This ACCV 2024 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2024
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

Cross-Modality Complementary Learning for Video-based Cloth-Changing Person Re-Identification

Vuong D. Nguyen, Pranav Mantini, and Shishir K. Shah

Quantitative Imaging Lab, Dept. of Computer Science, University of Houston
dnguy222@cougarnet.uh.edu

Abstract. Video-based Cloth-Changing Person Re-ID (VCCRe-ID) is a real-world Re-ID problem where individuals are observed in settings with a high likelihood of clothing changes between observations. To tackle this problem, capturing cloth-invariant modalities remains more effective than texture-based approaches. However, previous works extracted these modalities separately and directly leveraged the learned features for Re-ID, which is not effective since viewpoint changes and occlusion cause severe ambiguity in these modalities. To address this limitation, we propose a dual-branch framework that couples cloth-invariant modalities (i.e. shape and gait) with appearance by novelly exploiting the complementary relationship across them. In this work, we design a texture branch that enables body shape to complement the ambiguity in appearance caused by illumination variations or occlusions. Then texture and gait features are mutually learned at multiple levels, which helps to exchange beneficial information across branches for more discriminative person representations. We build a large-scale video-based cloth-changing dataset that contains the most cloth variations and is the first benchmark that mimics the real-world similar-clothing scenario. Extensive experiments show that our proposed framework outperforms state-of-the-art methods by a large margin. Code and dataset will be available at https://github.com/dustin-nguyen-qil/CCL-VCCReID

Keywords: Cloth-Changing Person Re-Identification · Cross-Attention · Spatio-Temporal Representation Learning · Gait Recognition

1 Introduction

Person Re-Identification (Re-ID) is the task of matching individuals across different cameras. Video-based Re-ID has been actively explored [34, 39, 70] since the advancement of deep learning [32, 44, 53]. These methods primarily rely on appearance and fail under clothing-confusion situations [51]. This has led to the development of methods to address the problem of Video-based Cloth-Changing Person Re-ID (VCCRe-ID).

To tackle Cloth-Changing Re-ID (CCRe-ID), several works mine clothing status or clothing templates as auxiliary pseudo-labels [20, 25, 77, 79]. This ap-

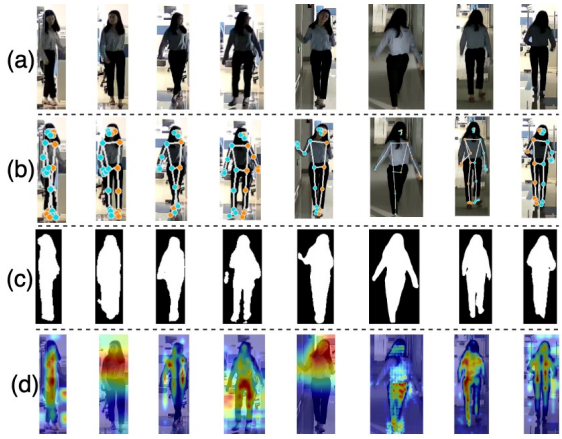

Fig. 1: Visualization of: (a) original sequence, (b) skeleton-based pose, (c) silhouettes. Viewpoint changes, occlusion, and illumination cause ambiguity in capturing identity-aware human structure from these 2D modalities and make cloth-invariant cues like face or hairstyle unobservable. We address this by adaptively attending to the most informative cues across modalities as shown in **(d) activation maps learnt by our proposed texture branch**, where red means larger feature scores and blue means smaller feature scores.

proach is not scalable under incremental data scenario with new clothing variations. Texture-based approaches such as [12, 16] extract cloth-invariant features from face or hairstyle. However, they fail under occlusions. Methods that rely on multiple modalities are more effective in such scenarios, where cloth-invariant modalities (e.g. body shape and gait) are captured and coupled with appearance. Body shape has been widely captured from 2D pose [45, 54, 63], silhouettes [19, 23, 26, 46, 48], or contour sketches [9, 76] for CCRe-ID. However, as shown in Figure 1(b)(c), viewpoint changes result in dissimilar poses and silhouettes for the same ID causing inaccuracy in capturing body shape in 2D space. More recent works [8,21,88] have proposed to extract 3D shape for Re-ID. However, these works do not model long-range dependencies from sequences, which can improve Re-ID learning. Moreover, they require 3D human datasets for regularization, leading to expensive training phases. Gait has been extracted from skeleton sequences [82], or 3D human structure [2] for CCRe-ID. However, these works employ off-the-shelf pose estimators [5, 59] or 3D human reconstruction models [27, 31], which are designed to produce only rough estimations and lack discriminability for Re-ID. Importantly, in all previous multi-modality methods, the constructed branches learn appearance, shape, and gait separately, then directly utilize those learned embeddings for matching individuals. This undermines the complementary relationship between modalities, while we found that it can significantly help to mitigate the influence of viewpoint changes, occlusions, and other in-the-wild Re-ID challenges.

To this end, based on the above finding, we propose a novel dual-branch framework for VCCRe-ID. We approach VCCRe-ID by extracting multiple modalities (i.e. appearance, shape, and gait), and the key contribution is our Cross-Modality Complementary Learning (CMCL) strategy. We leverage silhouette masks estimated from the RGB frames to capture body shape and gait, which has shown effectiveness in [19,23,26,37]. Then, our CMCL strategy is elaborately designed in the framework as follows.

As shown in Figure 1(a), occlusion and illumination variations significantly change the visual appearance of a person, affecting the ability to learn visual similarities from appearance. Thus, in texture branch, shape features encoded from silhouettes are used to complement appearance using 3D Cross-Attention blocks (3D-CAB). 3D-CAB comprises self-attention and cross-attention mechanisms that operate on sequences of appearance and shape feature maps at spatial-temporal level, enabling the model to adaptively attend to the most informative Re-ID cues under in-the-wild challenges. The effectiveness of our approach is illustrated in Figure 1(d), where the model knows to focus on body shape cues when faces and hairstyles are unobservable, while facial and appearance features are amplified in frames with visible appearance. For gait branch, we encode gait from silhouettes (estimated in the texture branch) instead of skeleton-based poses. The reasons are two-fold: (1) skeleton-based pose is sparse, which limits the capturing of semantic geometric information and temporal dependencies; and (2) we save on the computational cost of extracting another modality. Features from the two branches are fed into a mutual learning module, where beneficial information is exchanged between the branches (modalities) at multiple levels. This helps to maximize the robustness of our framework against in-the-wild Re-ID challenges. A problem with silhouette-based gait learning is the dissimilarity of silhouettes across the video sequence, which is caused by viewpoint variations as shown in Figure 1(c). To address this, inspired by [85], we leverage 3D knowledge based on the Skinned Multi-Person Linear Model (SMPL) [42] to normalize the frame-wise silhouettes onto a common latent space. SMPL provides 3D pose and viewpoint information, which can be leveraged for cross-view normalization of silhouette, and 3D shape information also helps to ensure the identity-awareness of normalized silhouettes.

There have been a limited number of public datasets proposed for VCCRe-ID (as will be shown in Table 1). Most datasets represent simplistic Re-ID scenarios such as frontal viewpoints or no occlusions. The range of clothing variations per identity in these datasets is small and unbalanced, which may result in degradation in Re-ID accuracy due to bias in matching. Moreover, there are no datasets that consider the situation of different identities wearing similar or same clothes. The visual similarities of this scenario may cause severe ambiguity in appearance, which helps further evaluate the effectiveness of cloth-invariant modalities like shape and gait. To this end, we propose **E-VCCR** dataset, in which we perform cross-identity cloth transfer on VCCR [21] using DG-Net [89]. E-VCCR is large-scale with more than 8k tracklets. It provides a wide and balanced range of clothing variations per identity and mimics the challenging similar-clothing

scenario (see Sec. 4). It also provides additional modalities including silhouette masks, 2D/3D poses, and 3D SMPL models.

