Pose Estimator
Pose Sequence $\mathbf{G}^\mathbf{i}$
ResGCN
Gait Embedding Network
Gait Feature z^i_{gait}
Fusion
Appearance Feature $z^i_{appearance}$
Fusion Feature z^i
[.,.,.]
Concatenate
Gait Feature z^i_{gait}

Figure 1: Overview of GAF-Net and its three main modules: the appearance feature module (with various backbones), the gait feature one, and the fusion one.

3.1 Gait Feature Extraction Module

To extract gait features, we, first, used the SOTA method (Teepe et al., 2021) to estimate the human pose from each video frame. This estimate is then fed to the ResGCN (Pei et al., 2021) to derive a graph representation.

3.1.1 Skeleton Graph Representation

To construct the graph representation of the gait, we start by estimating the $2D$ pose of the individual within each frame and then, combining these poses along the whole frame sequence. To achieve this, For each image in the sequence, we detect keypoints such as the nose, eyes, and knees using a pose estimator. Humans can be represented by a skeleton, modeled by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ denote node (joints) and edge (bones) sets, respectively. For $2D$ human pose estimation, we used YOLO-pose (Maji et al., 2022), rather than HRNET (Cheng et al., 2020), that, although used in GaitGraph (Teepe et al., 2021), has a higher computational cost. The model has been pre-trained on the COCO dataset (Lin et al., 2014) containing annotations for $N = 17$ keypoints in human posture.

Then, these $2D$ poses are used to model the gait as a series of graphs. Each joint of the body (node of a graph) is a C-dimensional feature vector $\mathbf{g}_{t,n}$ that contains the n joints position in frame t ($n \in [1, N]$ and $t \in [1, T]$) plus its confidence value. Then the gait of person Id i forms a $3D$ tensor $\mathbf{G}^i \in \mathbb{R}^{T \times N \times C}$, such as:

$$\mathbf{G}^i = \{\mathbf{g}^i_{t,n} \in \mathbb{R}^C \mid t, n \in \mathbb{Z}, 1 \leq t \leq T, 1 \leq n \leq N\}$$

A pose at t for person Id i is then given by matrix $G^i_t \in \mathbb{R}^{N \times C}$. In our case, we use $2D$ positions, then $C = 3$.

3.1.2 Gait Vector Representation

As in GraphGait (Teepe et al., 2021), the central architecture is based on the ResGCN network proposed by (Pei et al., 2021) that has been adapted for our gait recognition context. The ResGCN is fed by feature tensors $\mathbf{G}^i$ (see Section 3.1.1) of all tracklets. The architecture of a ResGCN block mainly consists of Graph Convolution and conventional $2D$ temporal convolution layers, plus a residual connection and optionally a bottleneck structure. The network is then constructed by a succession of multiple ResGCN blocks. Following this, an average pooling operation is applied, before a fully connected layer that generates the resulting feature vector. For the loss function, we used the supervised contrastive (SupCon) loss (Khosla et al., 2020).

Then, for each gait graph representation $\mathbf{G}^i$ of a given individual with Id i, we obtain a vector representation of gait $\mathbf{z}^i_{\text{gait}}$ such that:

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

3.2 Appearance Feature Extraction Module

Appearance features are obtained from commonly used backbone architectures for Re-ID, including ResNet-50 (He et al., 2016), the OSNet framework (Zhou et al., 2019), the precision-optimized MGH employing multi-granularity hypergraphs (Zhang et al., 2020), or the high-dimensional PiT model (Zang et al., 2022). These models are used to obtain the vector $\mathbf{z}^i_{\text{appearance}}$, that encodes the appearance of a person with Id i from $\mathbf{A}^i$ such that:

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

3.3 Feature Fusion Module

There are many ways to fuse features from different sets. It was shown in (Lu et al., 2022) that concatenation outperforms bitwise addition or Hadamard product for gait and appearance feature fusion.

From the previous extraction steps, for each person i, we obtain $\mathbf{z}^i_{\text{appearance}}$, which encodes the visual details of the tracklet, and $\mathbf{z}^i_{\text{gait}}$, which encodes gait pattern of the individual. We then perform the weighted concatenation these two vectors to obtain a unique feature vector representing the individual i. This fusion is expressed as:

$$\mathbf{z}^i = \phi_\lambda\left(\mathbf{z}^i_{\text{appearance}}, \mathbf{z}^i_{\text{gait}}\right) = \left[\mathbf{z}^i_{\text{appearance}}, \lambda \cdot \mathbf{z}^i_{\text{gait}}\right] \quad (4)$$

Where $\lambda \in [0,1]$ is a penalty factor, modulating the contribution of the gait features.

