Attention-Based Shape and Gait Representations Learning for Video-Based Cloth-Changing Person Re-Identification

Vuong D. Nguyen[a], Samiha Mirza[b], Pranav Mantini[c] and Shishir K. Shah[d]
Quantitative Imaging Lab, Dept. of Computer Science, University of Houston, Houston, Texas, U.S.A.

Keywords: Video-Based Person Re-Identification, Cloth-Changing Person Re-Identification, Gait Recognition, Graph Attention Networks, Spatial-Temporal Graph Learning.

Abstract: Current state-of-the-art Video-based Person Re-Identification (Re-ID) primarily relies on appearance features extracted by deep learning models. These methods are not applicable for long-term analysis in real-world scenarios where persons have changed clothes, making appearance information unreliable. In this work, we deal with the practical problem of Video-based Cloth-Changing Person Re-ID (VCCRe-ID) by proposing "**A**ttention-based **S**hape and **G**ait Representations **L**earning" (ASGL) for VCCRe-ID. Our ASGL framework improves Re-ID performance under clothing variations by learning clothing-invariant gait cues using a Spatial-Temporal Graph Attention Network (ST-GAT). Given the 3D-skeleton-based spatial-temporal graph, our proposed ST-GAT comprises multi-head attention modules, which are able to enhance the robustness of gait embeddings under viewpoint changes and occlusions. The ST-GAT amplifies the important motion ranges and reduces the influence of noisy poses. Then, the multi-head learning module effectively reserves beneficial local temporal dynamics of movement. We also boost discriminative power of person representations by learning body shape cues using a GAT. Experiments on two large-scale VCCRe-ID datasets demonstrate that our proposed framework outperforms state-of-the-art methods by 12.2% in rank-1 accuracy and 7.0% in mAP.

1 INTRODUCTION

Person Re-Identification (Re-ID) involves matching the same person across multiple non-overlapping cameras with variations in pose, lighting, or appearance. Video-based Person Re-ID has been actively researched for various applications such as surveillance, unmanned tracking, search and rescue, etc. Two main approaches have emerged: (1) Deep learning-based methods (Li et al., 2018a; Liu et al., 2019; Gu et al., 2020) that extract appearance features for Re-ID using Convolutional Neural Networks (CNNs); and (2) Graph-based methods (Yang et al., 2020; Wu et al., 2020; Khaldi et al., 2022) that capture spatial-temporal information using Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017). However, these methods primarily rely on appearance features, making them likely to suffer performance degradation in cloth-changing scenarios where texture information is unreliable. This leads to a more practical Re-ID task called Cloth-Changing Person Re-ID (CCRe-ID).

Several methods have been proposed for image-based CCRe-ID, which attempt to extract clothing-invariant modalities such as body shape (Qian et al., 2020; Li et al., 2021), contour sketches (Yang et al., 2021; Chen et al., 2022), or silhouettes (Hong et al., 2021). Although these cues are more stable than appearance in the long-term, extracting them from single-shot human image remains challenging. On the other hand, video-based data provides motion information that can improve matching ability of the Re-ID system. Video-based CCRe-ID (VCCRe-ID) has not been widely studied for two main reasons. First, there are only two public datasets: VCCR (Han et al., 2023), and CCVID (Gu et al., 2022) that are constructed from gait recognition datasets. Second, capturing identity-aware cloth-invariant cues from video sequences remains challenging in real-world scenarios. Texture-based works (Gu et al., 2022; Cui et al., 2023) have proposed to extract clothing-unrelated features like faces or hairstyles. However, these methods fail under occlusion. Gait recognition models

[a] https://orcid.org/0000-0002-2369-8793
[b] https://orcid.org/0000-0003-3754-6894
[c] https://orcid.org/0000-0001-8871-9068
[d] https://orcid.org/0000-0003-4093-6906

Nguyen, V., Mirza, S., Mantini, P. and Shah, S.
Attention-Based Shape and Gait Representations Learning for Video-Based Cloth-Changing Person Re-Identification.
DOI: 10.5220/0012315900003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 80-89
ISBN: 978-989-758-679-8; ISSN: 2184-4321

have been utilized (Zhang et al., 2018; Zhang et al., 2021) to assist the Re-ID systems. However, these works do not efficiently capture the local temporal features from video sequences. Moreover, they primarily rely on gait cues and overlook identity-relevant shape features. These shortcomings necessitate a more robust approach for VCCRe-ID.

In this work, we propose "**A**ttention-based **S**hape and **G**ait Represenatations **L**earning" (ASGL) framework for VCCRe-ID. Our framework aims to mitigate the influence of clothing changes by extracting texture-invariant body shape and gait cues simultaneously from 3D skeleton-based human poses. The key components of AGSL are the shape learning sub-branch and gait learning sub-branch, both of which are built on Graph Attention Networks (GAT). The shape learning sub-branch is a GAT that processes a 3D skeleton sequences to obtain shape embedding, which is unique to individuals under clothing variations. The gait learning sub-branch is a Spatial-Temporal GAT (ST-GAT) that encodes gait from the skeleton-based spatial-temporal graph by modeling the temporal dynamics from movement of the body parts. This is different from previous works which leverage simple GCNs (Teepe et al., 2021; Zhang et al., 2021; Khaldi et al., 2022). The multi-head attention mechanism in the proposed spatial-temporal graph attention blocks enables the framework to dynamically capture critical short-term movements by attending to important motion ranges in the sequence. This helps mitigate the influence of noisy frames caused by viewpoint changes or occlusion. We also reduce local feature redundancy in capturing motion patterns from pose sequence by narrowing the scope of self-attention operators, producing a discriminative gait embedding with beneficial information for Re-ID. Shape and gait are then coupled with appearance for the global person representation.

