Temporal 3D Shape Modeling for Video-based Cloth-Changing Person Re-Identification

Vuong D. Nguyen   Pranav Mantini   Shishir K. Shah
Quantitative Imaging Lab, Dept. of Computer Science, University of Houston
dnguyen170@uh.edu   pmantini@cs.uh.edu   sshah@central.uh.edu

Abstract

Video-based Cloth-Changing Person Re-ID (VCCRe-ID) refers to a real-world Re-ID problem where texture information like appearance or clothing becomes unreliable in long-term, limiting the applicability of traditional Re-ID methods. VCCRe-ID has not been well studied primarily due to (1) limited public datasets and (2) challenges related to extracting identity-related clothes-invariant cues from videos. Few existing works have heavily focused on gait-based features, which are severely affected under viewpoint changes and occlusions. In this work, we propose "Temporal 3D ShapE Modeling for VCCRe-ID" (SEMI), a lightweight end-to-end framework that addresses these issues by learning human 3D shape representations. The SEMI framework comprises of a Temporal 3D Shape Modeling branch, which extracts discriminative frame-wise 3D shape features using a temporal encoder, and an identity-aware 3D regressor. This is followed by a novel Attention-based Shape Aggregation (ASA) module that effectively aggregates frame-wise shape features for a fine-grained video-wise shape embedding. ASA leverages an attention mechanism to amplify the contribution of the most important frames while reducing redundancy during the aggregation process. Experiments on two VCCRe-ID datasets demonstrate that our proposed framework outperforms state-of-the-art methods by 10.7% *in rank-1 accuracy and* 7.4% *in mAP in cloth-changing setting.*

1. Introduction

Video-based Person Re-Identification (Re-ID) has become increasingly important in various applications, including video surveillance and forensic analysis. It aims to match individuals across different camera views. Traditional Re-ID methods [14, 22, 39] have primarily relied on Convolutional Neural Networks (CNNs) to extract texture information for Re-ID. Later works [19, 41, 44] have leveraged Graph Convolutional Networks (GCNs) to capture high-level spatio-temporal features for person representations. However, these approaches are unreliable in long-term scenarios as they heavily rely on appearance or visual similarities of body parts, which suffers severe degradation in performance under texture-confusing situations like clothing changes. Such shortcomings necessitate a robust long-term approach for the real-world Video-based Cloth-Changing Person Re-ID (VCCRe-ID) problem.

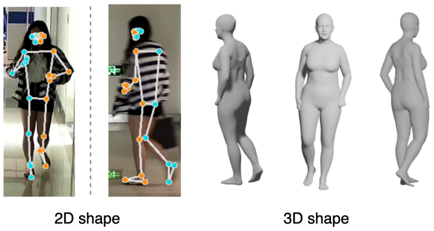


Figure 1. Visualization of (1): 2D skeleton-based shape under different viewpoints (same identity), and (2): 3D shapes generated from the image on the right by our proposed framework. While viewpoint changes result in highly dissimilar 2D shapes, 3D shape is more stable in long-term and invariant to viewpoints.

To address challenges of the general CCRe-ID problem, texture-based methods [8, 13] have proposed to attend to clothes-irrelevant features like face and hairstyle, which are likely to fail under occlusions where these cues are unobservable. 2D human geometric modalities such as 2D shape [3, 25, 34], sketches [6, 45], silhouettes [17], or gait [46, 47] have also been explored. However, as illustrated in Figure 1, viewpoint changes significantly limit the discriminative power of these 2D features for Re-ID.

In this work, we overcome the limitations of texture-based and 2D-based methods by utilizing human 3D geometric cues for VCCRe-ID. The motivations are two-fold. First, 3D geometric cues focus on the underlying structure of the person, which tends to remain stable in long-term, making them more invariant to clothing changes. Second, depth information in 3D space not only enhances robustness of features to viewpoint changes (Figure 1) but also captures with high fidelity the spatio-temporal relationships between body parts, leading to a more discriminative person representation [40].

Some pioneering works has resulted in the collection of 3D-based source data from kinect cameras [40] or radio signals [11]. Nonetheless, such sensing is costly and not feasible in real-world environments. Later methods [4, 6, 38, 51] have leveraged off-the-shelf 3D human estimation models to extract human 3D structure from video data. However, these off-the-shelf 3D models [18, 21] are designed to provide a rough estimate of the overall human body based on parametric models without identity-aware regularization, which fails to capture fine-grained details unique to individuals for Re-ID. Han *et al.* [15] proposed 3STA framework, in which the first stage learns a frame-wise 3D shape generator using a 3D human dataset for regularization while in the second stage, shape features are extracted and then aggregated. This framework is multi-stage and requires auxiliary large-scale datasets, resulting in complicated training.

To this end, we propose "Temporal 3D **S**hap**E** **M**odeling for VCCRe-**ID**" (SEMI) framework. To the best of our knowledge, we are the first to address VCCRe-ID using human 3D shape cues in an end-to-end manner. The overview of our framework is shown in Figure 2. To learn 3D shape representations under clothing changes, SEMI comprises of the Temporal 3D Shape Modeling (TSM) branch. TSM first captures the temporal dynamics of persons across video frames using a temporal encoder built on Gated Recurrent Units (GRUs). Then, the 3D regressor estimates frame-wise 3D shape parameters based on the Skinned Multi-person Linear (SMPL) model [28] using iterative regularization. To mitigate the tendency of generating neutral 3D shape, the regressor is guided to enhance the discriminability of 3D shape of different persons using an identification loss. To aggregate frame-wise shape features for a robust video-wise shape embedding, a novel Attention-based Shape Aggregation (ASA) module is introduced. ASA leverages GRUs and an attention mechanism to emphasize the most important frames and reduce the influence of noisy frames. ASA also minimizes redundancy of information brought by similar consecutive frames. Shape features are finally coupled with appearance features for the final person representation.