4 EXPERIMENTAL SETTINGS

Datasets. We evaluate GAF-Net on the iLIDS-VID dataset (Wang et al., 2014), which encompasses 600 videos from 300 pedestrians captured from two disjoint cameras within an airport. The sequence length of iLIDS-VID ranges from 23 to 192 frames with an average of 73 frames. This dataset presents significant Re-ID challenges, including background clutter, occlusions (self and mutual) and varying lighting conditions.

We deliberately choose not to evaluate our proposed method on the commonly used PRID2011 (Hirzer et al., 2011) and MARS (Zheng et al., 2016) datasets. Regarding the PRID2011 dataset, our choice is justified by the fact that it no longer poses a significant challenge for video-based person re-identification; state-of-the-art approaches have demonstrated exceptional performance, with rank-1 metrics surpassing 95% (Wang et al., 2021).

Turning to the MARS dataset, our decision is driven by the specific characteristics of its videos, featuring an average sequence length of only 13.2 frames. This brevity raises concerns, as many video sequences in MARS may not capture a complete gait cycle of pedestrians. Our aim is to ensure that our evaluation context aligns closely with the challenges and complexities inherent in person re-identification, leading us to focus on the iLIDS-VID dataset for a more relevant assessment of our GAF-Net approach.

Models and Implementation Details. In our experiments, we consider three backbones for appearance feature extraction: OSNet (Zhou et al., 2019), MGH (Zhang et al., 2020), and PiT (Zang et al., 2022).

- OSNet: We use the pre-trained architecture on ImageNet. Each input frame is resized to 256 × 128. Training the model for 60 epochs with a batch size of 6, we employ the Adam optimizer and cross-entropy loss. The learning rate is set to $10^{-3.5}$. For each tracklet, we use a sequence length of 8.
- MGH: For this architecture, images were resized to 256 × 128. Each batch randomly samples 8 sub-sequences from 8 individuals. The setup includes a hypergraph layer with $L = 2$ and $K = 3$ neighbors. Default spatial partitions are (1, 2, 4, 8), with temporal thresholds at (1, 3, 5). Adam optimizer is used with a 5×10^{-4} weight decay. Starting with a learning rate of $10^{-3.5}$, it is reduced by a factor of 10 every 100 epochs until 300 epochs are completed. Graph-level features are concatenated at the end, using cosine similarity as the matching metric.
- PiT: This model relies on a ViT-B16 transformer pre-trained on ImageNet. Images are resized to 256 × 128, with 8 images selected to represent each tracklet. We set the batch size and parameter m to 16 and 11, respectively. The convolution layer's kernel size $k = 16$, with a stride $s = 12$, gives an output feature embedding dimensions of $21 \times 10 \times 768$. We fine-tune trade-off parameters λ_1 and λ_2 to 1.0 and 1.5, respectively. Division parameters D_v, D_h, and D_p are fixed to 2, 3, and 6 respectively. Training uses standard Stochastic Gradient Descent (SGD) with momentum, starting from a learning rate of 10^{-2} for 120 epochs. The learning rate is adjusted using cosine annealing. To optimize training, convolution and transformer layers remain frozen during the initial five epochs, allowing refinement of the classifier.

Regarding the gait feature extraction module, we follow the implementation described in (Teepe et al., 2021). We fine-tune GaitGraph over CASIA-B on 60 frames from each tracklet, and has a 1-cycle learning rate and a 10^{-5} weight decay penalty. The initial cycle's maximum learning rate is 10^{-2} for 1000 epochs with a loss function temperature of 10^{-2}. In the second cycle, this rate is adjusted to 10^{-5} for 300 epochs, using a batch size of 128. We used PyTorch to build and train our models, and we ran these experiments on a QUADRO RTX 4000.

Evaluation Metrics. We use the original ten splits present in iLIDS-VID (Wang et al., 2014), dividing identities randomly into two groups, each one containing 150 identities, to create training and test sets.

