Contrastive Viewpoint-aware Shape Learning for Long-term Person Re-Identification

Vuong D. Nguyen[1] Khadija Khaldi[1] Dung Nguyen[2] Pranav Mantini[1] Shishir Shah[1]
[1]University of Houston [2]Hanoi University of Science and Technology
{dnguyen170, kkhaldi, sshah5}@uh.edu dungnt.samihust@gmail.com pmantini@cs.uh.edu

Abstract

*Traditional approaches for Person Re-identification (Re-ID) rely heavily on modeling the appearance of persons. This measure is unreliable over longer durations due to the possibility for changes in clothing or biometric information. Furthermore, viewpoint changes significantly degrade the matching ability of these methods. In this paper, we propose "**C**ontrastive **V**iewpoint-aware **S**hape **L**earning for Long-term Person Re-Identification" (CVSL) to address these challenges. Our method robustly extracts local and global texture-invariant human body shape cues from 2D pose using the Relational Shape Embedding branch, which consists of a pose estimator and a shape encoder built on a Graph Attention Network. To enhance the discriminability of the shape and appearance of identities under viewpoint variations, we propose Contrastive Viewpoint-aware Losses (CVL). CVL leverages contrastive learning to simultaneously minimize the intra-class gap under different viewpoints and maximize the inter-class gap under the same viewpoint. Extensive experiments demonstrate that our proposed framework outperforms state-of-the-art methods on long-term person Re-ID benchmarks.*

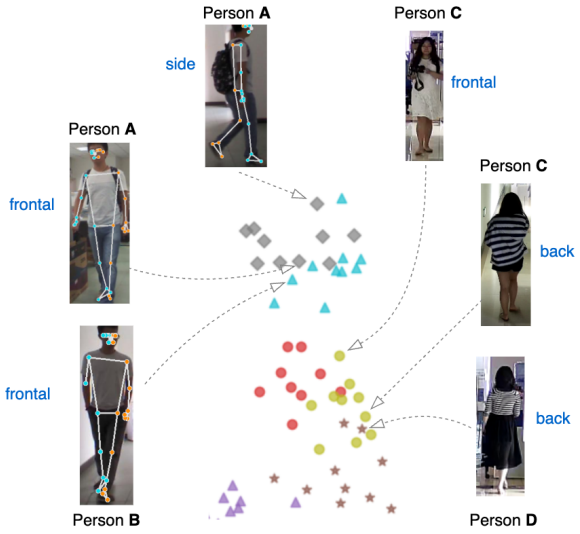

Figure 1. Features distribution on latent space visualized by t-SNE using model trained without contrastive viewpoint-aware losses. Features of different persons are not separable, showing large intra-class and small inter-class variations which degrades Re-ID accuracy. This forms the motivation for the proposed approach.

1. Introduction

Person Re-Identification (Re-ID) has emerged as a critical task in video surveillance applications that involves matching persons across multiple non-overlapping cameras. Traditional methods involve feature extraction [23, 32] and metric learning [20, 22], while recent approaches adopt deep learning techniques [21, 28]. These methods focus on encoding appearance features and achieve notable results on standard Re-ID datasets such as Market-1501 [38] and CUHK03 [18]. However, their performance is severely affected in two cases: (1) the person changes appearance by changing clothes, hairstyle, or covering the face (2) different persons wear similar clothing. Such shortcomings lead to poor Re-ID performance in real-world situations and necessitates a Long-term Person Re-ID (LRe-ID) approach that can deal with scenarios where appearance may become less reliable.

To address LRe-ID, cloth-invariant features that rely on face, hairstyle [7], or human body shape [11, 19] have been explored, as they tend to remain relatively unchanged over a longer period of time. However, explicitly extracting geometric cues from images is difficult due to pose variations or occlusions. Therefore, several LRe-ID methods rely on alternative modalities like 2D human posture keypoints [26], contour sketches [35], or gaits [15]. However, these works suffer from two major limitations. First, identity-relevant local shape features have not been well captured, leading to coarse-grained shape representation that has limited discriminability for re-identification. Second, these ap-

proaches overlook the impact of camera viewpoint on texture and shape information. This causes severe confusion in matching when different persons wear similar clothing under occluded viewpoints.