Our contributions in this paper are:

1. We conduct a thorough study of the practical VCCRe-ID task.
2. We propose the "Cross-modalities Complementary Learning" framework for VCCRe-ID in which appearance and cloth-invariant modalities are learned in a mutual and complementary manner at both spatial and temporal level, thus effectively handling clothing changes, viewpoint variations, and other in-the-wild Re-ID challenges.
3. We propose the large-scale E-VCCR dataset, which provides wide clothing variations and mimics more practical Re-ID scenarios, aiming to facilitate research in VCCRe-ID.
4. We perform extensive experiments on three VCCRe-ID datasets and show that our framework achieves state-of-the-art performance.

2 Related Work

2.1 Person Re-Identification (Re-ID)

Under short-term scenarios without clothing changes, person Re-ID has achieved remarkable success on standard datasets [71, 86, 87]. Convolutional Neural Networks have been widely adopted as deep feature extractors for image-based [60, 65, 91] and video-based [17, 34] Re-ID. Several works have tackled spatial misalignment posed by viewpoint changes [29, 90], occlusions [38, 43], or pose variations [10, 55]. Spatial-temporal information has been captured using Graph Convolutional Networks (GCNs) in [28, 72, 73] to promote video-based Re-ID. However, these methods are inapplicable in cloth-changing scenarios since they rely heavily on appearance for individual matching.

2.2 Image-based Cloth-Changing Person Re-ID

Recently, several image-based CCRe-ID datasets have been collected [24, 35, 54, 64, 74, 76, 79]. Existing methods for image-based CCRe-ID can be categorized into single-modality and multi-modality methods. The former only leverage the RGB image to extract biometric features via regional component reconstruction [12], hierarchical feature accumulation [78], or regional feature weighting [40]. The latter leverage cloth-irrelevant modalities that are more stable in long-term such as contour sketches [9], silhouettes [23, 26, 46, 48], and skeleton-based pose [36, 45, 54, 63]. Due to the confusion in 2D human geometric cues caused by viewpoint changes, 3D body structure is leveraged in [8]. However, the 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.

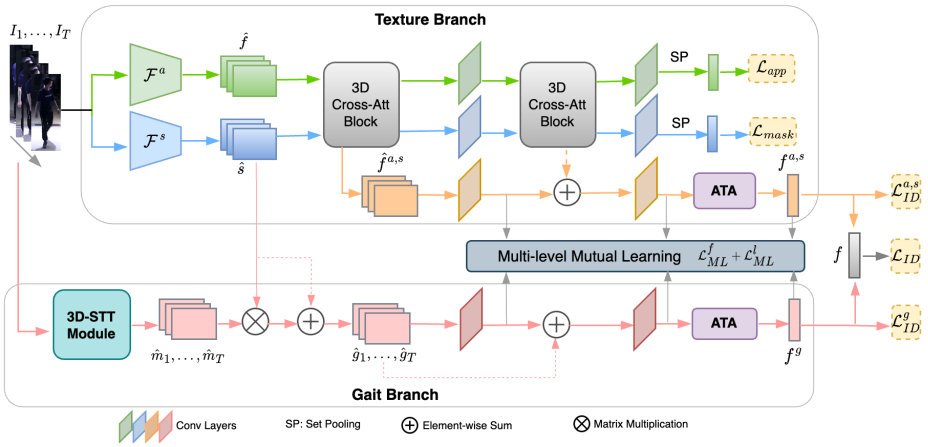

Fig. 2: Overview of our proposed framework. Texture branch encodes appearance and shape using $\mathcal{F}^a$ and $\mathcal{F}^s$, respectively, then capture their complementary relationship using 3D-CAB, producing texture-based embedding $f^{a,s}$. Gait branch learns transformation matrices using the 3D-STT module, then performs normalizing silhouette feature maps $\hat{s}$ for gait embedding f^g. Texture and gait are mutually learnt, then aggregated using ATA [49], and finally concatenated for person representation f.

2.3 Video-based Cloth-Changing Person Re-ID

There is limited research till date to address the challenges of VCCRe-ID. Texture-based methods [1,12,16] attempt to decouple features of faces or hairstyles from clothing status. Relying primarily on texture information significantly limits the ability for Re-ID under occlusion. In this work, we address this by adaptively transferring knowledge from human geometric features to complement the appearance features. Han *et al.* [21] proposed to capture 3D shape using an auxiliary 3D human dataset for regularization. This framework is multi-stage and requires heavy training. Several works [50, 67, 81, 82] leverage spatial-temporal information from video sequences to extract motion patterns as gait cues. For example, Zhang *et al.* [82] propose to use GCNs to model gait representation from the skeleton-based 3D pose sequence. However, these works rely heavily on the visual observation of body parts, which is affected by viewpoint changes. To tackle this, we leverage 3D knowledge to learn a latent space that minimizes intra-class gap and maximizes the inter-class gap of gait representations.

3 The Proposed Framework

An overview of our proposed framework is illustrated in Figure 2. Given a video sequence of T frames $X = \{I_i\}_{i=1}^{T}$ as input, the Texture branch learns frame-wise texture-based sequence $\hat{f}^{a,s}$ by maximizing the complementary relationship between appearance and shape using 3D-Cross-Attention Block (3D-CAB). Meanwhile, the Gait branch leverages 3D SMPL information to learn transformation

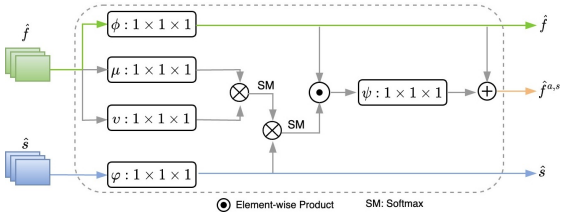

Fig. 3: 3D-CAB architecture. (Best viewed in color.)

matrices for normalizing the silhouette sequence in the latent space, producing frame-wise gait sequence $\hat{g}$. The Multi-level Mutual Learning module performs knowledge transfer between two branches. Attention-based Temporal Aggregation (ATA) module [49] then aggregates $\hat{f}^{a,s}$ and $\hat{g}$ for the texture-based embedding $f^{a,s}$ and gait embedding f^{g}, respectively, which are then concatenated to obtain the video-wise person representation f. During inference, the input video sequence is only passed to Texture branch to obtain embedding f for matching, where similarity scores are computed between the query embedding and every gallery embedding, then ranked to obtain the matched ID. Gait branch is only used during training to enhance robustness of Texture branch and will be discarded during inference to save computational cost.

3.1 Texture Branch

The goal of the Texture branch is to learn a robust texture-based representation by making shape information adaptively complement ambiguous appearance. As shown in Figure 2, Texture branch consists of a CNN backbone $\mathcal{F}^{a}(\cdot)$ that extracts mid-level frame-wise appearance features $\hat{f} = \{\mathcal{F}^{a}(I_i)\}_{i=1}^{T} = \{\hat{f}_i\}_{i=1}^{T}$. Meanwhile, a segmentation encoder $\mathcal{F}^{s}(\cdot)$ estimates mid-level silhouette features $\hat{s} = \{\mathcal{F}^{s}(I_i)\}_{i=1}^{T} = \{\hat{s}_i\}_{i=1}^{T}$. This suppresses background clutter in extracting geometric cues from body shape.