In summary, our contributions in this work are as follows: (1) we propose ASGL, a novel framework for the long-term VCCRe-ID task; (2) we propose a ST-GAT for gait learning and a GAT for shape learning, which helps to enhance the discriminative power of identity representations under clothing variations and viewpoint changes; and (3) we present extensive experiments on two large-scale public VCCRe-ID datasets and demonstrate that our framework significantly outperform state-of-the-art methods.

2 RELATED WORKS

2.1 Person Re-ID

Typically, early methods for image-based Re-ID include three main approaches: representation learning (Matsukawa et al., 2016; Wang et al., 2018), metric learning (Ma et al., 2014; Liao et al., 2015), and deep learning (Sun et al., 2018; Luo et al., 2019). Video-based Re-ID methods focus on aggregating frame-wise appearance features using 3D-CNN (Li et al., 2018a; Gu et al., 2020) and RNN-LSTM (Yan et al., 2016; Zhou et al., 2017), or capturing spatial-temporal information using GNNs (Yang et al., 2020; Wu et al., 2020; Khaldi et al., 2022). These methods produce comparable results on traditional person Re-ID datasets (Li et al., 2014; Zheng et al., 2015; Zheng et al., 2016; Li et al., 2018b). However, these datasets were collected in short-term scenarios and the identities present a consistency in appearance and clothing. Existing methods trained on these datasets focus on encoding appearance features and hence are inefficient in long-term scenarios.

2.2 Image-Based CCRe-ID

Recently, several datasets for image-based CCRe-ID have been published (Qian et al., 2020; Yang et al., 2021; Wan et al., 2020). Huang *et al.* (Huang et al., 2020) proposed to capture clothing variations within the same identity using vector-neuron capsules. Body shape cues are explicitly extracted by Qian *et al.* (Qian et al., 2020) using a shape-distillation module, and by Li *et al.* (Li et al., 2021) using adversarial learning. Other works also attempt to extract modalities that are stable in long-term, such as silhouettes (Hong et al., 2021; Jin et al., 2022), or contour sketches (Yang et al., 2021). These works mostly rely on 2D skeleton-based poses, which are affected by viewpoint changes, making extracted features ambiguous for Re-ID. Chen *et al.* (Chen et al., 2021) proposed to estimate and regularize 3D shape parameters using projected 2D pose and silhouettes. Zheng *et al.* (Zheng et al., 2022) leveraged 3D mesh to jointly learnt appearance and shape features. However, image-based setting is sensitive to the quality of Re-ID data and less tolerant to noise due to limited information contained in a single person image.

2.3 Video-Based CCRe-ID

Video-based CCRe-ID has not been widely studied due to the limited number of publicly available datasets. Previous works on VCCRe-ID can be cat-

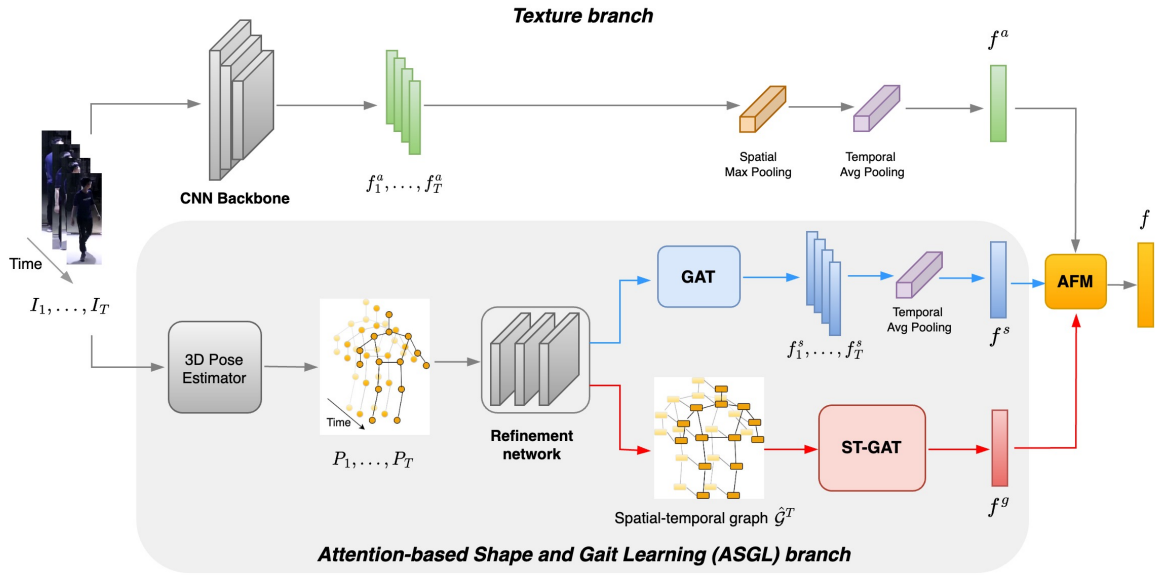

Figure 1: Overview of the proposed ASGL framework. Given a video sequence, for the ASGL branch, 3D pose sequence is first estimated and then refined. A GAT in shape learning sub-branch extracts frame-wise shape features, which are then aggregated for the video-wise shape embedding by a temporal average pooling layer (blue flow). Meanwhile, a spatial-temporal graph is constructed from the refined pose sequence, which is then processed by a ST-GAT to obtain gait embedding (red flow). Appearance, shape and gait are finally fused by the Adaptive Fusion module for the final person representation.

egorized into two main approaches. First, texture-based methods, where Gu *et al.* (Gu et al., 2022) designed clothes-based losses to eliminate the influence of clothes on global appearance. Bansal *et al.* (Bansal et al., 2022) and Cui *et al.* (Cui et al., 2023) leveraged self-attention to attend to appearance-based cloth-invariant features like face or hairstyle. These methods suffer severe performance degradation under occlusion. Second, gait-based methods (Zhang et al., 2018; Zhang et al., 2021; Wang et al., 2022), where motion patterns are captured as features for Re-ID. These works first assume constant walking trajectory from identities, which is not practical in real-world. Then gait cues are encoded from sequences of 2D poses by GNNs. There are two limitations to this approach: (1) viewpoint changes significantly limit the ability to capture body parts movement from 2D pose; and (2) simple GNNs do not capture local motion patterns efficiently. Han *et al.* (Han et al., 2023) proposed to extract human 3D shape cues using a two-stage framework, which needs additional large-scale datasets and results in a heavy training process. In this work, we address these issues by simultaneously extracting body shape along with gait from 3D pose using attention-based variations of GNNs.