In this work, we propose "**C**ontrastive **V**iewpoint-aware **S**hape **L**earning for Long-term Person Re-Identification" (CVSL) to overcome these challenges. Our framework aims to extract texture-invariant body shape cues, which can effectively represent persons in long-term by mitigating the influence of clothing. To achieve this, we incorporate Relational Shape Embedding (RSE) branch and train our framework using contrastive viewpoint-aware losses (CVL). RSE leverages 2D skeleton-based human postures to encode global body shape semantic information and exploit implicit local correlations between body parts, which is lightweight and robust in long-term scenarios compared to other modalities. RSE implements a refinement network and a Graph Attention Network (GAT) to effectively capture the shape representation of a person that is invariant to variations in clothing. Appearance remains a competitive cue when persons slightly change clothes, so we couple shape with appearance features for final representation.

As shown in Figure 1, persons A and B under the same viewpoint share similar skeleton-based shape, which results in close features of different persons and dispersive features of the same person on latent space. Mismatching also happens when persons C and D wear similar clothing. In this work, to enhance discriminability of shape and appearance cues under viewpoint variations, we introduce contrastive viewpoint-aware losses. These losses leverage contrastive learning to simultaneously minimize the intra-class gap under different viewpoints and maximize the inter-class gap under the same viewpoint. By incorporating viewpoint guidance, our method effectively learns to differentiate individuals under viewpoint changes, thus improving the robustness of person Re-ID systems in real-world scenarios where surveillance cameras can capture images of individuals from different angles.

The key contributions of our work can be summarized as follows: (1) we propose a strong baseline that jointly learns body shape embedding and appearance features under moderate texture variations; (2) we propose contrastive viewpoint-aware losses to enhance discriminability of shape and appearance under viewpoint changes; and (3) we present extensive experiment results, which demonstrate that our method significantly outperforms state-of-the-arts on LRe-ID benchmarks.

2. Related Works

2.1. Person Re-Identification (Re-ID)

Several approaches have been proposed for person Re-ID including feature extraction [23, 28, 32] or distance metric learning [20, 22]. Challenging factors that affect feature learning in ReID such as pose [5, 27] and occlusion [24, 41] have also been considered to reduce spatial misalignment. However, these methods rely substantially on the assumption of consistent clothing over long-term, which is not practical in real-world scenarios.

2.2. Long-term Person Re-ID (LRe-ID)

LRe-ID has not been widely studied and there exists few published datasets [26, 31, 34, 35] to help in the design and development of novel approaches. Qian *et al.* [26] proposed to extract shape embedding by a cloth-elimination shape-distillation module. [11, 19] utilized shape cues to support feature learning. However, these works overlooked the local structural cues of human. Self-attention [2], regularization [15] and clothes-based losses [7] are applied in LRe-ID frameworks to attend to cloth-irrelevant features like face and hairstyle. Recent works [8, 13, 36] propose to utilize clothing status and labels for augmenting Re-ID features in the latent space. As they solely rely on texture, occluded viewpoints significantly limit these models' ability for re-identification. To enhance robustness against such challenges, our LRe-ID framework explicitly captures human structural cues while further boosting the appearance feature learning for an accurate Re-ID model. Additionally, we fully leverage the spatial relationships between body parts to learn more discriminative shape feature representation.

2.3. Viewpoint-aware Person Re-ID

Viewpoint changes severely affect appearance and shape of a person. Sun *et al.* [27] leveraged viewpoints to determine the level of gradient update in back propagation. However, relying solely on texture information brought by frontal viewpoint is not applicable in LRe-ID. VTM [39] proposed a view transform feature extraction method. Inspired by VTM, Zhu *et al.* [40] proposed a method to cluster persons of the same identity based on a viewpoint-aware hyper-sphere. Multi-feature fusion frameworks driven by viewpoint-aware loss were designed in [1, 14]. Contrastive learning has been applied in traditional Re-ID task [16]. In this paper, we propose contrastive viewpoint-aware losses, which leverage contrastive learning to deal with viewpoint variations. These losses aim to learn a latent space in which features of the same person under different viewpoints are pulled closer while features of different persons under the same viewpoint are pushed away from each other.

3. Methodology

3.1. Framework and Notation

The overview of our proposed framework is shown in Figure 2. In each training batch, the framework takes in a set of N RGB images $X = \{x_i\}_{i=1}^{N}$. Relational Shape Em-