Cross-Attention has shown great effectiveness in Re-ID [4, 56, 68], while 3D Cross-Attention [69] has paved the way for multiple computer vision tasks [18]. In this work, we specifically design a **3D Cross-Attention Block (3D-CAB)**, which takes in appearance and shape feature maps and makes them complement each other at both frame-level and sequence-level (Figure 3). It consists of five linear embeddings implemented as $1 \times 1 \times 1$ convolutions [69] named $\phi, \mu, \upsilon, \varphi, \psi$, which help capture non-local features at pixel-level by enlarging the receptive field within a frame while modeling temporal dependencies across frames. Given the frame-wise appearance and silhouette feature maps $\hat{f}, \hat{s} \in \mathbb{R}^{T \times H \times W \times C}$, we first apply self-attention mechanism [80] to the appearance feature maps:

$$\delta(k) = \mathrm{SM}\left(\sum_{q}\left(\mu\left(\hat{f}(k)\right)^{T} \upsilon\left(\hat{f}(q)\right)\right)\right), \tag{1}$$

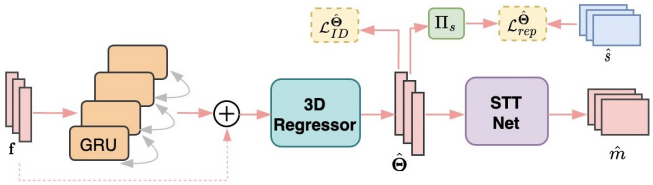

Fig. 4: Architecture of the 3D-STT module.

where $\delta(k)$ is the accumulative response at position (t, x, y) with respect to all possible positions q in the input feature maps, $\delta \in \mathbb{R}^{THW \times THW}$, and SM denotes softmax activation. A mask embedding φ is applied to silhouettes maps $\hat{s}$ to filter noise from background. The filtered map is then multiplied with the attention map to capture long-range dependencies between appearance and shape features:

$$\xi(k) = \mathrm{SM}\left(\sum_{q}\left(\delta(k)^{T}\varphi(\hat{s}(q))\right)\right), \tag{2}$$

where $\xi \in \mathbb{R}^{THW \times \overline{C}}$, $\overline{C} = C/2$. ξ is then multiplied with the latent representation of each position in the input appearance feature map, formulated as:

$$\zeta = \xi \odot \phi\left(\hat{f}\right). \tag{3}$$

Another linear embedding ψ is then applied on ζ to produce non-local feature maps and re-project ζ onto original space, which is then summed with the original input feature map using a residual connection [22] as:

$$\hat{f}^{a,s} = \psi(\zeta) + \hat{f}. \tag{4}$$

This results in the cross-shape-appearance feature map $\hat{f}^{a,s} \in \mathbb{R}^{T \times H \times W \times C}$. Note that the original appearance and silhouettes feature maps $\hat{f}$ and $\hat{s}$ are not altered by 3D-CAB. As shown in Figure 2, the feature maps $\hat{f}$, $\hat{s}$, and $\hat{f}^{a,s}$ are then fed into the corresponding convolutional layers that decrease channel size to reduce computation, shown by green, blue, and orange flows. Another 3D-CAB is cascaded to progressively refine the features extracted from the previous block, which aids in reducing noise and enhancing feature discrimination. We then obtain the video-wise appearance and shape feature vectors f^a, f^s using Set Pooling [6] (SP). Given the identity label, f^a and f^s are fed into the cross-entropy (CE) losses $\mathcal{L}_{app}$ and $\mathcal{L}_{mask}$ for identity-guidance in training. The CE losses are based on identity labels. The texture-based feature vector $f^{a,s}$ is obtained using the Attention-based Temporal Aggregation module employed from [49].

3.2 Gait Branch

The goal of the Gait branch is to learn identity-aware 3D knowledge from the input video tracklet, then leverage it to normalize the silhouette sequence, and finally aggregate the sequence for gait representation.

3D Spatial-Temporal Transformation (3D-STT). The 3D-STT module (Figure 4) learns transformation matrices from 3D SMPL knowledge for gait normalization using STT Network. Given frame I_i, SMPL represents the 3D human body as a 85-dimensional vector $\hat{\Theta}_i = \{\theta, \beta, \omega\}$ where $\theta \in \mathbb{R}^{69}$, $\beta \in \mathbb{R}^{10}$, and $\omega \in \mathbb{R}^6$ denote pose, shape, and viewpoint parameter sets. To estimate $\hat{\Theta}_i$, previous works [2, 8, 85] directly employ off-the-shelf 3D human reconstruction models [27, 61], which lack the modeling of temporal dependencies and discriminability across individuals since these models are designed to produce neutral models. To address this, we propose identity-aware 3D SMPL Temporal Learning. Specifically, for frame I_i, feature vector $\mathbf{f}_i \in \mathbb{R}^{2048}$ is extracted using the same encoder $\mathcal{F}^a$ from Texture branch, i.e. $\mathbf{f}_i = \mathcal{F}^a(I_i)$. First, the temporal encoder $\mathcal{G}(\cdot)$ consists of several GRUs [11] layers (can be replaced with other temporal encoding techniques), which produces the latent code $\mathbf{g}_i = \mathcal{G}(\mathbf{f}_i)$ that captures the long-range dependencies across frames. The 3D regressor $\mathcal{R}(\cdot)$, employed from [27], then takes the residual sum $\mathbf{h}_i = \mathbf{f}_i + \mathbf{g}_i$ as input and outputs $\hat{\Theta}_i$, i.e. $\hat{\Theta}_i = \mathcal{R}(\mathbf{h}_i)$. $\mathcal{R}$ consists of several fully-connected layers and is initialized with the template SMPL model to ensure the realisticity of the estimated 3D human body. To enhance the **identity-awareness** of $\hat{\Theta}_i$, we first use a cross-entropy identification loss $\mathcal{L}_{ID}^{\hat{\Theta}}$, then we feed the silhouette feature map $\hat{s}_i$ and 2D silhouette reprojection $\Pi_{sil}(\hat{\Theta}_i)$ into the L1 reprojection loss $\mathcal{L}_{rep}^{\hat{\Theta}}$, formulated as:

$$\mathcal{L}_{rep}^{\hat{\Theta}} = \|\hat{s}_i - \Pi_{sil}(\hat{\Theta}_i)\|_1. \tag{5}$$

The STT Network $\mathcal{M}(\cdot) : \mathbb{R}^{85} \rightarrow \mathbb{R}^{HWC}$ comprises of several fully-connected layers that takes in the frame-level SMPL model $\hat{\Theta}_i$ and learns the transformation vector m_i for frame I_i, formulated as: $m_i = \mathcal{M}(\hat{\Theta}_i)$. m_i is then reshaped to matrix $\hat{m}_i \in \mathbb{R}^{C \times W \times H}$. We then apply zero padding on the short edge of $\hat{m}_i$ and $\hat{s}_i$ to expand them to square matrices for convenience in computation.

For **3D Gait Normalization**, first, we multiply the transformation matrix with the corresponding silhouette feature map at frame-wise level. Then, we apply residual connection to the original feature map sequences in order to preserve important high-level information from the input features while leveraging the temporal dynamics, given as:

$$\hat{g}_i = \left(\hat{s}_i^T \cdot (\mathbf{I} + \hat{m}_i)\right)^T, \tag{6}$$

where $\hat{g}_i \in \mathbb{R}^{H \times H \times C}$, and $\mathbf{I}$ denotes the identity matrix.