3 PROPOSED FRAMEWORK

3.1 Overview

An overview of the proposed ASGL framework is demonstrated in Figure 1 Given a T-frame video sequence $\{I_i\}_{i=1}^{T}$ as input, we first employ an off-the-shelf 3D pose estimation model to estimate frame-wise 3D pose sequence $P = \{P_i\}_{i=1}^{T}$. P is then fed into the refinement network $\mathcal{R}(.)$, giving refined sequence of frame-wise skeleton-based features $\hat{J} = \{\hat{J}_{P_i}\}_{i=1}^{T}$. The shape learning sub-branch which comprises of a Graph Attention Network extracts frame-wise shape feature vectors $\{f_i^s\}_{i=1}^{T}$ from $\hat{J}$, then aggregates $\{f_i^s\}_{i=1}^{T}$ for the video-wise shape embedding f^s using a temporal average pooling layer. Meanwhile, the gait learning sub-branch first connects $\hat{J}$ to yield the spatial-temporal motion graph $\mathcal{G}^{st}$, then uses the proposed Spatial-Temporal Graph Attention Network to encode gait embedding f^g from $\mathcal{G}^{st}$. Since texture information is still important in the cases of slight clothing changes, we extract appearance embedding f^a using a CNN backbone. Finally, f^a, f^s and f^g are fused by the Adaptive Fusion Module for the final person representation f, which is then fed into the cross-entropy loss and pair-wise triplet loss functions for training the framework. In testing stage, matching is performed by comparing the video-wise representations based on cosine distance.

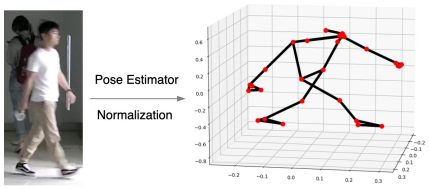

Figure 2: Illustration of 3D pose estimation. Pose is first estimated using an off-the-shelf pose estimator, then normalized to an unified view.

3.2 Attention-Based Shape and Gait Learning Branch

The goal of the Attention-based Shape and Gait Learning (ASGL) branch is to learn body shape and gait features, which serve as complementary information to appearance features for a robust person representation in long-term scenarios. ASGL comprises of a 3D pose estimator, a refinement network to refine the frame-wise skeleton-based pose sequence, a shape learning sub-branch built on a GAT and a gait learning sub-branch built on a ST-GAT.

3.2.1 Pose Estimator and Refinement Network

In contrast to existing methods (Qian et al., 2020; Teepe et al., 2021) that use 2D pose, we utilize 3D pose for learning shape and gait. 3D pose is more robust to camera viewpoint changes and occlusions. We employ an off-the-shelf 3D pose estimator (Bazarevsky et al., 2020) to obtain frame-wise pose sequence $P = \{P_i\}_{i=1}^{T}$. The joint set $J_{P_i} = \{j_i\}_{i=1}^{k}$ corresponding to pose P_i contains k 3D keypoints as illustrated in Figure 2. Each estimated keypoint j_i is represented as a set of three coordinates (x_i, y_i, z_i) indicating the location of certain body parts.

To avoid misalignment in capturing motion patterns caused by camera viewpoint variations, for every keypoint $j_i \in J_{P_i}$, we first translate the raw keypoints to the origin of coordinate:

$$(x_i, y_i, z_i) \rightarrow (x_i + \Delta_{i,x}, y_i + \Delta_{i,y}, z_i + \Delta_{i,z}) \quad (1)$$

where $(\Delta_{i,x}, \Delta_{i,y}, \Delta_{i,z})$ is the translation offset. We then normalize the translated keypoints to an unified view as follows: $\bar{j}_i = \left(\frac{x_i+\Delta_{i,x}}{h}, \frac{y_i+\Delta_{i,y}}{w}, \frac{z_i+\Delta_{i,z}}{h\times w}\right)$, where (h, w) is the size of input frame. The normalized keypoint set is then refined to obtain $\hat{j}_i \in \mathbb{R}^d$ using the refinement network $\mathcal{R}(\cdot)$, i.e. $\hat{j}_i = \mathcal{R}(\bar{j}_i)$. $\mathcal{R}(\cdot)$ consists of a sequence of fully connected layers, which aims to capture fine-grained details of the person's body shape via the high-dimensional keypoint-wise feature vector set $\hat{J}_{P_i} = \{\hat{j}_i\}_{i=1}^{k}$. Output of the refinement network is the refined frame-wise pose sequence $\hat{J} = \{\hat{J}_{P_i}\}_{i=1}^{T}$.