Table 1: Person Re-ID performances (rank-1/rank-5/rank-10/rank-20) on iLIDS-VID dataset depending on the model and the integration or not of gait information.

Backbone	Appearance				Appearance + gait			
	rank-1	rank-5	rank-10	rank-20	rank-1	rank-5	rank-10	rank-20
OSNet (Zhou et al., 2019)	59.20	82.60	89.34	94.80	70.93	88.40	93.00	96.54
MGH (Zhang et al., 2020)	85.60	97.1	99.00	99.30	90.40	98.66	98.99	99.66
PiT (Zang et al., 2022)	92.07	98.93	99.80	100	93.07	99.27	99.74	99.94

For a consistent evaluation standard, we use Cumulative Matching Characteristic (CMC) curves ranging from rank-1 (R-1) to rank-20 (R-20). Re-ranking is consistently not applied.

5 EXPERIMENTAL RESULTS

In our experiments, we concentrated on three aspects. Firstly, we examined the impact of incorporating gait information on the performance of a given appearance feature extraction backbone network. Secondly, we compared GAF-Net, utilizing the state-of-the-art PiT backbone, against existing Person Re-Id methods. Thirdly, we conducted an ablation study to explore the effect of the fusion factor λ on GAF-Net's performance

5.1 Fusion Improvements

Table 1 gives comparative results obtained for the integration of gait features into various appearance feature backbones, each one optimized with a specific λ value used for the fusion (see Eq. 4). We can see that the fusion with the GaitGraph architecture improved all results:

- OSNet, with its compact architecture producing 512-dimensional representations, initially achieved a rank-1 accuracy of 59.20%, which was increased by 11.73%.
- The MGH architecture, generating 5120-dimensional vectors, registered a 4.7% improvement in rank-1 accuracy.
- The PiT method, which outputs a 9216-dimensional vector, has an initial rank-1 accuracy of 90.34% that was increased by 2.73%.

5.2 Comparison with the SOTA Methods

In this section, we compare GAF-Net(PiT) to other state-of-the-art methods. These methods mainly exploit appearance and visual cues within the sequences. The evaluation was performed on the iLIDS-VID dataset. Comparative results are given in Table 2.

Table 2: Comparison with the state-of-the-arts on iLIDS-VID.

Backbone	References	Results			
		rank-1	rank-5	rank-10	rank-20
PiT	(Zang et al., 2022)	92.07	98.93	**99.80**	**100**
PSTA	(Wang et al., 2021)	91.5	98.1	-	-
STRF	(Aich et al., 2021)	89.3	-	-	-
MGH	(Yan et al., 2020)	85.6	97.1	-	99.5
GAF-Net	(Ours)	**93.07**	**99.27**	99.74	99.94

Several conclusions emerged from our study:

Firstly, we can see that the integration of gait information greatly improves performance. Our GAF-Net outperformed existing methods according to the results given in the PiT paper, increasing by 1% rank-1 and by 0.34% rank-5 accuracies. Compared to the Pyramid Spatial-Temporal Aggregation (PSTA —a method that amalgamates frame-level features with hierarchical temporal elements to craft a nuanced video-level representation—) our GAF-Net gives higher scores. It surpasses PSTA by 1.57% at rank-1 and gives comparable scores at rank-5. Compared to STRF, a technique that uses spatio-temporal representation coupled with 3D CNN, our methodology increases by 3.77% the rank-1 score. Finally, compared to Multi-Granular Hypergraph (MGH), GAF-Net demonstrably increases by 7.47% the rank-1 and by 2.17% the rank-5 scores.

5.3 Analysis of the Fusion Factor λ

Figure 2 shows the impact of varying the λ value of Eq. 4 on the rank-1 accuracy for the three architectures: OSNet, PiT, and MGH.

For the OSNet architecture, we can see its rank-1 accuracy increases with λ until it reaches $\lambda = 0.8$, with a rank-1 accuracy of 70.93%. This accuracy stays a little stable, then decreases for higher λ values.

For the MGH architecture, the rank-1 accuracy also increases with λ, but more slowly until $\lambda = 0.8$, reaching an accuracy of 90%, which stays stable with higher values for λ.