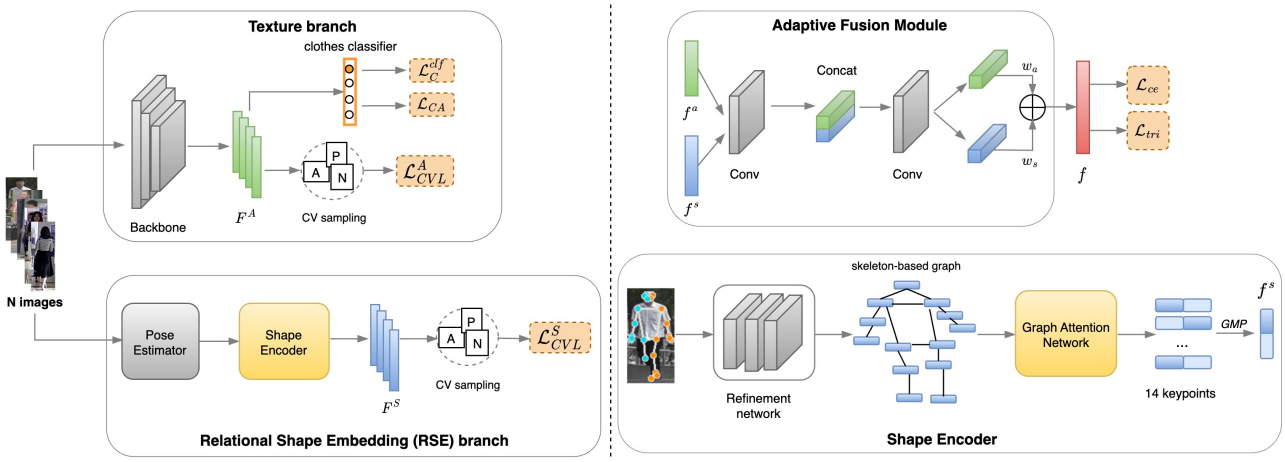

Figure 2. Overall model architecture (left), Adaptive Fusion module (upper right) and Shape Encoder (lower right). In a training batch, given N images, RSE branch outputs shape feature set F^S, while Texture branch outputs appearance feature set F^A. F^S and F^A are then fed into contrastive viewpoint-aware sampling module to obtain inputs for the proposed contrastive viewpoint-aware losses (CVL).

bedding (RSE) branch first estimates skeleton-based pose for each image using a pose estimator and then outputs a set of shape feature vectors $F^S = \{f_i^s\}_{i=1}^N$ using a shape encoder built on a Graph Attention Network. Appearance still provides valuable semantic information about the person. Therefore, we utilize a CNN backbone to obtain a set of appearance feature vectors $F^A = \{f_i^a\}_{i=1}^N$. Feature sets F^S and F^A are then used to sample inputs for the proposed contrastive viewpoint-aware losses $\mathcal{L}_{CVL}^S$ and $\mathcal{L}_{CVL}^A$. Appearance and shape features are then fused by Adaptive Fusion module to obtain final representation sets $F = \{f_i\}_{i=1}^N$.

Given X, we denote the corresponding identity label set as $Y^{ID} = \{y_i^{ID}\}_{i=1}^N$, clothes label set as $Y^C = \{y_i^C\}_{i=1}^N$ and estimated viewpoint label set as $Y^V = \{y_i^V\}_{i=1}^N$, where $x_i \in \mathbb{R}^{3\times h\times w}$, $y_i^{ID} \in \mathbb{N}$, $y_i^C \in \mathbb{N}$, and $y_i^V \in \mathbb{N}$. The total number of clothing classes is the cumulative number of suits worn by all individuals in the training set.

3.2. Relational Shape Embedding (RSE) branch

Body shape is an important cue in LRe-ID. It remains relatively stable and can be used to distinguish persons from one another, unlike clothing which can vary significantly in color, style, and pattern. Shape refers to the geometric form of the human body and can be described using the skeleton representation. We incorporate body shape information and improve the accuracy and robustness of our Re-ID framework for long-term scenarios.

3.2.1 Pose Estimation

RSE employs OpenPose [3], which is an off-the-shelf pose estimator that has achieved decent performance for pose estimation. Given an image $x \in X$, OpenPose outputs a set $J = \{j_i\}_{i=1}^k$ of k body joints, which is used to represent body pose of the person in the image x. Each joint is represented by the relative position of the joint in the input image, i.e. a set of two coordinates (x_i, y_i) indicating the position of the pixel corresponding to the location of the joint.