3.2.2 Shape Representation Learning

Body shape remains relatively stable in long-term, thus it can serve as an important cue for CCRe-ID. Using the 3D skeleton representation, shape describes the geometric form of the human body. Intuitively, body shape can not be captured via a single keypoint-wise feature vector. In this work, we propose a Graph Attention Network (GAT) (Velickovic et al., 2017) to model the relations between pairs of connected keypoints. GAT is a type of GCN, which uses message passing to learn features across neighborhoods from the skeleton-based graph. GAT applies an attention mechanism in the aggregation and updating process across several graph attention layers. This helps to exploit the local relationships between body parts for a discriminative shape embedding of the person.

Specifically, a graph that represents the body pose is first constructed from the refined keypoint set $\hat{J}_{P_i} = \{\hat{j}_i\}_{i=1}^{k}$, in which keypoints are nodes and bones are edges of the graph. Each node $\hat{j}_i$ corresponds to a set Q_i containing indices of neighbors of $\hat{j}_i$. Q_i can be constructed via adjacency matrix. Then, the l^{th} layer of GAT $\mathcal{G}$ updates $\hat{j}_i$ by aggregating information from Q_i, given as:

$$\hat{j}_i^{(l+1)} = \sigma\left(\sum_{j\in Q_i} \mathbf{W}_{ij}\theta^{(l)}\hat{j}_i^{(l)}\right) \quad (2)$$

where $\mathbf{W} = (a_{ij}), \mathbf{W} \in \mathbb{R}^{k\times k}$, a_{ij} stores the weighting between $\hat{j}_i$ and $\hat{j}_j$ (i.e. the importance of joint $\hat{j}_j$ to joint $\hat{j}_i$). $\theta^{(l)} \in \mathbb{R}^{d^{(l+1)}\times d^{(l)}}$ is the weight matrix of the l^{th} graph attention layer $\mathcal{G}^{(l)}$, where $d^{(l)}$ is the dimension of layer $\mathcal{G}^{(l)}$. σ is an activation function. GAT implicitly amplifies importances of each joint to its neighbors. To do this, unlike traditional GCNs in which $\mathbf{W}$ is explicitly defined, GAT $\mathcal{G}$ implicitly computes $a_{ij} \in \mathbf{W}$ by:

$$a_{ij} = \text{softmax}_j\, h\left(\theta\hat{j}_i, \theta\hat{j}_j\right) \quad (3)$$

where $h : \mathbb{R}^{d^{(l+1)}} \times \mathbb{R}^{d^{(l+1)}} \rightarrow \mathbb{R}$ is a byproduct of an attentional mechanism. We employ a global max pooling layer to aggregate the higher-order representations of joint set $\hat{J}_{P_i}^{(L-1)}$ after L GAT layers for the frame-wise shape embedding f_i^s of the i^{th} frame as follows:

$$f_i^s = \text{GMP}\left(\hat{J}_{P_i}^{(L-1)}\right) \quad (4)$$

where GMP denotes global max pooling and summarizes the discriminative information in the graph. The frame-wise shape embedding set $\{f_i^s\}_{i=1}^{T}$ is finally fed into a temporal average pooling layer to obtain the video-wise shape representation f^s.

3.2.3 Gait Representation Learning

Unlike previous works (Teepe et al., 2021; Zhang et al., 2021) that encode gait using traditional variations of Spatial-Temporal Graph Convolutional Networks (ST-GCN) (Yan et al., 2018), we propose to learn gait cues using a Spatial-Temporal Graph Attention Network (ST-GAT) $\mathbf{G}_{sta}$. By incorporating an attention mechanism, we enable the network to effectively amplify the important motion patterns and reduce the influence of noisy motion ranges, producing gait representations with high discriminative power.

As shown in Figure 3, input for the gait learning sub-branch is the spatial-temporal graph $\hat{\mathcal{G}}^T \in \mathbb{R}^{T\times k\times d}$ constructed from the refined skeleton sequence $\hat{J} = \left\{\hat{J}_{P_i}\right\}_{i=1}^{T}$. In this work, we follow the spatial-temporal connection as in (Yan et al., 2018), allowing for direct aggregation of moving information from consecutive frames. Our ST-GAT $\mathbf{G}_{sta}$ consists of B consecutive Spatial-Temporal Attention (STA) blocks, followed by a global average pooling to output gait representation f^g. For each STA block, we first employ multi-head attention modules $\mathcal{H}$. As occlusions may lead to noisy frames with unobservable movements of body parts, $\mathcal{H}$ allows to attend to different ranges of motion patterns simultaneously. This helps amplify the most contributing frame-wise skeletons to the global gait encoding and reduce the influence of noisy frames. In this work, $\mathcal{H}$ consists of S independent self-attention modules, which capture the spatial-temporal dynamics of the input spatial-temporal graph:

$$\hat{\mathcal{G}}^{att} = \sigma\left(\frac{1}{S}\sum_{s=1}^{S} W_s \hat{\mathcal{G}}^T \mathbf{A}_s\right) \tag{5}$$

where W_s is the learnable attention matrix of the s^{th} head which weights the edge importance, i.e. the relationships among connected joints, $\mathbf{A}_s$ is the adjacency matrix of the s^{th} head, σ is the activation function and $\hat{\mathcal{G}}^{att}$ is the accumulated output of all heads.

$\hat{\mathcal{G}}^{att}$ is then fed into the multi-head learning module which consists of several 1×1 convolutional layers. Compared to the single-frame spatial graph, the size of the spatial-temporal graph increases T times. This limits the ability of the self-attention operators to adaptively construct relationships between joints and neighbors. Therefore, in the multi-head learning module, we first partition the graph into several small groups to limit the number of neighbor nodes for each joint:

$$\mathcal{N}\left(\hat{j}_i\right) = \left\{\hat{j}_j \,|\, d\left(\hat{j}_i, \hat{j}_i\right) \le D\right\} \tag{6}$$

where $\mathcal{N}$ is the neighbor set, $d(\cdot,\cdot)$ denotes the shortest graph distance between two nodes. In this work, we set $D = 3$. Then, the ST-GAT $\mathbf{G}_{sta}$ only weighs edges within groups, thus helps to reserve beneficial local motion patterns and reduce computation costs.