The key contributions of our work can be summarized as follows: (1) we propose a novel end-to-end framework that integrates identity-aware temporal 3D shape representation learning for VCCRe-ID; (2) we introduce an Attention-based Shape Aggregation module that effectively aggregates shape information to obtain robust shape embeddings for Re-ID from video; and (3) we report results on two large-scale VCCRe-ID datasets, demonstrating the superiority of our proposed SEMI framework over state-of-the-art methods by a significant margin in all evaluation settings which mimic real-world scenarios for person Re-ID.

Dataset	#IDs	#Videos	#Suits/ID	Public
Motion-ReID [46]	30	240	-	✗
CVID-reID [47]	90	2980	-	✗
SCCVRe-ID [38]	333	9620	$2 \sim 37$	✗
RCCVRe-ID [38]	34	6948	$2 \sim 10$	✗
CCVID [13]	226	2856	$2 \sim 5$	✓
VCCR† [15]	392	4384	$2 \sim 10$	✓

Table 1. A summary of existing Video-based Cloth-Changing Person Re-ID datasets. † means the dataset contains distractor identities who do not change clothes.

2. Related Works

2.1. Person Re-ID

Early methods for image-based person Re-ID involved feature representation learning [31, 37] and distance metric learning [26, 30], while more recent methods have adopted deep learning [29, 36]. For video-based person Re-ID, spatio-temporal information has been exploited by using 3D-CNN [14, 22], RNN-LSTM [43, 52] or attention mechanisms [12, 48]. Graph Convolutional Networks (GCNs) have also been applied [41, 42, 44] to promote video-wise person representations. These methods have achieved notable results on standard image-based [24, 50] and video-based [23, 49] Re-ID benchmarks. However, their applicability in real-world scenarios can be limiting for two reasons: (1) the standard benchmarks datasets have been collected over short-term, thus presenting an impractical consistency in clothing that would not hold true for longer duration Re-ID; and (2) these models rely heavily on appearance features to identify the persons, which is also unreliable in long-term scenarios.

2.2. Image-based Cloth-Changing Person Re-ID

Thanks to recently published datasets [34, 45], several approaches have been proposed to address Image-based Cloth-Changing Person Re-ID. For texture-based methods, Gu *et al.* [13] proposed CAL, which extracts clothes-irrelevant features like face and hairstyle by using clothes-based adversarial loss. Cui *et al.* [8] disentangled clothes-irrelevant features based on the reconstruction of human component regions. For shape-based methods, Qian *et al.* [34] proposed to extract 2D shape using a cloth-elimination shape-distillation module. Li *et al.* [25] leveraged adversarial learning to capture shape from RGB images and augmented gray images. Since 2D shape is severely affected by viewpoint changes and occlusions, recent works have utilized human 3D geometric cues. Chen *et al.* [5] proposed a framework built on HMR [18] to estimate and regularize 3D shape parameters. Zheng *et al.* [51] leveraged 3D mesh to construct a KNN graph, and then used graph convolution

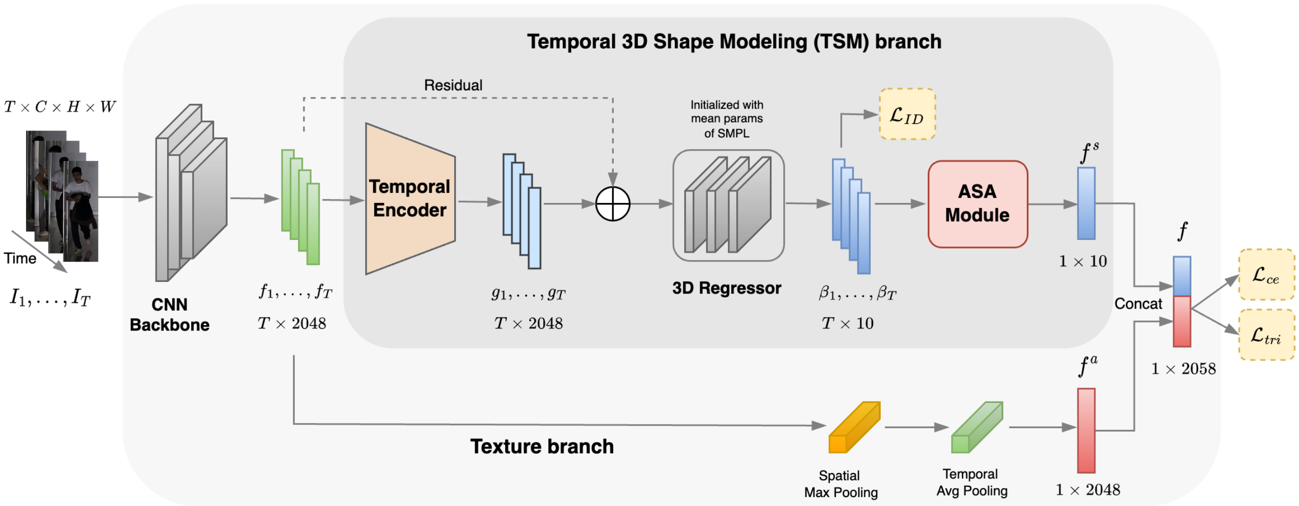


Figure 2. **Overview of our SEMI framework.** A video sequence of T frames is passed through a CNN backbone to obtain global feature set F. Given F, TSM branch outputs a video-wise shape feature vector f^s, while texture branch outputs a video-wise appearance feature vector f^a. f^s and f^a are then concatenated for final person representation.

to exploit shape features. These methods lack the modeling of temporal dynamics of persons over video sequences, thus limiting the information that can be extracted to assist VCCRe-ID.