We also employ the ATA module from [49] to produce the video-wise gait embedding $\hat{f}^g$. The cross-entropy losses, $\mathcal{L}_{ID}^{a,s}$ and $\mathcal{L}_{ID}^{g}$, are used to enhance identity-awareness of each branch. Texture and gait embeddings are concatenated, giving the final person representation $f = [f^{a,s}, f^g]$ supervised by $\mathcal{L}_{ID}$, which is the sum of a cross-entropy loss $\mathcal{L}_{ce}$ and a pair-wise triplet loss $\mathcal{L}_{tri}$.

3.3 Multi-level Mutual Learning

Simply concatenating the texture and gait embeddings for person representations tends to undermine the complementary knowledge of intermediate-level

features between the two branches. Inspired by [83], we propose Multi-level Mutual Learning (MML) module, in which each branch mutually serves as a teacher to drive the cloth-irrelevant feature extraction of the other via knowledge transfer at multiple levels. Given a batch of N tracklets, the learned gait knowledge is represented using feature similarity matrix as:

$$\mathbf{S}^g = \sum_{k=1}^{N}\sum_{l=1}^{N} \text{AvgPool}\left(\hat{g}^k\right) \cdot \text{AvgPool}\left(\hat{g}^l\right), \tag{7}$$

where AvgPool denotes average pooling. The feature similarity matrix of texture branch is similarly represented by $\mathbf{S}^{a,s}$. Then, mutual learning is performed across two 3D-CABs at feature level by minimizing the distance between the learnt knowledge of two branches, formulated as:

$$\mathcal{L}_{ML}^{f} = \sum_{m=1}^{2} \left(\mathbf{S}_m^{a,s} - \mathbf{S}_m^{g}\right), \tag{8}$$

where $\mathbf{S}_m$ denotes the similarity matrix after the m^{th} 3D-CAB. Experimentally, models using only gait embeddings for Re-ID fail to achieve comparable performance (see Sec. 5.4 for more details). Thus, MML module also encourages a high-level semantic consistency between two branches by a logit-level mutual learning loss, given by:

$$\mathcal{L}_{ML}^{l} = D_{KL}\left(\hat{p}^{a,s} \| \, \hat{p}^{g}\right) + D_{KL}\left(\hat{p}^{g} \| \, \hat{p}^{a,s}\right), \tag{9}$$

where $D_{KL}\left(\hat{p}^{a,s} \| \, \hat{p}^{g}\right)$ is the Kullback-Leibler distance from output class probabilities $\hat{p}^{a,s}$ of texture branch to output class probabilities $\hat{p}^{g}$ of gait branch.

Our framework is trained by the total loss $\mathcal{L}$, given as:

$$\begin{aligned}\mathcal{L} = {} & \lambda_1\left(\mathcal{L}_{app} + \mathcal{L}_{mask}\right) + \lambda_2\left(\mathcal{L}_{rep}^{\hat{\Theta}} + \mathcal{L}_{ID}^{\hat{\Theta}}\right) \\ & + \lambda_3\left(\mathcal{L}_{ML}^{f} + \mathcal{L}_{ML}^{l}\right) + \lambda_4\left(\mathcal{L}_{ID} + \mathcal{L}_{ID}^{a,s} + \mathcal{L}_{ID}^{g}\right),\end{aligned} \tag{10}$$

where $\lambda_1, ..., \lambda_4$ are scalars controlling the scale of the losses.

4 The Proposed E-VCCR Dataset

4.1 VCCRe-ID Datasets

A summary of existing datasets for VCCRe-ID is reported in Table 1. Among the few public datasets [16, 21, 36], CCVID [16] is constructed from the Front View Gait dataset [84], thus mimics simplistic scenarios with frontal viewpoints, clearly visible faces, and slight clothing changes. CCPG [36] dataset was captured in a fixed scene under good lighting and no background clutter or occlusion but presents substantial viewpoint variations and clothing changes. VCCR [21]

Dataset	#IDs	#Tracklets	#Clothes/ID	Environment	Modalities	Distractors	Public
Motion-ReID [81]	30	240	-	Indoor	RGB	✗	✗
CVID-reID [82]	90	2,980	-	Outdoor	RGB	✗	✗
CCVID [16]	226	2,856	$2 \sim 5$	Outdoor	RGB	✗	✓
CCPG [36]	200	16,566	4	Indoor	RGB, Sils	✗	✓
VCCR [21]	392	4,384	avg. 3.3	Indoor	RGB	✗	✓
E-VCCR	392	8,396	**10**	Indoor	RGB,Sils,2D/3DPose,3DSMPL	✓	✓

Table 1: A summary of datasets for VCCRe-ID and our proposed E-VCCR dataset.

constructed from the RAP dataset [33] poses realistic challenges for Re-ID including illlumination variations, occlusions, and broad range of pose variations across IDs. Hence, CCPG and VCCR are suitable for baseline experimental studies. However, these datasets contain a small and unbalanced range of cloth variations per identity, limiting the understanding of the impact of cloth-changes in VCCRe-ID. The unbalanced number of cloth changes per ID may also confound the matching results where a subset of the IDs may be easier to match over others. Further, they do not explicitly mimic the real-world similar-clothing scenario. The similarity in visual appearance across IDs may cause severe ambiguity in matching, which helps further evaluate the effectiveness of clothing-invariant modalities like shape and gait. To address these, we propose to build E-VCCR dataset by synthesizing cloth-transfer tracklets across identities.

4.2 Construction of E-VCCR Dataset

Acquiring and annotating CCRe-ID datasets is expensive. Moreover, synthesizing data using generative models is becoming more favorable for person Re-ID task [7,14,75]. Thus, we construct E-VCCR by performing cloth-transfer synthesis across identities using DG-Net [89] on VCCR. First, we aim to generate more cloth-changing samples and balance the number of suits per identity for training. For each tracklet in the original training set, we transfer clothes from a tracklet of a different identity, giving a total of $5,746$ tracklets in 10 suits per identity in the new training set of E-VCCR. Second, for evaluation, for each probe in the query set, we create distractors of *different identity* but *similar clothing* in the gallery set. The new gallery set of E-VCCR contains $2,154$ tracklets. To serve multimodal research purposes, we also provide other modalities in E-VCCR, which are acquired as follows: (1) 3D SMPL are estimated using our proposed SMPL Learning module (Sec. 3.2), (2) 2D silhouettes are obtained using HR-Net [66], and (3) 2D/3D pose skeletons are estimated using MediaPipe [3]. An illustration and details of E-VCCR are provided in the **Supp. Mat.**.