STA block then aggregates the temporal information learnt by S self-attention heads using a temporal convolutional layer. The coarse-grained gait encoding after the B^{th} STA blocks is finally summarized by a global average pooling layer, giving the fine-grained video-wise gait representation f^g.

3.3 Adaptive Fusion Module

When the identities slightly change clothes, appearance remains competitive in visual similarities. Frame-wise appearance feature set is first extracted by a CNN backbone, then aggregated by a spatial max pooling and temporal average pooling to obtain the video-wise appearance embedding f^a (Figure 1).

We finally fuse appearance, shape, and gait embeddings using the Adaptive Fusion Module (AFM) as illustrated in Figure 4. Embeddings are first projected onto a common latent space. Then, they are stacked using concatenation and fed into a convolutional layer which aims to optimize the embeddings in parallel by making them refer to each other. The corresponding weights w_a, w_s and w_g are estimated to adaptively amplify the importance of each embedding for the final person representation f. This is useful since viewpoint changes and occlusions bring different level of semantic information from appearance, gait, and shape for certain input videos.

Our ASGL framework is supervised by the sum of two identification loss functions:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ce} + \lambda_2 \mathcal{L}_{tri} \tag{7}$$

where $\mathcal{L}_{ce}$ is cross-entropy loss, $\mathcal{L}_{tri}$ is pair-wise triplet loss and λ_1, λ_2 are weighting parameters.

4 EXPERIMENTAL SETUP

4.1 Datasets and Evaluation Protocols

Datasets. We validate the performance of our proposed framework on two public VCCRe-ID datasets, VCCR (Han et al., 2023) and CCVID (Gu et al., 2022). **VCCR** contains 4,384 tracklets with 392 identities, with $2 \sim 10$ different suits per identity. **CCVID** contains 2,856 tracklets with 226 identities, with $2 \sim 5$ different suits per identity. In Figure 5, we report a relative comparison in challenges for Re-ID posed by the two datasets by showing samples randomly selected and viewpoint variations. It can be seen that CCVID mimics simplistic Re-ID scenarios

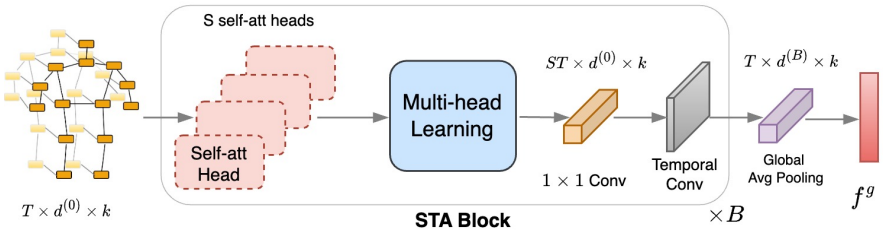

Figure 3: Architecture of the proposed Spatial-Temporal Graph Attention Network for encoding skeleton-based gait.

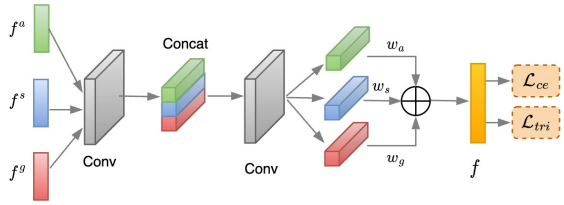

Figure 4: Architecture of the Adaptive Fusion Module.

Figure 5: Samples from VCCR (top) and CCVID (bottom). For VCCR, we randomly collect 3 tracklets from the **same identity** under different clothing. For CCVID, we randomly choose 3 identities with 2 tracklets each under different clothing. VCCR poses realistic challenges for Re-ID like entire clothing changes, viewpoint variations, and occlusions, while CCVID contains only frontal images, clearly visible faces, no occlusion and slight clothing change with identities carrying bags.

such as frontal viewpoints, clearly visible faces, or no occlusion, while VCCR poses complex real-world scenarios for Re-ID. Therefore, we focus on validating the effectiveness of our framework on VCCR.

Evaluation Protocols. Two Re-ID metrics are used to evaluate the effectiveness of our method: Rank-k (R-k) accuracy and mean Average Precision (mAP). We compute testing accuracy in three settings: (1) **Cloth-Changing** (CC), i.e. the test sets contains only cloth-changing ground truth samples; (2) **Standard**, i.e. the test sets contain both cloth-changing and cloth-consistent ground truth samples; and (3) **Same-clothes** (SC), i.e. the test sets contain only cloth-consistent ground truth samples.

4.2 Implementation Details

Model Architecture. We choose Resnet-50 (He et al., 2015) pretrained on ImageNet (Deng et al., 2009) as the CNN backbone for texture branch. For the ASGL branch, we employ MediaPipe (Bazarevsky et al., 2020), an off-the-shelf estimator for 3D pose estimation, which outputs 33 keypoints in world coordinates. As we focus on capturing moving patterns of major joints and do not need features from noses, eyes, fingers, or heels. We average keypoints on face, hand and foot to single keypoints, leaving the skeleton graph with 14 keypoints. The refinement networks consists of 3 fully connected layers of size $[128, 512, 2048]$ respectively. For shape learning sub-branch, the GAT consists of two graph attention layers. For gait learning sub-branch, our ST-GAT consists of 2 STA blocks with their channels in $[128, 256]$. Implementation is in PyTorch (Paszke et al., 2019).