2.3. Video-based Cloth-Changing Person Re-ID

Video-based Cloth-Changing Person Re-ID (VCCRe-ID) has received little research attention with few public datasets. We report a summary of existing VCCRe-ID datasets in Table 1. Zhang *et al.* [46] used trajectory-aligned feature descriptors to encode motion features based on the assumption that persons keep constant walking patterns. SpTSkM is proposed in [47] to learn skeleton-based gait cues from using GCNs. Wang *et al.* [38] enhanced gait learning by proposing a confidence-guided re-ranking strategy. These works attempt to extract motion patterns as features for Re-ID. However, viewpoints changes limit the capturing of walking patterns and occlusions can hinder the modeling of body parts movement, making gait-based features ambiguous for Re-ID. Han *et al.* [15] proposed a multi-stage framework in which the first stage generates 3D shape using an auxiliary 3D human dataset for regularization, which requires heavy training. In this paper, we propose an end-to-end framework that effectively learns and aggregates 3D shape features over video sequences, producing robust person representations for Re-ID.

3. The Proposed Framework

3.1. Overview

The overview of our proposed SEMI framework is shown in Figure 2. Given a video sequence of T frames $X = \{I_i\}_{i=1}^{T}$, a CNN backbone $\mathcal{F}(\cdot) : \mathbb{R}^{C\times H\times W} \rightarrow \mathbb{R}^{d_f}$ first extracts global frame-wise feature set $F = \{f_i\}_{i=1}^{T}$. Then, for Temporal 3D Shape Modeling (TSM) branch, a temporal encoder $\mathcal{G}(\cdot)$ takes F as input and outputs latent frame-wise feature set $G = \{g_i\}_{i=1}^{T}$. We then apply residual connection, where $H = \{f_i + g_i\}_{i=1}^{T}$ is used as input to a 3D regressor $\mathcal{R}(\cdot)$ to yield frame-wise 3D shape parameters set $\beta = \{\beta_i\}_{i=1}^{T}$. Attention-based Shape Aggregation module then aggregates β to obtain a video-wise shape representation f^s. Meanwhile, texture branch summarizes appearance information over frames from F and outputs a video-wise appearance representation f^a. We finally concatenate f^s and f^a for the final person representation.

3.2. Temporal 3D Shape Modeling branch

3D human body can be encoded as a function of pose and shape. In this work, we only leverage shape since pose may not necessarily be unique to individuals and thus not discriminative for Re-ID. Given global feature set F, we learn robust video-wise 3D shape representation of a person using the Temporal 3D Shape Modeling (TSM) branch, which comprises a temporal encoder, a 3D regressor and a shape aggregation module as illustrated in Figure 2.

Temporal Encoder. The global frame-wise feature set F only provides visual representations of each frame and lacks temporal information. The temporal evolution of a person over time is crucial for VCCRe-ID since informative frames can compliment the ambiguity of appearance and shape in other frames where the body is partially occluded or affected by viewpoint changes. Therefore, we use a temporal encoder to capture the temporal dynamics present in the input video sequence. Instead of using traditional Recurrent Neural Networks, the temporal encoder $\mathcal{G} : \mathbb{R}^{d_f} \rightarrow \mathbb{R}^{d_g}$ consists of several Gated Recurrent Units [7] layers that can selectively capture longer-term dependencies across frames. Given global feature vector f_i of frame I_i, $\mathcal{G}$ yields latent feature vector $g_i = \mathcal{G}(f_i), g_i \in \mathbb{R}^{d_g}$ based on the information from previous frames $\{I_1, .., I_{i-1}\}$.

3D Human Parametric Modeling. In this work, we estimate 3D human body shape based on the Skinned Multi-Person Linear (SMPL) [28] model. SMPL is first assigned to a mean shape in the standard zero pose, which serves as a template to capture shape through the variations in height, weight, and body proportions. Then, SMPL represents shape by formulating the shape displacements from the template shape via a linear combination of K coefficients of a PCA shape space:

$$B_S(\beta_i) = \sum_{k=1}^{K} \beta_i^k S \tag{1}$$

where $\beta_i = \left[\beta_i^1, \ldots, \beta_i^K\right]$ are the shape parameters estimated from frame I_i of input sequence, $B_S(\cdot)$ denotes the blending shape function, $S \in \mathbb{R}^{3V}$ denotes the orthonormal principal components of shape offsets, V is the number of vertices on the mesh. Following [33], in this work, K is set to 10, which sufficiently represents body shape. Each shape parameter $\beta^k, k = 1, ..., 10$ controls certain aspect of body shape such as body length, hip size, etc.

Identity-aware 3D Regressor. The goal of the regressor $\mathcal{R}(\cdot) : \mathbb{R}^{d_g} \rightarrow \mathbb{R}^{10}$ is to output frame-wise 3D shape parameters $\beta_i = [\beta_i^1, \ldots, \beta_i^{10}]$ from image encodings. We apply a residual connection, where we sum up latent encoding g_i with the global feature f_i as $h_i = f_i + g_i$ to be input to the regressor, i.e. $\beta_i = \mathcal{R}(h_i)$. By doing this, we enable the regressor to preserve important high-level information from the input features while leveraging the temporal dynamics. The regressor comprises of several fully connected layers with dropout in between every two layers. The final layer is responsible for decomposing shape parameters from high-dimensional body encoding. Note that we need to ensure the validity of generated shape parameters so they can represent true body shape. Therefore, to provide the regressor with prior knowledge about the neutral human body shape, we assign the mean shape parameters $\overline{\beta}$ from SMPL [28] model to the regressor. Then, the regressor is biased towards generating realistic human body shape by aligning estimated shape with the average shape using MSE loss $\mathcal{L}^s_{val}$ formulated as:

$$\mathcal{L}^S_{val} = \sum^{N} \left(\frac{1}{T} \sum_{i=1}^{T} \left(\beta_i - \overline{\beta}_i\right)^2 \right) \tag{2}$$

where T is the sequence length and N is the batch size.