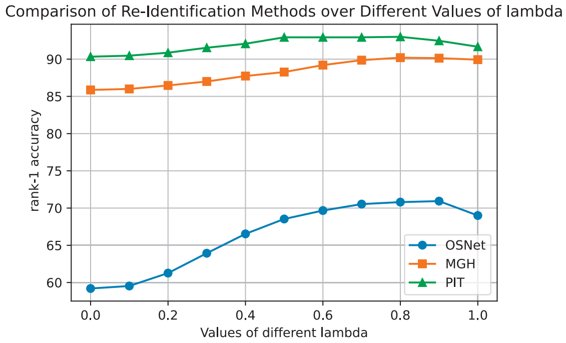

Figure 2: Impact of the fusion factor value λ (varying from 0 to 1) on the rank-1 accuracy.

The PiT model from a high rank-1 accuracy (90.33%) without adding any gait information ($\lambda = 0$). This accuracy slowly increases, reaching 93.00% with λ until $\lambda = 0.7$, then it decreases.

To summarize, each of the three architecture's accuracies are increased, up to 10%, by adding gait information. Even the SOTA PiT accuracy was improved to 3%. test have shown that we should choose $\lambda \in [0.6, 0.8]$.

6 CONCLUSION

In this paper, we proposed GAF-Net, a novel video-based person Re-ID approach that combines appearance features with gait features extracted from skeletal structures. Our approach is generic and could be easily combined with different appearance feature extractor to boost their performance. Two key findings emerge from our study: 1) Integrating gait information into a fixed backbone consistently enhances performance. 2) GAF-Net (PiT) outperforms state-of-the-art approaches in both rank-1 and rank-5 metrics on the iLIDS-VID dataset.

The experimental results demonstrate how effectively gait features, extracted from skeletal structures, enhance person Re-ID system performance. We believe that skeletal-based gait fusion holds the promise of outperforming silhouette-based counterparts, owing to its adeptness in handling complex scenes with multiple individuals. We intend to validate our claim in a future work.

REFERENCES

Aich, A., Zheng, M., Karanam, S., Chen, T., Roy-Chowdhury, A. K., and Wu, Z. (2021). Spatio-temporal representation factorization for video-based person re-identification. In *International conference on computer vision*, pages 152–162.

An, W., Yu, S., Makihara, Y., Wu, X., Xu, C., Yu, Y., Liao, R., and Yagi, Y. (2020). Performance evaluation of model-based gait on multi-view very large population database with pose sequences. *IEEE transactions on biometrics, behavior, and identity science*, 2(4):421–430.

Babaee, M., Li, L., and Rigoll, G. (2018). Gait energy image reconstruction from degraded gait cycle using deep learning. In *European conference on computer vision workshops*.

Bedagkar-Gala, A. and Shah, S. K. (2014). Gait-assisted person re-identification in wide area surveillance. In *Asian conference on computer vision*, pages 633–649. Springer.

Chao, H., He, Y., Zhang, J., and Feng, J. (2019). Gait-set: Regarding gait as a set for cross-view gait recognition. In *AAAI conference on artificial intelligence*, volume 33, pages 8126–8133.

Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., and Zhang, L. (2020). Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Conference on computer vision and pattern recognition*, pages 5386–5395.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale.

Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., and He, Z. (2020). Gaitpart: Temporal part-based model for gait recognition. In *Conference on computer vision and pattern recognition*, pages 14225–14233.

Frikha, M., Chtourou, I., Fendri, E., and Hammami, M. (2021). Bimper: A novel bi-model person re-identification method based on the appearance and the gait features. *Procedia Computer Science*, 192:913–922.

Fu, Y., Wang, X., Wei, Y., and Huang, T. (2019). Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI conference on artificial intelligence*, volume 33, pages 8287–8294.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Conference on computer vision and pattern recognition*, pages 770–778.

Hirzer, M., Beleznai, C., Roth, P. M., and Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *Scandinavian Conference on Image Analysis*, pages 91–102. Springer.

Iguernaissi, R., Merad, D., Aziz, K., and Drap, P. (2019). People tracking in multi-camera systems: a review. *Multimedia Tools and Applications*, 78:10773–10793.

Jin, X., He, T., Zheng, K., Yin, Z., Shen, X., Huang, Z., Feng, R., Huang, J., Chen, Z., and Hua, X.-S. (2022). Cloth-changing person re-identification from a single image with gait prediction and regularization. In *Conference on Computer Vision and Pattern Recognition*.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Kim, J., Shin, W., Park, H., and Baek, J. (2023). Addressing the occlusion problem in multi-camera people tracking with human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5462–5468.