Training and Testing. Input clips for training and testing are formed by randomly sampling 8 frames from each original tracklet with a stride of 2 for VCCR and 4 for CCVID. Frames are resized to 256×128, then horizontal flipping is applied for data augmentation. The batch size is set to 32, each batch randomly selects 8 identities and 4 clips per identity. The model is trained using Adam (Kingma and Ba, 2017) optimizer for 120 epochs. Learning rate is initialized at $5 \times e^{-3}$ and reduced by a factor of 0.1 after every 40 epochs. We set $\lambda_1 = 0.7, \lambda_2 = 0.3$.

5 EXPERIMENTAL RESULTS

5.1 Results on VCCR

A comparison of the quantitative results on VCCR (Han et al., 2023) dataset is reported in in Table 1. State-of-the-art results categorized by method types are presented as benchmark. These include image-based short-term Re-ID (i.e. PCB (Sun et al., 2018)), video-based short-term Re-ID (i.e. AP3D (Gu et al.,

Table 1: Quantitative results on VCCR. ASGL outperforms SOTAs in all evaluation settings by a significant margin.

Method	Method type	Modalities	CC		Standard		SC	
			R-1	mAP	R-1	mAP	R-1	mAP
PCB (Sun et al., 2018)	Image-based	RGB	18.8	15.6	55.6	36.6	-	-
AP3D (Gu et al., 2020)	Video-based	RGB	35.9	31.6	78.0	52.1	-	-
GRL (Liu et al., 2021)	Video-based	RGB	35.7	31.8	76.9	51.4	-	-
SPS (Shu et al., 2021)	Image-based CC	RGB	34.5	30.5	76.5	50.6	-	-
CAL (Gu et al., 2022)	Video-based CC	RGB	36.6	31.9	78.9	52.9	79.1	63.8
3STA (Han et al., 2023)	Video-based CC	RGB + 3D shape	40.7	36.2	80.5	54.3	-	-
ASGL (Ours)	Video-based CC	RGB + shape + gait	**52.9**	**43.2**	**88.1**	**65.8**	**89.9**	**79.5**

Table 2: Comparison of quantitative results on CCVID.

Method	CC		Standard	
	R-1	mAP	R-1	mAP
InsightFace (Deng et al., 2020)	73.5	-	**95.3**	-
ReFace (Arkushin et al., 2022)	**90.5**	-	89.2	-
CAL (Gu et al., 2022)	81.7	79.6	82.6	81.3
DCR-ReID (Cui et al., 2023)	83.6	81.4	84.7	82.7
ASGL (Ours)	83.9	**82.2**	86.1	82.5

2020) and GRL (Liu et al., 2021)), image-based CCRe-ID (i.e. SPS (Shu et al., 2021)) and video-based CCRe-ID (i.e. CAL (Gu et al., 2022) and 3STA (Han et al., 2023)).

Overall, ASGL outperforms previous methods on VCCR in all evaluation protocols. Compared to image-based methods, our method achieves better performance, indicating the importance of spatial-temporal information for Re-ID. The texture-based method CAL (Gu et al., 2022) is outperformed by our framework by 16.3% in rank-1 accuracy and 11.3% in mAP in cloth-changing setting, which shows the effectiveness of ASGL in coupling auxiliary modalities (i.e. shape and gait) with appearance for VCCRe-ID. In same-clothes setting, which mimics the short-term Re-ID environment, ASGL is also superior to CAL, which demonstrates the robustness of ASGL in real-world scenarios. It can be reasoned that severe occlusion and viewpoint changes posed by VCCR hinders the ability of CAL to capture face and hairstyle.

Compared to the 3STA (Han et al., 2023) framework, in cloth-changing setting, ASGL achieves a remarkable improvement of 12.2% in rank-1 and 7.0% in mAP. It can be seen that the combination of shape and gait cues significantly improve the discriminative power of person representations. Moreover, 3STA framework is multi-stage and requires heavy training, shown by a number of 250 training epochs for the first stage and 30000 training epochs for the second stage as reported in (Han et al., 2023). Our framework instead can be trained in an end-to-end manner with only 120 epochs.

5.2 Results on CCVID

On CCVID (Gu et al., 2022) dataset, we report experimental results of our method along with two texture-based methods including CAL (Gu et al., 2022) and DCR-ReID (Cui et al., 2023) in Table 2. It can be seen that the models that focus on extracting appearance features achieve comparable performance on CCVID. Moreover, the results are close to saturation. The limitation of these methods is that they rely heavily on the assumption that input frames contain clearly visible persons. Importantly, CCVID mimics unrealistic Re-ID environment, in which all identities walk towards camera, giving frontal viewpoint, no occlusion. Clothing variations only include carrying a bag or wearing a cap, leading to very slight clothing changes. Therefore, we do not focus on validating the effectiveness of our ASGL framework on CCVID.

5.3 Ablation Study

In ablation study, we validate the effectiveness of: (1) shape and gait embeddings produced by the Attention-based Shape and Gait Learning (ASGL) branch; (2) the proposed GAT in modeling shape and ST-GAT in modeling gait compared to traditional GCN and ST-GCN; and (3) using 3D pose compared to 2D pose.

ASGL Branch. To validate the effectiveness of the various modules in the proposed framework, we carried out training with the following model settings: Texture branch (only appearance embeddings are extracted), ASGL branch (only shape and gait embed-

Table 3: Ablation study of the ASGL branch on VCCR and CCVID.

Method	VCCR				CCVID			
	CC		Standard		CC		Standard	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
Texture branch (Appearance)	32.8	29.3	74.3	46.7	78.5	75.3	79.7	76.9
ASGL branch (Shape and Gait)	29.1	27.4	68.2	43.9	71.3	70.3	72.1	70.8
Appearance and Shape	38.7	32.3	77.2	50.7	79.6	75.6	79.9	77.2
Appearance and Gait	41.5	35.5	79.6	53.1	79.2	76.1	80.6	77.9
Joint (the proposed ASGL)	**52.9**	**43.2**	**88.1**	**65.8**	**83.9**	**82.2**	**86.1**	**82.5**

dings are extracted), and joint representations. The experimental results are reported in Table 3.