5 Experiment Setup

5.1 Evaluation Protocols

We use VCCR [21], CCPG [36], and E-VCCR for experiments. CMC score at rank-1 (R-1 accuracy) and mean Average Precision (mAP) are computed for

Method	Method type	VCCR				CCPG				E-VCCR			
		CC		Standard		CC		Top/Bottom		CC		Standard	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
PCB [60]	Image-based	18.8	15.6	55.6	36.6	-	-	-	-	-	-	-	-
AP3D [17]†	Video-based	35.9	31.6	78.0	52.1	80.4	53.1	51.1	33.8	30.2	24.1	70.9	43.3
GRL [41]	Video-based	35.7	31.8	76.9	51.4	-	-	-	-	-	-	-	-
STCA [4]	Video-based	35.8	32.0	77.1	52.0	-	-	-	-	-	-	-	-
SPS [57]	Image-based CC	34.5	30.5	76.5	50.6	-	-	-	-	-	-	-	-
CCPG [47]	Image-based CC	34.9	30.8	76.8	50.9	-	-	-	-	-	-	-	-
CAL [16]†	Video-based CC	36.6	31.9	78.9	52.9	82.3	54.5	54.9	35.1	34.1	26.1	73.2	44.8
3STA [21]	Video-based CC	40.7	36.2	80.5	54.3	-	-	-	-	-	-	-	-
SEMI [49]†	Video-based CC	51.4	43.6	86.2	65.2	<u>83.5</u>	<u>56.1</u>	<u>57.8</u>	<u>37.6</u>	<u>40.1</u>	<u>34.5</u>	<u>79.0</u>	<u>51.3</u>
ASGL [50]	Video-based CC	<u>52.9</u>	<u>43.2</u>	<u>88.1</u>	<u>65.8</u>	-	-	-	-	-	-	-	-
Ours	Video-based CC	**54.2**	**44.2**	**89.3**	**66.9**	**84.1**	**56.2**	**59.4**	**38.9**	**46.3**	**38.6**	**80.7**	**52.0**

Table 2: Comparison between our method and current SOTA methods. "†" denotes results are reproduced on CCPG and E-VCCR based on the open-source code of the method. "-" denotes results are not reported or codes are not available for reproducibility. Best results are shown in **bold**, while second-to-best results are <u>underlined</u>.

evaluation. For VCCR and E-VCCR, we use the following evaluation settings: (1) *Cloth-changing (CC)*, i.e. only cloth-changing samples are used; (2) *Standard*, i.e. both cloth-changing and cloth-consistent samples are used. For CCPG, due to its nature, two settings are used: (1) *Cloth-changing (CC)*, i.e. only the entire-outfit-changing samples are considered; (2) *Top/Bottom-changing (Top/Bottom)*, i.e. only top-changing or pant-changing samples are considered.

5.2 Implementation Details

Architecture. We adopt ResNet-50 [22] pretrained on ImageNet [13] as the CNN backbone $\mathcal{F}^a$, while the pretrained HRNetV2-W48 [66] is employed as the segmentation encoder $\mathcal{F}^s$. For the gait branch, the temporal encoder $\mathcal{G}$ consists of two GRU layers with 1024 neurons each, followed by a linear projection layer. The 3D regressor $\mathcal{R}$ consists of three 1024 fully-connected layers with a dropout layer in between. $\mathcal{R}$ is initialized with pretrained weights from SPIN [31]. The 3D-STT module consists of three 1024 fully-connected layers. For the ATA module, we employ 2-layer GRUs of size 1024, followed by a self-attention mechanism with an MLP layer of size 1024. Implementation is in PyTorch [52].

Training and Testing. To form input clips for training, 8 frames are randomly sampled from each tracklet with a stride of 2 for VCCR, E-VCCR and 4 for CCPG. We first resized each frame to 256×128, then applied horizontal flipping as augmentation. The batch size is set to 32, each batch randomly selects 8 IDs and 4 clips per ID. The model is trained for 120 epochs using Adam optimizer [30]. Learning rate is initialized at $5e^{-3}$ and reduced by a factor of 0.1 after every 40 epochs. We set $\lambda_1 = 0.1$, $\lambda_2 = 0.5$, $\lambda_3 = 0.5$, $\lambda_4 = 1$ in the total loss function $\mathcal{L}$. In testing stage, we applied the same sampling strategy on all datasets to form 8-frame input clips. To save computation costs of 3D SMPL estimation in gait branch for fast inference, only the texture branch is used to obtain person representations for matching. Given a query, a ranking list is computed from gallery set based on pair-wise cosine distances.

Ablation study	Validate the effectiveness of
Texture vs Gait vs Joint	Each branch and each particular modality
Shape-Appearance Complementary Learning	3D-CAB in cross-modality learning in Texture branch
Gait Representation Learning	Normalizing gait using 3D knowledge
Identity-guidance for Gait Learning	Identity-awareness in 3D SMPL estimation
Multi-level Mutual Learning	Cross-modality mutual learning
Number of cloth variations per ID	The number of cloth variations present in data

Table 3: Design of ablation studies.

5.3 Comparison with State-of-the-Art

A comparison with the state-of-the-art methods (SOTAs) is reported in Table 2. Overall, our framework outperforms SOTAs on every dataset in all evaluation protocols by a large margin.

On VCCR, we achieve a significant boost of 1.3%/1.2% in R-1 accuracy and 1.0%/1.1% in mAP in CC/standard setting, respectively. The multi-stage 3STA [21] framework requires auxiliary large-scale 3D human datasets for regularization. Moreover, it requires a heavy training process of 30000 training epochs for the second stage as reported in [21]. Our framework is trained in an end-to-end manner for 120 epochs without any additional datasets. Compared to CAL [16] which relies on only appearance, our method shows superiority by complementing appearance with shape and gait to mitigate the influence of viewpoint changes and occlusions. Importantly, compared with STCA [4] which also leverages cross-attention for learning video-based person embeddings, we outperform STCA by more than 10% in R-1 accuracy in both evaluation settings. This demonstrate the effectiveness of our training strategy where we feed appearance and shape feature maps into the cross-attention module and capture their complementary relationship for Re-ID.

On CCPG, for AP3D [17], we report the results reproduced by using the experimental configurations provided in the original paper [36]. Despite the saturation caused by the simplistic Re-ID scenarios in the dataset, we achieve an improvement in performance in both evaluation settings.

On E-VCCR, the generated distractors result in a sharp performance drop of texture-based methods [16,17]. However, by coupling 3D gait cues with texture, we significantly improve Re-ID accuracy under cloth-changing situations with severe confusion. Compared to SEMI [49] which only uses 3D shape features, it is further shown that Re-ID performance is boosted by our framework via also using gait. Our method importantly shows robustness under the similar-clothing scenario, shown by a visualization of retrieval ranking list in the **Supp. Mat.**.

5.4 Ablation Study

To better understand the specific contribution of each component in our framework, we perform ablation studies as summarized in Table 3. Further analysis on feature fusion and the ATA module are provided in the **Supp. Mat.**.

Texture vs Gait vs Joint. In Table 4, we show experiments on the following model settings: texture branch only (T(Ours)), gait branch only (G), and joint

Method	VCCR		CCPG		E-VCCR	
	R-1	mAP	R-1	mAP	R-1	mAP
T (C2D [69])	34.7	30.5	78.9	52.1	29.0	22.8
T (Ours)	40.1	35.0	81.0	54.2	36.5	27.1
G	25.6	23.3	70.8	46.2	24.1	19.9
T (C2D [69]) + G	47.0	39.1	82.9	55.0	42.1	35.5
T (Ours) + G	**54.2**	**44.2**	**84.1**	**56.2**	**46.3**	**38.6**
T (1 3D-CAB) + G	53.1	43.3	82.0	54.6	45.3	37.7
T (3 3D-CAB) + G	54.0	43.9	83.8	56.0	46.1	38.5

Table 4: Ablation studies on: (1) contribution of each branch, where T denotes Texture and G denotes Gait branch, and (2) effectiveness of SACL and 3D-CAB.