Direct estimation only yields coarse-grained shape parameters with limited discriminability. To facilitate the Re-ID, we need to ensure a small intra-class gap and large inter-class gap of shape parameters in the latent space. Therefore, we further supervise the regressor using an identification loss $\mathcal{L}^S_{ID}$, which is a cross-entropy-based classification loss given the number of classes (i.e. identities) in the training set. By optimizing the TSM branch with $\mathcal{L}^S = \mathcal{L}^S_{val} + \mathcal{L}^S_{ID}$, we force the framework to satisfy the validity and enhance the discriminative power of estimated shape parameters for more accurate re-identification.

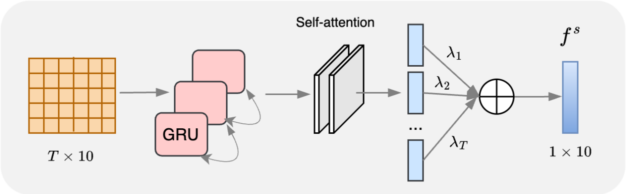


Figure 3. Architecture of ASA module, which comprises GRUs layers and self-attention layers to output video-wise shape representation f^s from frame-wise shape parameters β.

Attention-based Shape Aggregation module. To aggregate frame-wise shape parameter set $\beta = \{\beta_i\}_{i=1}^T$ for a video-wise shape embedding f^s, using traditional aggregation methods such as max or average pooling suffers from two limitations: (1) temporal relationships between consecutive frames are ignored since each frame is treated independently; and (2) all frames are treated equally, leading to potential loss of discriminative power in the aggregated shape representation since not every frame is equally informative due to viewpoint changes or occlusions.

In this work, we propose the Attention-based Shape Aggregation (ASA) module illustrated in Figure 3. To address (1), we first feed $\beta \in \mathbb{R}^{T \times 10}$ into a multi-layer GRUs, which estimates a latent code $\hat{\beta}_i$ at each time step i (i.e. frame i^{th} in the sequence). The GRU layers help capture the fine-grained variations in shape across frames. Then, we address (2) by using an attention mechanism [2] to consider the importance of each frame to the video-wise shape representation. Specifically, a sequence of self-attention layers are used to aggregate hidden states $\left[\hat{\beta}_1, ..., \hat{\beta}_T\right]$. Formally:

$$f^s = \sum_{i=1}^{T} \lambda_i \hat{\beta}_i, \tag{3}$$

where λ_i is the weight assigned to $\hat{\beta}_i$. We adopt a MLP layer $\varphi(\cdot)$ to learn $\theta_i, i = 1, ..., T$, which is then normalized using softmax to obtain $\lambda_i, i = 1, ..., T$, given as:

$$\theta_i = \varphi\left(\hat{\beta}_i\right), \quad \lambda_i = \frac{e^{\theta_i}}{\sum_{j=1}^{T} e^{\theta_j}}. \tag{4}$$

By doing this, ASA module is able to amplify the contribution of most important frames to the aggregated shape representation while reducing the influence of noisy frames caused by occlusions or viewpoint changes. Moreover, ASA module reduces information redundancy brought by sequences of consecutive frames, making the model more lightweight and robust.

3.3. Texture branch

Appearance remains a competitive cue for Re-ID in the cases of slight clothing change. Thus, we couple 3D shape with appearance for a more discriminative global person representation. Given frame-wise feature set $F = \{f_i\}_{i=1}^{T}$, following [14], we first use spatial max pooling and then temporal average pooling to obtain the video-wise appearance embedding f^a (Figure 2). We then simply concatenate appearance and shape embeddings to obtain the final video-wise person representation:

$$f = [f^a, f^s]. \tag{5}$$

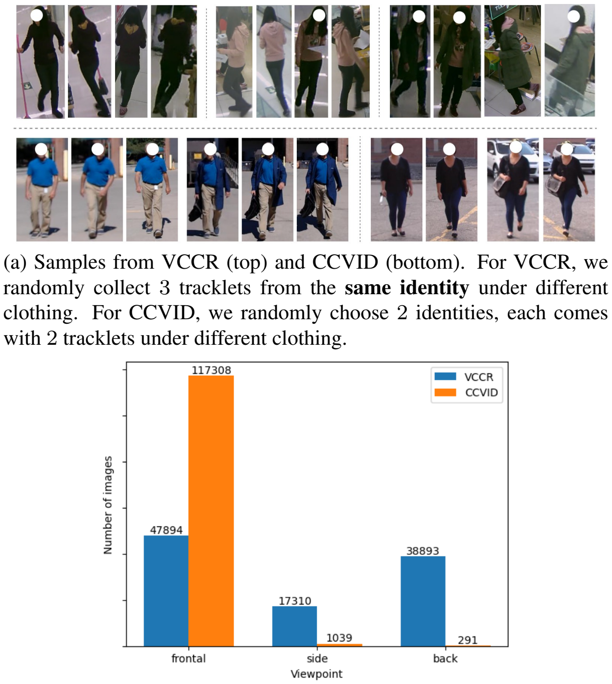
Finally, we feed f into a cross-entropy loss $\mathcal{L}_{ce}$ and a triplet loss $\mathcal{L}_{tri}$, giving the total loss for training our framework, given as:

$$\mathcal{L}_{total} = \mathcal{L}^S + \mathcal{L}_{ce} + \mathcal{L}_{tri}. \tag{6}$$