Li, J., Zhang, S., and Huang, T. (2019). Multi-scale 3d convolution network for video based person re-identification. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 8618–8625.

Liao, R., Cao, C., Garcia, E. B., Yu, S., and Huang, Y. (2017). Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In *Chinese conference on biometric recognition*, pages 474–483. Springer.

Liao, R., Yu, S., An, W., and Huang, Y. (2020). A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognition*, 98:107069.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer.

Liu, Z., Zhang, Z., Wu, Q., and Wang, Y. (2015). Enhancing person re-identification by integrating gait biometric. *Neurocomputing*, 168:1144–1156.

Lu, X., Li, X., Sheng, W., and Ge, S. S. (2022). Long-term person re-identification based on appearance and gait feature fusion under covariate changes. *Processes*, 10(4):770.

Maji, D., Nagori, S., Mathew, M., and Poddar, D. (2022). Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Conference on Computer Vision and Pattern Recognition*.

McLaughlin, N., Del Rincon, J. M., and Miller, P. (2016). Recurrent convolutional network for video-based person re-identification. In *Conference on computer vision and pattern recognition*, pages 1325–1334.

Merad, D., Aziz, K.-E., Iguernaissi, R., Fertil, B., and Drap, P. (2016). Tracking multiple persons under partial and global occlusions: Application to customers' behavior analysis. *Pattern Recognition Letters*, 81:11–20.

Pei, Y., Huang, T., van Ipenburg, W., and Pechenizkiy, M. (2021). Resgcn: attention-based deep residual modeling for anomaly detection on attributed networks. In *International Conference on Data Science and Advanced Analytics*, pages 1–2. IEEE.

Song, C., Huang, Y., Huang, Y., Jia, N., and Wang, L. (2019). Gaitnet: An end-to-end network for gait based human identification. *Pattern recognition*, 96.

Suh, Y., Wang, J., Tang, S., Mei, T., and Lee, K. M. (2018). Part-aligned bilinear representations for person re-identification. In *European conference on computer vision*, pages 402–419.

Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., and Rigoll, G. (2021). Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *IEEE International Conference on Image Processing*, pages 2314–2318.

Thapar, D., Nigam, A., Aggarwal, D., and Agarwal, P. (2018). Vgr-net: A view invariant gait recognition network. In *international conference on identity, security, and behavior analysis (ISBA)*, pages 1–8.

Wang, T., Gong, S., Zhu, X., and Wang, S. (2014). Person re-identification by video ranking. In *European conference on computer vision*, pages 688–703. Springer.

Wang, Y., Zhang, P., Gao, S., Geng, X., Lu, H., and Wang, D. (2021). Pyramid spatial-temporal aggregation for video-based person re-identification. In *International conference on computer vision*, pages 12026–12035.

Yan, Y., Qin, J., Chen, J., Liu, L., Zhu, F., Tai, Y., and Shao, L. (2020). Learning multi-granular hypergraphs for video-based person re-identification. In *Conference on computer vision and pattern recognition*.

Yang, J., Zheng, W.-S., Yang, Q., Chen, Y.-C., and Tian, Q. (2020). Spatial-temporal graph convolutional network for video-based person re-identification. In *Conference on computer vision and pattern recognition*, pages 3289–3299.

Yu, S., Tan, D., and Tan, T. (2006). A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *International conference on pattern recognition*, volume 4, pages 441–444. IEEE.

Zang, X., Li, G., and Gao, W. (2022). Multidirection and multiscale pyramid in transformer for video-based pedestrian retrieval. *IEEE Transactions on Industrial Informatics*, 18(12):8776–8785.

Zhang, Z., Lan, C., Zeng, W., and Chen, Z. (2020). Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *Conference on computer vision and pattern recognition*, pages 10407–10416.

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., and Tian, Q. (2016). Mars: A video benchmark for large-scale person re-identification. In *European conference on computer vision*, pages 868–884. Springer.

Zhou, K., Yang, Y., Cavallaro, A., and Xiang, T. (2019). Omni-scale feature learning for person re-identification. In *International Conference on Computer Vision*, pages 3702–3712.