Method	VCCR		CCPG		E-VCCR	
	R-1	mAP	R-1	mAP	R-1	mAP
GaitSet [6]	47.2	39.6	81.9	55.4	42.1	34.8
GaitGraph [62]	44.6	36.9	81.4	54.8	39.2	31.1
3STGE (Ours)	**52.2**	**42.8**	**83.0**	**55.4**	**44.0**	**36.5**

(a)

Num. clothes	CC		Standard	
	R-1	mAP	R-1	mAP
Four	41.6	36.1	80.2	53.3
Eight	46.3	38.5	84.7	57.1
Nine	46.3	38.6	84.7	57.0

(b)

Table 5: (a) Ablation study on the effectiveness of Gait branch (MML is not applied for fair comparison); and (b) Analysis on E-VCCR about the impact of number of cloth variations per ID on model's generalizability.

(T(Ours) + G). It can be seen that relying on only gait representations for Re-ID does not produce satisfactory results, since texture information remains more competitive in visual similarities when identities slightly change clothes, and capturing gait features relies on the performance of the segmentation model in estimating silhouettes. Re-ID accuracy is maximized by coupling gait with texture, showing the effectiveness of our dual-branch framework.

Shape-Appearance Complementary Learning (SACL). As shown in Table 4, Re-ID accuracy is remarkably improved by using SACL in our Texture branch (T(Ours)) compared to C2DResNet-50 [69] (T(C2D)). For example, on VCCR, SACL helps to achieve 7.2%/5.1% higher R-1 accuracy and mAP in the joint model setting (T(Ours) + G *vs.* T(C2D) + G). This shows the effectiveness of the 3D-CAB in capturing the spatial-temporal complementary relationship between appearance and shape features to tackle viewpoint changes and occlusions. Moreover, from the last three rows in Table 4, it can be seen that cascading two 3D-CAB in the texture branch helps to yield the highest Re-ID performance compared to using one or three 3D-CAB.

Gait Representation Learning. We compare our gait branch with two gait recognition models: silhouette-based GaitSet [6] and pose-based GaitGraph [62]. Both models are reimplemented based on OpenGait [15] toolbox. GaitSet directly aggregates features learned from silhouette sequence using a shallow CNN. GaitGraph adopts ResGCN [58] to operate on 2D skeleton-based graph, which suffers from low accuracy of pose estimators caused by occlusion. From Table 5a, it can be seen that we significantly improve the discriminative power of gait representations by performing cross-view normalization on silhouette sequences using 3D SMPL knowledge.

Method	$\mathcal{L}_{rep}^{\hat{\Theta}}$	$\mathcal{L}_{ID}^{\hat{\Theta}}$	VCCR		E-VCCR	
			R-1	mAP	R-1	mAP
1	-	-	51.0	42.1	43.5	36.0
2	✓	-	52.9	43.0	44.8	37.2
Ours	✓	✓	**54.2**	**44.2**	**46.3**	**38.6**

(a) Identity-guidance for gait learning.

Method	$\mathcal{L}_{ML}^{f}$	$\mathcal{L}_{ML}^{l}$	VCCR		E-VCCR	
			R-1	mAP	R-1	mAP
1	-	-	52.2	42.8	44.0	36.5
2	✓	-	53.5	43.6	45.8	37.5
Ours	✓	✓	**54.2**	**44.2**	**46.3**	**38.6**

(b) MML.

Table 6: Ablation studies of: a) identity-guidance for gait learning and b) Multi-level Mutual Learning (MML).

How many cloth variations per ID can lead to good generalizability of model? It is essential to analyze the trade-off between the model's generalizability and the practicality of collecting and labeling cloth-changing Re-ID data. Thus, we proposed the enriched E-VCCR in which the number of clothing variations per identity is upscaled and balanced. Then, we gradually increase the number of variations during training and monitor the model's performance on test sets to look for a saturation point. As can be seen in Table 5b, the Re-ID accuracy begins to saturate at 8 clothing variations. This suggests that additional variations only lead to an increase in data collection and annotation workload rather than in performance improvements.

Multi-level Mutual Learning (MML). Table 6b shows that capturing the complementary relationship between modalities using MML is beneficial for Re-ID. Specifically, mid-level feature transfer via $\mathcal{L}_{ML}^{f}$ boosts rank-1 accuracy and mAP on VCCR by 1.3% and 0.8%, which are further boosted by 0.7% and 0.6% by constraining the semantic consistency between branches using $\mathcal{L}_{ML}^{l}$.

Identity-guidance for Gait Learning. In Table 6a, we study how identity-guidance can help enhancing discriminative power of gait features. It can be seen that R-1 accuracy is improved by 1.9%/1.3% on VCCR/E-VCCR when adding the reprojection loss $\mathcal{L}_{rep}^{\hat{\Theta}}$, which is further boosted by adding the ID loss $\mathcal{L}_{ID}^{\hat{\Theta}}$.

6 Conclusion

In this paper, we tackle VCCRe-ID by proposing the Cross-Modality Complementary Learning framework, which fully exploits the cloth-invariant modalities in a collaborative manner. The proposed 3D Cross-Attention block enables the texture branch to mine body shape information when appearance cues are unreliable due to occlusion or illumination. 3D knowledge enables gait branch to address viewpoint variations in sequences of silhouette while preserving discriminative power for Re-ID. Texture and gait features are effectively learned and aggregated, shown by superiority of our framework over existing methods on all experimented VCCRe-ID datasets. Furthermore, we build the large-scale E-VCCR dataset which better suits real-world VCCRe-ID and provides a wide range of clothing variations and modalities, facilitating more active research in this important Re-ID task.

References

1. Arkushin, D., Cohen, B., Peleg, S., Fried, O.: Reface: Improving clothes-changing re-identification with face features. arXiv preprint arXiv:2211.13807 (2022)
2. Bansal, V., Micheloni, C., Foresti, G., Martinel, N.: Spatio-temporal attention for cloth-changing reid in videos. In: ECCVW. pp. 353–368 (2023)
3. Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., Grundmann, M.: Blazepose: On-device real-time body pose tracking. arXiv preprint arXiv:2006.10204 (2020)
4. Bhuiyan, A., Huang, J.X.: Stca: Utilizing a spatio-temporal cross-attention network for enhancing video person re-identification. Image and Vision Computing **123**, 104474 (2022)
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. pp. 7291–7299 (2017)
6. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: AAAI. vol. 33, pp. 8126–8133 (2019)
7. Chen, H., Wang, Y., Lagadec, B., Dantcheva, A., Bremond, F.: Joint generative and contrastive learning for unsupervised person re-identification. In: CVPR. pp. 2004–2013 (2021)
8. Chen, J., Jiang, X., Wang, F., Zhang, J., Zheng, F., Sun, X., Zheng, W.S.: Learning 3d shape feature for texture-insensitive person re-identification. In: CVPR. pp. 8142–8151 (2021). https://doi.org/10.1109/CVPR46437.2021.00805
9. Chen, J., Zheng, W.S., Yang, Q., Meng, J., Hong, R., Tian, Q.: Deep shape-aware person re-identification for overcoming moderate clothing changes. IEEE TMM **24**, 4285–4300 (2022). https://doi.org/10.1109/TMM.2021.3114539
10. Cho, Y.J., Yoon, K.J.: Pamm: Pose-aware multi-shot matching for improving person re-identification. IEEE TIP **27**(8), 3739–3752 (2018). https://doi.org/10.1109/TIP.2018.2815840
11. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
12. Cui, Z., Zhou, J., Peng, Y., Zhang, S., Wang, Y.: Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. IEEE TCSVT (2023). https://doi.org/10.1109/TCSVT.2023.3241988
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
14. Eom, C., Lee, W., Lee, G., Ham, B.: Disentangled representations for short-term and long-term person re-identification. IEEE TPAMI **44**(12), 8975–8991 (2022). https://doi.org/10.1109/TPAMI.2021.3122444
15. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., Yu, S.: Opengait: Revisiting gait recognition towards better practicality. In: CVPR. pp. 9707–9716 (June 2023)
16. Gu, X., Chang, H., Ma, B., Bai, S., Shan, S., Chen, X.: Clothes-changing person re-identification with rgb modality only. In: CVPR. pp. 1050–1059 (2022). https://doi.org/10.1109/CVPR52688.2022.00113
17. Gu, X., Chang, H., Ma, B., Zhang, H., Chen, X.: Appearance-preserving 3d convolution for video-based person re-identification). In: ECCV. p. 228–243 (2020)
18. Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M., Hu, S.M.: Attention mechanisms in computer vision: A survey. Computational visual media **8**(3), 331–368 (2022)