4. Experimental setup

4.1. Datasets and Evaluation Protocols

Datasets. Two public VCCRe-ID datasets, VCCR [15] and CCVID [13] are used for experiments. **VCCR** contains $4,384$ tracklets with 392 identities. Each identity wears $2 \sim 10$ different suits (with an average of 3.3). $2,873$ tracklets of 150 cloth-changing identities make up the training set. The query set contains 496 tracklets of 82 cloth-changing identities, and the gallery set contains the remaining 718 tracklets of these 82 identities along with 297 tracklets of 160 distractors. **CCVID** contains $2,856$ tracklets with 226 identities. Each identity has $2 \sim 5$ different suits. No distractors are present in CCVID. The training set contains 968 tracklets of 75 identities, while 834 tracklets are used as query set, and the remaining 1074 tracklets build the gallery set. In Figure 4, 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 unrealistic Re-ID scenarios such as frontal viewpoints, clearly visible faces, or no occlusion, while VCCR poses real-world challenges for Re-ID. Therefore, in this work, we *focus on validating the effectiveness of our framework on VCCR*.



(a) 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 2 identities, each comes with 2 tracklets under different clothing.

(b) Comparison in viewpoint variations.

Figure 4. Comparison between VCCR and CCVID. VCCR poses realistic challenges for Re-ID like clothing changes, viewpoint variations, and occlusions, while CCVID contains only frontal images, showing no occlusion and slight clothing changes. (Best viewed in color).

Evaluation protocols. Rank-k accuracy and mean average precision (mAP) are used to evaluate the performance of our method. 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

Architecture. We adopted Resnet-50 [16] pretrained on ImageNet [9] as the CNN backbone $\mathcal{F}$, which outputs $f_i \in \mathbb{R}^{2048}$ for each frame I_i. For the Temporal Shape Estimation module, the temporal encoder $\mathcal{G}$ consists of 2 GRUs layers with 1024 neurons each, followed by a linear projection layer, which produces $g_i \in \mathbb{R}^{2048}$. The 3D regressor $\mathcal{R}$ consists of two 1024 fully-connected layers with a dropout layer in between, followed by a final layer that out-

Method	Method type	Modalities	CC			Standard			SC		
			R-1	R-5	mAP	R-1	R-5	mAP	R-1	R-5	mAP
PCB [36]	Image-based	RGB	18.8	38.6	15.6	55.6	75.2	36.6	-	-	-
AP3D [14]	Video-based	RGB	35.9	55.8	31.6	78.0	88.4	52.1	-	-	-
GRL [27]	Video-based	RGB	35.7	55.3	31.8	76.9	88.2	51.4	-	-	-
SPS [35]	Image-based CC	RGB	34.5	54.1	30.5	76.5	85.5	50.6	-	-	-
CAL [13]	Video-based CC	RGB	36.6	56.1	31.9	78.9	89.2	52.9	79.1	89.8	63.8
3STA [15]	Video-based CC	RGB + 3D shape	40.7	58.7	36.2	80.5	90.2	54.3	-	-	-
SEMI (Ours)	Video-based CC	RGB + 3D shape	**51.4**	**71.5**	**43.6**	**86.2**	**92.2**	**65.2**	**90.6**	**96.0**	**81.8**

Table 2. Comparison of quantitative results on VCCR. SEMI outperforms SOTAs by a significant margin in all evaluation settings.

Method	CC		Standard	
	R-1	mAP	R-1	mAP
InsightFace [10]	73.5	-	**95.3**	-
CAL [13]	81.7	79.6	82.6	81.3
ReFace (CAL + Face) [1]	**90.5**	-	89.2	-
DCR-ReID [8]	83.6	81.4	84.7	**82.7**
SEMI (Ours)	82.5	**81.9**	83.1	81.8

Table 3. Comparison of quantitative results on CCVID.

puts shape features $\beta_i \in \mathbb{R}^{10}$. $\mathcal{R}$ it is initialized with pretrained weights from SPIN [21]. For the Attention-based Shape Aggregation module, to output video-wise shape embedding $f^s \in \mathbb{R}^{10}$, we first employ 2-layer GRUs of size 1024, followed by a self-attention mechanism with 2 MLP layers of size 1024. Texture branch performs 2 steps of spatial max pooling and temporal average pooling to output video-wise appearance embedding $f^a \in \mathbb{R}^{2048}$. Embeddings f^s and f^a are concatenated for final person representation $f_i \in \mathbb{R}^{2058}$.

Training and Testing. To form input clips for training, 8 frames are randomly sampled from each original tracklet with a stride of 2 for VCCR and 4 for CCVID. We first resized each image in the clip to 256 × 128, then applied horizontal flipping for data augmentation following [14]. The batch size is set to 16 due to GPU memory limit. We randomly select 4 identities and 4 clips per identity for each batch. The model was trained for 120 epochs using Adam [20] optimizer. Learning rate is initialized to $5e^{-3}$ and reduced by a factor of 0.1 after every 40 epochs. We trained SEMI in an end-to-end manner on a single NVIDIA GeForce GTX 1080 16GB RAM GPU, which took around 6 hours. In testing stage, we applied the same sampling strategy on both datasets to form 8-frame input clips. The input clip is passed to the trained CNN backbone only, then the output frame-wise feature set is averaged as the video-wise person representation, which is then used to compute pair-wise similarities for matching stage. Implementation is done in PyTorch [32].