It can be observed that the model using only ASGL branch performs worse than that using only Texture branch on VCCR with a gap of 3.7% in rank-1 accuracy and 1.9% in mAP in cloth-changing setting. The reasons is that appearance features are still more competitive in visual similarities than pose-based modalities when identities do not change or slightly change clothes. Moreover, the discriminative power of shape and gait embeddings rely on the accuracy and robustness of the off-the-shelf pose estimator, which may suffer poor estimation results under severe occlusions. Overall, the matching ability of our Re-ID framework is maximized when appearance is coupled with shape and gait embeddings extracted by ASGL branch, shown by a large performance gap of 20.1%/23.8% in rank-1 accuracy and 13.9%/15.8% in mAP between the joint model and the single-branch texture/ASGL models.

We further analyze the contribution of each pose-based cue (i.e. shape and gait) to the discriminability of person representations, in which two models are trained: appearance coupled with shape and appearance coupled with gait. On VCCR, appearance-gait model achieves higher rank-1 accuracy in both evaluation protocols than appearance-shape model, while on CCVID, there is no significant difference in performance between two models. This can be reasoned by the challenges posed by the two datasets. Video tracklets in CCVID only contain frontal walking trajectories (i.e. people walking towards camera) and no occlusion, while VCCR poses greater viewpoint variations and occlusions. Thus, VCCR brings richer gait information and its nature hinders the extraction of fine-grained shape embeddings. It is also worth noting that coupling both shape and gait cues with appearance brings the most discriminative power for Re-ID in real-world scenarios, shown by the results of our proposed ASGL framework.

GAT and ST-GAT. In Table 4, we report the comparison results on VCCR between the two models using GCN (Kipf and Welling, 2017) and ST-GCN (Yan et al., 2018) and our proposed GAT and ST-GAT for shape and gait learning, respectively. By incorporating GAT and ST-GAT, our proposed model improves Re-ID performance in cloth-changing and standard settings by 5.1% and 7.7% in mAP. Unlike traditional GCN or ST-GCN which operate on graphs by treating every node equally, we leverage attention mechanism, which allows for attending to local shape cues and local motion ranges, then amplifying the most important shape features and motion patterns. This helps mitigates the influence of viewpoint changes and occlusions on shape and gait, giving more discriminative final person representations.

Table 4: Ablation study of the proposed GAT for learning shape and ST-GAT for learning gait on VCCR.

Method	CC		Standard	
	R-1	mAP	R-1	mAP
GCN & ST-GCN	45.2	38.1	82.6	58.1
GAT & ST-GAT	**52.9**	**43.2**	**88.1**	**65.8**

Table 5: Ablation study of 3D pose on VCCR.

Method	CC		Standard	
	R-1	mAP	R-1	mAP
ASGL w/ 2D pose	47.2	39.9	84.1	60.3
ASGL w/ 3D pose	**52.9**	**43.2**	**88.1**	**65.8**

3D Pose. In Table 5, we provide an insight on the effectiveness of 3D pose compared to 2D pose in Re-ID. We can observe that using 3D pose leads to an improvement of 3.3%/5.5% in mAP in cloth-changing/standard setting. Compared to 2D pose, 3D pose contains richer spatial information. Moreover, 3D pose is less affected by viewpoint changes, thus temporal dynamics from a person's trajectory can be more accurately captured. The superiority of 3D pose over 2D pose demonstrates the effectiveness of our ASGL framework for VCCRe-ID task.

5.4 Discussion

In Table 2, we also report the performance on CCVID of two models that focus on capturing facial features for Re-ID: (1) the face model InsightFace (Deng et al., 2020); and (2) ReFace (Arkushin et al., 2022). It

can be observed that in standard setting, InsightFace achieves the highest rank-1 accuracy of 95%, showing that only by extracting facial features, Re-ID performance on CCVID is close to saturation. In cloth-changing setting, ReFace outperforms other works. It is worth noting that ReFace is built upon CAL (Gu et al., 2022), which combines clothes-based loss functions with explicit facial feature extraction. In our future research, we would consider coupling biometric cues like faces with other structural cues like body shape and gait for VCCRe-ID task.

6 CONCLUSION

In this paper, we proposed Attention-based Shape and Gait Representations Learning, an end-to-end framework for Video-based Cloth-Changing Person Re-Identification. By extracting body shape and gait cues, we enhance the robustness of Re-ID features under clothing-change situations where appearance is unreliable. We proposed a Spatial-Temporal Graph Attention Network (ST-GAT) that encodes gait embedding from 3D-skeleton-based pose sequence. Our ST-GAT is able to amplify important motion ranges as well as capture beneficial local motion patterns for a discriminative gait representation. We showed that by leveraging 3D pose and attention mechanism in our framework, Re-ID accuracy under confusing clothing variations is significantly improved, compared to using 2D pose and traditional Graph Convolutional Networks. Our framework also effectively deals with viewpoint variations and occlusions, shown by state-of-the-art experimental results on the large-scale VCCR dataset which mimics real-world Re-ID scenarios.

REFERENCES

Arkushin, D., Cohen, B., Peleg, S., and Fried, O. (2022). Reface: Improving clothes-changing re-identification with face features.

Bansal, V., Foresti, G. L., and Martinel, N. (2022). Cloth-changing person re-identification with self-attention. In *Winter Conference on Applications of Computer Vision Workshop*, pages 602–610.

Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., and Grundmann, M. (2020). Blazepose: On-device real-time body pose tracking.