19. Gupta, A., Chellappa, R.: You can run but not hide: Improving gait recognition with intrinsic occlusion type awareness. In: WACV. pp. 5893–5902 (2024)
20. Han, K., Gong, S., Huang, Y., Wang, L., Tan, T.: Clothing-change feature augmentation for person re-identification. In: CVPR. pp. 22066–22075 (2023)
21. Han, K., Huang, Y., Gong, S., Huang, Y., Wang, L., Tan, T.: 3d shape temporal aggregation for video-based clothing-change person re-identification. In: ACCV. pp. 71–88 (2022)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (June 2016)
23. Hong, P., Wu, T., Wu, A., Han, X., Zheng, W.S.: Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In: CVPR. pp. 10508–10517 (2021). https://doi.org/10.1109/CVPR46437.2021.01037
24. Huang, Y., Wu, Q., Xu, J., Zhong, Y.: Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In: International Joint Conference on Neural Networks. pp. 1–8 (2019). https://doi.org/10.1109/IJCNN.2019.8851957
25. Huang, Y., Wu, Q., Xu, J., Zhong, Y., Zhang, Z.: Clothing status awareness for long-term person re-identification. In: ICCV. pp. 11875–11884 (2021). https://doi.org/10.1109/ICCV48922.2021.01168
26. Jin, X., He, T., Zheng, K., Yin, Z., Shen, X., Huang, Z., Feng, R., Huang, J., Chen, Z., Hua, X.S.: Cloth-changing person re-identification from a single image with gait prediction and regularization. In: CVPR. pp. 14258–14267 (2022). https://doi.org/10.1109/CVPR52688.2022.01388
27. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR. pp. 7122–7131 (2018)
28. Khaldi, K., Mantini, P., Shah, S.K.: Unsupervised person re-identification based on skeleton joints using graph convolutional networks. In: Image Analysis and Processing. pp. 135–146 (2022)
29. Khaldi, K., Nguyen, V.D., Mantini, P., Shah, S.: Unsupervised person re-identification in aerial imagery. In: WACV Workshops. pp. 260–269 (2024)
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
31. Kolotouros, N., Pavlakos, G., Black, M., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV. pp. 2252–2261 (2019)
32. Le, N., Pham, T., Do, T., Tjiputra, E., Tran, Q.D., Nguyen, A.: Music-driven group choreography. In: CVPR. pp. 8673–8682 (2023)
33. Li, D., Zhang, Z., Chen, X., Huang, K.: A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios. IEEE TIP **28**(4), 1575–1590 (2019). https://doi.org/10.1109/TIP.2018.2878349
34. Li, J., Zhang, S., Huang, T.: Multi-scale 3d convolution network for video based person re-identification. AAAI p. 8618–8625 (2019). https://doi.org/10.1609/aaai.v33i01.33018618
35. Li, S., Chen, H., Yu, S., He, Z., Zhu, F., Zhao, R., Chen, J., Qiao, Y.: Cocas+: Large-scale clothes-changing person re-identification with clothes templates. IEEE TCSVT **33**(4), 1839–1853 (2023). https://doi.org/10.1109/TCSVT.2022.3216769
36. Li, W., Hou, S., Zhang, C., Cao, C., Liu, X., Huang, Y., Zhao, Y.: An in-depth exploration of person re-identification and gait recognition in cloth-changing conditions. In: CVPR. pp. 13824–13833 (2023)

37. Li, Y.J., Weng, X., Kitani, K.M.: Learning shape representations for person re-identification under clothing change. In: WACV. pp. 2431–2440 (2021)
38. Li, Y., He, J., Zhang, T., Liu, X., Zhang, Y., Wu, F.: Diverse part discovery: Occluded person re-identification with part-aware transformer. In: CVPR. pp. 2897–2906 (2021). https://doi.org/10.1109/CVPR46437.2021.00292
39. Liu, C.T., Wu, C.W., Wang, Y.C.F., Chien, S.Y.: Spatially and temporally efficient non-local attention network for video-based person re-identification. arXiv preprint arXiv:1908.01683 (2019)
40. Liu, F., Ye, M., Du, B.: Dual level adaptive weighting for cloth-changing person re-identification. IEEE TIP **32**, 5075–5086 (2023). https://doi.org/10.1109/TIP.2023.3310307
41. Liu, X., Zhang, P., Yu, C., Lu, H., Yang, X.: Watching you: Global-guided reciprocal learning for video-based person re-identification. In: CVPR. pp. 13329–13338 (2021)
42. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM TOG **34**(6) (2015). https://doi.org/10.1145/2816795.2818013
43. Miao, J., Wu, Y., Liu, P., Ding, Y., Yang, Y.: Pose-guided feature alignment for occluded person re-identification. In: ICCV. pp. 542–551 (2019). https://doi.org/10.1109/ICCV.2019.00063
44. Nguyen, T.P., Pham, T.T., Nguyen, T., Le, H., Nguyen, D., Lam, H., Nguyen, P., Fowler, J., Tran, M.T., Le, N.: Embryosformer: Deformable transformer and collaborative encoding-decoding for embryos stage development classification. In: WACV. pp. 1981–1990 (2023)
45. Nguyen, V.D., Khaldi, K., Nguyen, D., Mantini, P., Shah, S.: Contrastive viewpoint-aware shape learning for long-term person re-identification. In: WACV. pp. 1041–1049 (2024)
46. Nguyen, V.D., Mantini, P., Shah, S.K.: Acml: Attention-based cross-modality learning for cloth-changing and occluded person re-identification. In: 2024 IEEE International Conference on Image Processing (ICIP). pp. 2396–2402 (2024). https://doi.org/10.1109/ICIP51287.2024.10647794
47. Nguyen, V.D., Mantini, P., Shah, S.K.: Contrastive clothing and pose generation for cloth-changing person re-identification. In: CVPRW. pp. 7541–7549 (2024)
48. Nguyen, V.D., Mantini, P., Shah, S.K.: Occluded cloth-changing person re-identification via occlusion-aware appearance and shape reasoning. In: 2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–8 (2024). https://doi.org/10.1109/AVSS61716.2024.10672564
49. Nguyen, V.D., Mantini, P., Shah, S.K.: Temporal 3d shape modeling for video-based cloth-changing person re-identification. In: WACV Workshops. pp. 173–182 (2024)
50. Nguyen, V.D., Mirza, S., Mantini, P., Shah, S.K.: Attention-based shape and gait representations learning for video-based cloth-changing person re-identification. In: VISIGRAPP (2: VISAPP). pp. 80–89 (2024)
51. Nguyen, V.D., Mirza, S., Zakeri, A., Gupta, A., Khaldi, K., Aloui, R., Mantini, P., Shah, S.K., Merchant, F.: Tackling domain shift in person re-identification: A survey and analysis. In: CVPRW. pp. 4149–4159 (2024)
52. 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., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library (2019)