5. Results

5.1. Quantitative results on VCCR

We report the quantitative results on VCCR [15] dataset in Table 2. We compare our SEMI framework with current state-of-the-art approaches (SOTAs) categorized by method types, including image-based short-term Re-ID (i.e. PCB [36]), video-based short-term Re-ID (i.e. AP3D [14] and GRL [27], image-based CCRe-ID (i.e. SPS [35]) and video-based CCRe-ID (i.e. CAL [13] and 3STA [15]). Overall, SEMI outperforms previous methods on VCCR in all evaluation settings. Specifically, in cloth-changing setting, SEMI achieves a remarkable improvement of 10.7% in rank-1, 12.8% in rank-5 and 7.4% in mAP compared to 3STA [15], which is the closest approach to ours. The multi-stage 3STA framework requires heavy and complicated training processes, shown by the number of training epochs for the first stage to be 250 and training epochs for the second stage to be 30000 as reported in [15]. Our framework instead can be trained in an end-to-end manner with only 120 epochs.

Compared to texture-based approach, CAL [13], SEMI outperforms CAL in cloth-changing and standard settings without the need for an additional clothes classifier and clothes-based losses as CAL. This shows the effectiveness of coupling 3D shape with appearance for VCCRe-ID. In same-clothes setting, which mimics a short-term Re-ID dataset, SEMI performs clearly better than CAL. The reason lies in the occlusion and viewpoint changes posed by VCCR, which hinders the ability of CAL to capture features from face or hairstyle. Results on VCCR demonstrate the robustness of SEMI in real-world scenarios.

5.2. Quantitative results on CCVID

Quantitative results on CCVID [13] dataset are reported in Table 3, where we compare our framework with InsightFace [10], CAL [13], ReFace [1] and DCR-ReID [8]. The face model InsightFace [10] achieved the highest rank-1 accuracy in standard setting, while in cloth-changing setting, ReFace [1], that simultaneously extracts face and clothes-irrelevant features outperformed other methods. This is due to the nature of CCVID which represents unrealistic situations in Re-ID such as frontal viewpoint, clearly visible faces, no occlusion, slight clothing changes, and high-quality cropped bounding boxes as shown in Figure

Method	VCCR				CCVID			
	CC		Standard		CC		Standard	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
Appearance (Resnet50 [16])	32.8	29.3	74.3	46.7	78.5	75.3	79.7	76.9
Shape (by HMR [18])	20.4	19.5	59.1	36.6	69.1	64.4	67.2	61.3
Shape (by 3STA [15])	21.3	20.6	62.8	39.2	-	-	-	-
Shape (by proposed TSM)	24.7	22.3	65.2	40.7	69.6	65.2	68.1	61.7
Joint (Shape by HMR [18])	39.5	35.1	79.1	52.6	78.7	75.4	80.1	76.9
Joint (3STA [15])	40.7	36.2	80.5	54.3	-	-	-	-
Joint (SEMI) w/o. $\mathcal{L}_{ID}^{S}$	50.1	42.3	84.0	63.6	80.6	79.8	81.1	79.9
Joint (SEMI) w/. $\mathcal{L}_{ID}^{S}$	**51.4**	**43.6**	**86.2**	**65.2**	**82.5**	**81.9**	**83.1**	**81.8**

Table 4. Ablation study on the effectiveness of 3D shape features produced by our Temporal 3D Shape Modeling (TSM) branch with and without identity-guidance loss $\mathcal{L}_{ID}^{S}$ compared to HMR [18] and 3STA [15] on VCCR and CCVID. Our 3D shape modeling method outperforms the SOTAs by a large margin.

4. While these settings lend to limited challenges, we see that our framework still achieves comparable results to CAL [13], showing that coupling our estimated 3D shape with global texture information is more competitive and robust for Re-ID than solely relying on clothes-irrelevant features like face and hairstyle.

6. Ablation study

We perform an ablation study for the proposed framework to validate the effectiveness of: (1) coupling 3D shape with appearance for Re-ID, (2) the Temporal 3D Shape Modeling (TSM) branch, (3) the Attention-based Shape Aggregation module, and (4) concatenation of representations for appearance and shape fusion.

Appearance vs Shape vs Joint. In Table 4, we report experimental results that we carried out with three model settings: appearance, shape and joint representations. The model in appearance setting only comprises the texture branch with Resnet-50 [16] backbone. In shape setting, only shape features are used as person representations. It can be observed that shape models perform worse than appearance model on both VCCR and CCVID. The reasons can be two-fold. First, 3D shape features are only 10-dimensional, which limits the ability to model a global person representation. Second, when the identities do not change or slightly change clothes, exploiting visual similarities from appearance remains more competitive than 3D shape features. The large performance gap on VCCR between the joint models and the individual representation models demonstrate the effectiveness of 3D shape when coupled with appearance in real-world scenarios. Appearance and shape can bring richer information by complementing each other. This facilitates the Re-ID model in both cloth-consistent and cloth-changing environment.