Chen, J., Jiang, X., Wang, F., Zhang, J., Zheng, F., Sun, X., and Zheng, W.-S. (2021). Learning 3d shape feature for texture-insensitive person re-identification. In *CVPR*, pages 8142–8151.

Chen, J., Zheng, W.-S., Yang, Q., Meng, J., Hong, R., and Tian, Q. (2022). Deep shape-aware person re-identification for overcoming moderate clothing changes. *IEEE TMM*, 24:4285–4300.

Cui, Z., Zhou, J., Peng, Y., Zhang, S., and Wang, Y. (2023). Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):4415–4428.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255.

Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S. (2020). Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, pages 5202–5211.

Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., and Chen, X. (2022). Clothes-changing person re-identification with rgb modality only. In *CVPR*, pages 1050–1059.

Gu, X., Chang, H., Ma, B., Zhang, H., and Chen, X. (2020). Appearance-preserving 3d convolution for video-based person re-identification.

Han, K., Huang, Y., Gong, S., Huang, Y., Wang, L., and Tan, T. (2023). 3d shape temporal aggregation for video-based clothing-change person re-identification. In *ACCV*, pages 71–88.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Hong, P., Wu, T., Wu, A., Han, X., and Zheng, W.-S. (2021). Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *CVPR*, pages 10508–10517.

Huang, Y., Xu, J., Wu, Q., Zhong, Y., Zhang, P., and Zhang, Z. (2020). Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3459–3471.

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 *CVPR*, pages 14258–14267.

Khaldi, K., Mantini, P., and Shah, S. K. (2022). Unsupervised person re-identification based on skeleton joints using graph convolutional networks. In *Image Analysis and Processing*, pages 135–146.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization.

Kipf, T. N. and Welling, M. (2017). Semi-supervised classification with graph convolutional networks.

Li, J., Zhang, S., and Huang, T. (2018a). Multi-scale 3d convolution network for video based person re-identification.

Li, M., Zhu, X., and Gong, S. (2018b). Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, pages 772–788.

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159.

Li, Y.-J., Weng, X., and Kitani, K. M. (2021). Learning shape representations for person re-identification un-

der clothing change. In *Winter Conference on Applications of Computer Vision*, pages 2431–2440.

Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015). Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206.

Liu, C.-T., Wu, C.-W., Wang, Y.-C. F., and Chien, S.-Y. (2019). Spatially and temporally efficient non-local attention network for video-based person re-identification.

Liu, X., Zhang, P., Yu, C., Lu, H., and Yang, X. (2021). Watching you: Global-guided reciprocal learning for video-based person re-identification. In *CVPR*, pages 13329–13338.

Luo, H., Gu, Y., Liao, X., Lai, S., and Jiang, W. (2019). Bag of tricks and a strong baseline for deep person re-identification.

Ma, L., Yang, X., and Tao, D. (2014). Person re-identification over camera networks using multi-task distance metric learning. *IEEE TIP*, 23(8):3656–3670.

Matsukawa, T., Okabe, T., Suzuki, E., and Sato, Y. (2016). Hierarchical gaussian descriptor for person re-identification. In *CVPR*, pages 1363–1372.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library.

Qian, X., Wang, W., Zhang, L., Zhu, F., Fu, Y., Xiang, T., Jiang, Y.-G., and Xue, X. (2020). Long-term cloth-changing person re-identification.

Shu, X., Li, G., Wang, X., Ruan, W., and Tian, Q. (2021). Semantic-guided pixel sampling for cloth-changing person re-identification. *IEEE Signal Processing Letters*, 28:1365–1369.

Sun, Y., Zheng, L., Yang, Y., Tian, Q., and Wang, S. (2018). Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, page 501–518.

Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., and Rigoll, G. (2021). GaitGraph: Graph convolutional network for skeleton-based gait recognition. In *ICIP*, pages 2314–2318.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al. (2017). Graph attention networks. *stat*, 1050(20):10–48550.

Wan, F., Wu, Y., Qian, X., Chen, Y., and Fu, Y. (2020). When person re-identification meets changing clothes. In *CVPRW*, pages 3620–3628.

Wang, G., Yuan, Y., Chen, X., Li, J., and Zhou, X. (2018). Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*.

Wang, L., Zhang, X., Han, R., Yang, J., Li, X., Feng, W., and Wang, S. (2022). A benchmark of video-based clothes-changing person re-identification.

Wu, Y., Bourahla, O. E. F., Li, X., Wu, F., Tian, Q., and Zhou, X. (2020). Adaptive graph representation learning for video person re-identification. *IEEE TIP*, 29:8821–8830.

Yan, S., Xiong, Y., and Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition.

Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., and Yang, X. (2016). Person re-identification via recurrent feature aggregation. In *ECCV*, pages 701–716.

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 *CVPR*, pages 3289–3299.

Yang, Q., Wu, A., and Zheng, W.-S. (2021). Person re-identification by contour sketch under moderate clothing change. *IEEE TPAMI*, 43(6):2029–2046.

Zhang, P., Wu, Q., Xu, J., and Zhang, J. (2018). Long-term person re-identification using true motion from videos. In *Winter Conference on Applications of Computer Vision*, pages 494–502.

Zhang, P., Xu, J., Wu, Q., Huang, Y., and Ben, X. (2021). Learning spatial-temporal representations over walking tracklet for long-term person re-identification in the wild. *IEEE TMM*, 23:3562–3576.

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 *ECCV*, page 868–884.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124.

Zheng, Z., Wang, X., Zheng, N., and Yang, Y. (2022). Parameter-efficient person re-identification in the 3d space. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14.

Zhou, Z., Huang, Y., Wang, W., Wang, L., and Tan, T. (2017). See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, pages 6776–6785.