53. Pham, T.T., Brecheisen, J., Nguyen, A., Nguyen, H., Le, N.: I-ai: A controllable & interpretable ai system for decoding radiologists' intense focus for accurate cxr diagnoses. In: WACV. pp. 7850–7859 (2024)
54. Qian, X., Wang, W., Zhang, L., Zhu, F., Fu, Y., Xiang, T., Jiang, Y.G., Xue, X.: Long-term cloth-changing person re-identification. In: ACCV. pp. 71–88 (2021)
55. Sarfraz, M.S., Schumann, A., Eberle, A., Stiefelhagen, R.: A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In: CVPR. pp. 420–429 (2017)
56. Sarker, P.K., Zhao, Q.: Enhanced visible–infrared person re-identification based on cross-attention multiscale residual vision transformer. Pattern Recognition **149**, 110288 (2024)
57. Shu, X., Li, G., Wang, X., Ruan, W., Tian, Q.: Semantic-guided pixel sampling for cloth-changing person re-identification. IEEE Signal Processing Letters **28**, 1365–1369 (2021). https://doi.org/10.1109/lsp.2021.3091924
58. Song, Y.F., Zhang, Z., Shan, C., Wang, L.: Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In: ACM MM. p. 1625–1633 (2020). https://doi.org/10.1145/3394171.3413802
59. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR. pp. 5693–5703 (2019)
60. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: ECCV. p. 501–518 (2018)
61. Sun, Y., Bao, Q., Liu, W., Fu, Y., Black, M.J., Mei, T.: Monocular, one-stage, regression of multiple 3d people. In: ICCV. pp. 11159–11168 (2021). https://doi.org/10.1109/ICCV48922.2021.01099
62. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: GaitGraph: Graph convolutional network for skeleton-based gait recognition. In: ICIP. pp. 2314–2318 (2021). https://doi.org/10.1109/ICIP42928.2021.9506717
63. Trinh, Q.H., Bui, N.T., Hoang, D.H., Thi, P.T.V., Nguyen, H.D., Jha, D., Bagci, U., Le, N., Tran, M.T.: Pgds: Pose-guidance deep supervision for mitigating clothes-changing in person re-identification. In: 2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). pp. 1–8 (2024). https://doi.org/10.1109/AVSS61716.2024.10672607
64. Wan, F., Wu, Y., Qian, X., Chen, Y., Fu, Y.: When person re-identification meets changing clothes. In: CVPRW. pp. 3620–3628 (2020)
65. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. In: ACM MM. p. 274–282 (2018). https://doi.org/10.1145/3240508.3240552
66. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. IEEE TPAMI **43**(10), 3349–3364 (2021). https://doi.org/10.1109/TPAMI.2020.2983686
67. Wang, L., Zhang, X., Han, R., Yang, J., Li, X., Feng, W., Wang, S.: A benchmark of video-based clothes-changing person re-identification. arXiv preprint arXiv:2211.11165 (2022)
68. Wang, Q., Qian, X., Fu, Y., Xue, X.: Co-attention aligned mutual cross-attention for cloth-changing person re-identification. In: ACCV. pp. 2270–2288 (2022)
69. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR. pp. 7794–7803 (2018)

70. Wang, Y., Zhang, P., Gao, S., Geng, X., Lu, H., Wang, D.: Pyramid spatial-temporal aggregation for video-based person re-identification. In: ICCV. pp. 12006–12015 (2021). https://doi.org/10.1109/ICCV48922.2021.01181
71. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: CVPR. pp. 79–88 (2018). https://doi.org/10.1109/CVPR.2018.00016
72. Wu, Y., Bourahla, O.E.F., Li, X., Wu, F., Tian, Q., Zhou, X.: Adaptive graph representation learning for video person re-identification. IEEE TIP **29**, 8821–8830 (2020)
73. Xian, Y., Yang, J., Yu, F., Zhang, J., Sun, X.: Graph-based self-learning for robust person re-identification. In: WACV. pp. 4789–4798 (2023)
74. Xu, P., Zhu, X.: Deepchange: A long-term person re-identification benchmark with clothes change. In: ICCV. pp. 11196–11205 (2023)
75. Xu, W., Liu, H., Shi, W., Miao, Z., Lu, Z., Chen, F.: Adversarial feature disentanglement for long-term person re-identification. In: IJCAI. pp. 1201–1207 (2021). https://doi.org/10.24963/ijcai.2021/166
76. Yang, Q., Wu, A., Zheng, W.S.: Person re-identification by contour sketch under moderate clothing change. IEEE TPAMI **43**(6), 2029–2046 (2021). https://doi.org/10.1109/tpami.2019.2960509
77. Yang, Z., Lin, M., Zhong, X., Wu, Y., Wang, Z.: Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In: CVPR. pp. 1472–1481 (2023). https://doi.org/10.1109/CVPR52729.2023.00148
78. Yang, Z., Zhong, X., Zhong, Z., Liu, H., Wang, Z., Satoh, S.: Win-win by competition: Auxiliary-free cloth-changing person re-identification. IEEE TIP **32**, 2985–2999 (2023). https://doi.org/10.1109/TIP.2023.3277389
79. Yu, S., Li, S., Chen, D., Zhao, R., Yan, J., Qiao, Y.: Cocas: A large-scale clothes changing person dataset for re-identification. In: CVPR. pp. 3397–3406 (2020). https://doi.org/10.1109/CVPR42600.2020.00346
80. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. vol. 97, pp. 7354–7363 (2019)
81. Zhang, P., Wu, Q., Xu, J., Zhang, J.: Long-term person re-identification using true motion from videos. In: WACV. pp. 494–502 (2018). https://doi.org/10.1109/WACV.2018.00060
82. Zhang, P., Xu, J., Wu, Q., Huang, Y., Ben, X.: Learning spatial-temporal representations over walking tracklet for long-term person re-identification in the wild. IEEE TMM **23**, 3562–3576 (2021). https://doi.org/10.1109/TMM.2020.3028461
83. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: CVPR. pp. 4320–4328 (2018)
84. Zhang, Z., Tran, L., Yin, X., Atoum, Y., Liu, X., Wan, J., Wang, N.: Gait recognition via disentangled representation learning. In: CVPR. pp. 4705–4714 (2019). https://doi.org/10.1109/CVPR.2019.00484
85. Zheng, J., Liu, X., Liu, W., He, L., Yan, C., Mei, T.: Gait recognition in the wild with dense 3d representations and a benchmark. In: CVPR. pp. 20196–20205 (2022). https://doi.org/10.1109/CVPR52688.2022.01959
86. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: ECCV (2016)
87. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV. pp. 1116–1124 (2015). https://doi.org/10.1109/ICCV.2015.133

88. Zheng, Z., Wang, X., Zheng, N., Yang, Y.: Parameter-efficient person re-identification in the 3d space. IEEE Transactions on Neural Networks and Learning Systems pp. 1–14 (2022). https://doi.org/10.1109/tnnls.2022.3214834
89. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. In: CVPR. pp. 2133–2142 (2019)
90. Zhihui, Z., Jiang, X., Zheng, F., Guo, X., Huang, F., Sun, X., Zheng, W.: Viewpoint-aware loss with angular regularization for person re-identification. AAAI **34**, 13114–13121 (2020). https://doi.org/10.1609/aaai.v34i07.7014
91. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: ICCV. pp. 3701–3711 (2019). https://doi.org/10.1109/ICCV.2019.00380