Temporal 3D Shape Modeling (TSM) branch. The effectiveness of 3D shape features produced by our proposed TSM branch is compared with the off-the-self 3D human estimation model HMR [18] and 3STA framework [15] in Table 4. Due to the presence of a broader range of challenges being represented in the VCCR dataset, we focus on comparing the methods on VCCR. TSM outperforms HMR and 3STA in both shape and joint model settings. For example, the rank-1 accuracy in cloth-changing/standard setting, the shape model with TSM achieves $4.3\%/6.1\%$ higher than the shape model with HMR, while our joint SEMI framework achieves $10.7\%/5.7\%$ higher than the joint 3STA framework. Note that same as our SEMI framework, 3STA also couples shape with appearance features extracted using Resnet-50 [16] backbone. For HMR, it roughly estimates 3D shape for each frame without modeling the temporal information and identity-aware regularization. For 3STA, the second shape extraction stage relies heavily on the quality of generated pseudo shape labels in the first stage, which needs auxiliary datasets for regularization. We also enhance the discriminative power of shape features using the identification loss $\mathcal{L}_{ID}^{S}$, shown by higher Re-ID accuracy on both datasets. Besides quantitative comparison, in Figure 5, we visualize the distribution on latent space of 3D shape features estimated by the three methods from 3 videos of 3 identities sampled from VCCR. It can be seen that framewise shape features produced by our proposed TSM are more separable. This further demonstrates the validity and

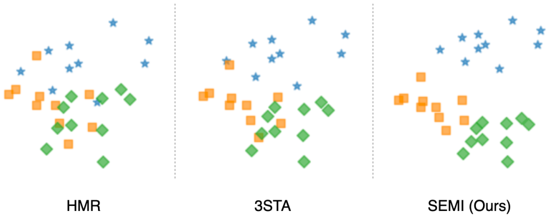


Figure 5. t-SNE visualization of distribution on latent space of frame-wise 3D shape features estimated by HMR [18], 3STA [15], and our proposed TSM. We randomly sample 3 clips of 3 different identities from VCCR, each clip is 10-frame long.

Method	CC		Standard	
	R-1	mAP	R-1	mAP
TSM w/ Average Agg	40.1	33.6	78.4	51.7
TSM w/ DSA [15]	47.1	39.2	83.5	61.1
TSM w/ ASA (Ours)	**51.4**	**43.6**	**86.2**	**65.2**

Table 5. Ablation study of the ASA module on VCCR.

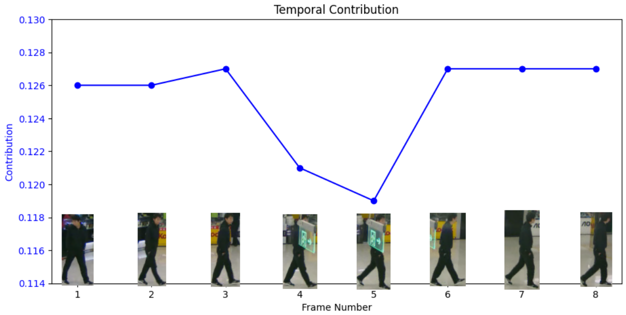


Figure 6. Visualization of contributions of frame-wise shape features to the video-wise shape embedding, represented by attention scores $\lambda_i, i = 1, ..., 8$ learnt by our proposed ASA module. Informative frames are emphasized, while the influence of occluded frames 4 and 5 is effectively reduced.

effectiveness of our proposed identity-guidance 3D shape modeling method.

Attention-based Shape Aggregation module. To validate the effectiveness of our proposed Attention-based Shape Aggregation (ASA) module, in Table 5, we report the results on VCCR of our SEMI framework using three different aggregation methods: traditional averaging, Difference-aware Shape Aggregation (DSA) [15], and ASA, respectively. It can be seen that ASA makes a significant improvement in both evaluation settings compared to averaging and DSA. This is because unlike our ASA module with GRUs, averaging and DSA lack the implicit capturing of temporal dependencies in frame-wise shape sequences. Moreover, ASA is able to produce discriminative shape embeddings by amplifying the contribution of the most important frames using an attention mechanism. As shown in Figure 6, ASA attends to the most informative frames by assigning high attention scores, while noisy frames caused by occlusion receive little attention, which helps to minimize their impact on the aggregated video-wise shape representation.

Appearance and Shape Fusion. To fuse appearance and shape embeddings for final person representation, in this work, we simply apply concatenation, and compare it with the Weight Prediction Fusion (WPF) proposed in [15]. As shown in Table 6, SEMI with concatenation outperforms WPF by 2% in mAP in both evaluation settings on VCCR. For WPF, the embeddings $f^a \in \mathbb{R}^{2048}$ and $f^s \in \mathbb{R}^{10}$ are scaled to 512-dimensional vectors, then summed up with weights predicted by a convolutional layer, resulting in $f \in \mathbb{R}^{512}$. However, as each value $f^s_i, i = 1, ..., 10$ in f^s represents a certain aspect of shape like hip size and shoulder length, upscaling f^s to a high-dimensional vector potentially causes information loss of the body shape. Therefore, we apply concatenation to preserve the discriminative power of the shape embedding.

Method	CC		Standard	
	R-1	mAP	R-1	mAP
SEMI w/ WFP [15]	49.5	40.8	84.3	62.0
SEMI w/ concat (Ours)	**51.4**	**43.6**	**86.2**	**65.2**

Table 6. Comparison in appearance and shape fusion methods on VCCR: concatenation and Weight Prediction Fusion [15].

7. Conclusion

In this paper, we address the challenging problem of VCCRe-ID, where texture-based methods are limiting due to changes in clothing. We propose “Temporal 3D Shape Modeling for VCCRe-ID” (SEMI), a novel end-to-end framework that leverages human 3D shape to overcome the limitations of previous works. A temporal encoder first captures the temporal dynamics from the video sequences, then an identity-aware 3D regressor estimates frame-wise shape parameters. We introduce the Attention-based Shape Aggregation (ASA) module, which aggregates frame-wise shape features using GRUs and an attention mechanism. ASA allows for the amplification of the most informative frames, resulting in a robust and discriminative video-wise shape representation. Experimental results on two large-scale VCCRe-ID datasets demonstrate the superiority of our SEMI framework over state-of-the-art methods in real-world scenarios, achieving a significant improvement in all evaluation settings.

References

[1] Daniel Arkushin, Bar Cohen, Shmuel Peleg, and Ohad Fried. Reface: Improving clothes-changing re-identification with face features. *arXiv preprint arXiv:2211.13807*, 2022. 6

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 4

[3] Vaibhav Bansal, Gian Luca Foresti, and Niki Martinel. Cloth-changing person re-identification with self-attention. In *WACVW*, pages 602–610, 2022. 1

[4] Vaibhav Bansal, Christian Micheloni, Gianluca Foresti, and Niki Martinel. Spatio-temporal attention for cloth-changing reid in videos. In *ECCVW*, pages 353–368, 2023. 2

[5] Jiaxing Chen, Xinyang Jiang, Fudong Wang, Jun Zhang, Feng Zheng, Xing Sun, and Wei-Shi Zheng. Learning 3d shape feature for texture-insensitive person re-identification. In *CVPR*, pages 8142–8151, 2021. 2

[6] Jiaxing Chen, Wei-Shi Zheng, Qize Yang, Jingke Meng, Richang Hong, and Qi Tian. Deep shape-aware person re-identification for overcoming moderate clothing changes. *IEEE TMM*, 24:4285–4300, 2022. 1, 2

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 4

[8] Zhenyu Cui, Jiahuan Zhou, Yuxin Peng, Shiliang Zhang, and Yaowei Wang. Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. *IEEE TCSVT*, 2023. 1, 2, 6

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5

[10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, pages 5202–5211, 2020. 6

[11] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. In *CVPR*, pages 10696–10706, 2020. 2

[12] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*, number 01, pages 8287–8294, 2019. 2

[13] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *CVPR*, pages 1050–1059, 2022. 1, 2, 5, 6, 7

[14] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *ECCV*, pages 228–243, 2020. 1, 2, 5, 6

[15] Ke Han, Yan Huang, Shaogang Gong, Yan Huang, Liang Wang, and Tieniu Tan. 3d shape temporal aggregation for video-based clothing-change person re-identification. In *ACCV*, pages 71–88, 2022. 2, 3, 5, 6, 7, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016. 5, 7

[17] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *CVPR*, pages 10508–10517, 2021. 1

[18] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2017. 2, 7

[19] Khadija Khaldi, Pranav Mantini, and Shishir K. Shah. Unsupervised person re-identification based on skeleton joints using graph convolutional networks. In *Image Analysis and Processing*, pages 135–146, 2022. 1

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Nikos Kolotouros, Georgios Pavlakos, Michael Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261, 2019. 2, 6

[22] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. *AAAI*, page 8618–8625, 2019. 1, 2

[23] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, pages 772–788, 2018. 2

[24] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deep-reid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 2

[25] Yu-Jhe Li, Xinshuo Weng, and Kris M. Kitani. Learning shape representations for person re-identification under clothing change. In *WACV*, pages 2431–2440, 2021. 1, 2

[26] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015. 2

[27] Xuehu Liu, Pingping Zhang, Chenyang Yu, Huchuan Lu, and Xiaoyun Yang. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *CVPR*, pages 13329–13338, 2021. 6

[28] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6), 2015. 2, 4

[29] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR*, 2019. 2

[30] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multitask distance metric learning. *IEEE TIP*, 23(8):3656–3670, 2014. 2

[31] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, pages 1363–1372, 2016. 2

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. 6

[33] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019. 4

[34] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. In *ACCV*, pages 71–88, 2021. 1, 2

[35] Xiujun Shu, Ge Li, Xiao Wang, Weijian Ruan, and Qi Tian. Semantic-guided pixel sampling for cloth-changing person re-identification. *IEEE Signal Processing Letters*, 28:1365–1369, 2021. 6

[36] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 2, 6

[37] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*. ACM, 2018. 2

[38] Likai Wang, Xiangqun Zhang, Ruize Han, Jialin Yang, Xiaoyu Li, Wei Feng, and Song Wang. A benchmark of video-based clothes-changing person re-identification. *arXiv preprint arXiv:2211.11165*, 2022. 2, 3

[39] Yingquan Wang, Pingping Zhang, Shang Gao, Xia Geng, Hu Lu, and Dong Wang. Pyramid spatial-temporal aggregation for video-based person re-identification. In *ICCV*, pages 12006–12015, 2021. 1

[40] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Robust depth-based person re-identification. *IEEE TIP*, 26(6):2588–2603, 2017. 2

[41] Yiming Wu, Omar El Farouk Bourahla, Xi Li, Fei Wu, Qi Tian, and Xue Zhou. Adaptive graph representation learning for video person re-identification. *IEEE TIP*, 29:8821–8830, 2020. 1, 2

[42] Yuqiao Xian, Jinrui Yang, Fufu Yu, Jun Zhang, and Xing Sun. Graph-based self-learning for robust person re-identification. In *WACV*, pages 4789–4798, 2023. 2

[43] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, pages 701–716, 2016. 2

[44] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *CVPR*, pages 3289–3299, 2020. 1, 2

[45] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE TPAMI*, 43(6):2029–2046, 2021. 1, 2

[46] Peng Zhang, Qiang Wu, Jingsong Xu, and Jian Zhang. Long-term person re-identification using true motion from videos. In *WACV*, pages 494–502, 2018. 1, 2, 3

[47] Peng Zhang, Jingsong Xu, Qiang Wu, Yan Huang, and Xianye Ben. Learning spatial-temporal representations over walking tracklet for long-term person re-identification in the wild. *IEEE TMM*, 23:3562–3576, 2021. 1, 2, 3

[48] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10407–10416, 2020. 2

[49] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 2

[50] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 2

[51] Zhedong Zheng, Xiaohan Wang, Nenggan Zheng, and Yi Yang. Parameter-efficient person re-identification in the 3d space. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2022. 2

[52] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, pages 6776–6785, 2017. 2