3.2.2 Shape Encoder

Shape embeddings are extracted using an attention-based encoder with the architecture shown in Figure 2. Given (w, h) as the original width and height of the image, the coordinates of all joints in J are then normalized, giving the representation of joint $j_i = \left(\frac{x_i}{w}, \frac{y_i}{h}, \frac{w}{h}\right), i = 1, ..., k$. We then refine the body joints representation set J by passing them through a refinement network $\mathcal{R}(\cdot)$. The refinement network $\mathcal{R}(\cdot)$ is designed to output a feature vector for each keypoint to capture the fine-grained details of the person's body shape. As shown in Figure 2, our refinement network $\mathcal{R}(\cdot)$ consists of a sequence of fully connected layers. $\mathcal{R}(\cdot)$ refines joint $j_i \in J$ by:

$$\hat{j}_i = \mathcal{R}(j_i) \tag{1}$$

where $\hat{j}_i \in \mathbb{R}^d$ is the higher-dimensional refined representation of j_i and d is the dimension of the last layer in $\mathcal{R}$.

Intuitively, the feature of a single joint is not sufficient to capture the information of body shape. We amplify the relations between pairs of keypoints to capture local geometric features of a person's body. To this end, we propose a Graph Attention Network (GAT) [30] to exploit local relationships between body parts and obtain a more discriminative shape embedding of the person, especially when input image suffers from severe occlusion and global shape can not be fully captured. GAT is a type of Graph Convolutional Networks (GCN) that operates on graphs and is designed

to learn features that are aggregated across neighborhoods in a graph using message passing. GAT employs attention mechanism [29] to perform aggregation and updating for several graph attention layers, allowing the network to capture high-order relationships between keypoints.

Specifically, as shown in Figure 2, the refined keypoint feature set $\hat{J} = \{\hat{j}_i\}_{i=1}^{k}$ are then used to construct a graph that represents the person's body shape. Each keypoint is treated as a node in the graph, and edges are added between nodes that are connected in a body skeleton (i.e. elbow and shoulder, torso and knee, etc.). Our GAT $\mathcal{G}$ consists of L graph attention layers $\mathcal{G}^{(l)}, l = 0, ..., L-1$. Consider a node $\hat{j}_i \in \hat{J}$, then the l^{th} layer $\mathcal{G}^{(l)}$ operates on $\hat{j}_i$ by aggregating features and updating $\hat{j}_i$ using set $\mathcal{N}_i$ containing indices of neighboring nodes of $\hat{j}_i$, given as:

$$\hat{j}_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}_i} \mathbf{M}_{ij} \mathbf{W}^{(l)} \hat{j}_j^{(l)}\right), \tag{2}$$

where $\mathbf{M} = (\alpha_{ij}), \mathbf{M} \in \mathbb{R}^{k \times k}$, α_{ij} specifies the weight between $\hat{j}_i$ and $\hat{j}_j$ (i.e. the importance of joint $\hat{j}_j$ to joint $\hat{j}_i$), σ is an activation function and $\mathcal{N}_i$ can be defined via adjacency matrix. Denote $d^{(l)}$ as the dimension of node-level feature vector at layer $\mathcal{G}^{(l)}$ (i.e. $d^{(0)} = d$), then $\mathbf{W}^{(l)} \in \mathbb{R}^{d^{(l+1)} \times d^{(l)}}$ is the weight matrix of layer $\mathcal{G}^{(l)}$. Note that unlike in traditional GCN where $\mathbf{M}$ is explicitly defined, GAT $\mathcal{G}$ implicitly computes $\alpha_{ij} \in \mathbf{M}$ by:

$$\alpha_{ij} = \text{softmax}_j\left(h\left(\mathbf{W}\hat{j}_i, \mathbf{W}\hat{j}_j\right)\right), \tag{3}$$

where $h : \mathbb{R}^{d^{(l+1)}} \times \mathbb{R}^{d^{(l+1)}} \rightarrow \mathbb{R}$ is a byproduct of an attentional mechanism. GAT trivially attends over neighborhoods and implicitly amplifies importance of each joint to its different neighbors, thus enables our shape encoder to exploit local relations between body parts. Finally, after L graph attention layers, a global max pooling layer is employed to aggregate the high-order representations of joint set $\hat{J}^{(L-1)} = \left\{\hat{j}_i^{(L-1)}\right\}_{i=1}^{N}$, producing a fixed-length vector that summarizes information from the skeleton graph:

$$f^s = \text{GMP}\left(\hat{J}^{(L-1)}\right), \tag{4}$$

where GMP denotes global max pooling, f^s is the image-wise global shape representation.

3.3. Texture branch

The texture branch consists of a CNN backbone $\mathcal{F}_\theta(.)$ with parameters θ and a clothes classifier $\mathcal{C}_\phi^{clf}$ with parameters ϕ. Given training batch of images X, we first extract global appearance feature f_i^a of image $x_i \in X$ by $f_i^a = \mathcal{F}_\theta(x_i)$. Then, we capture local clothes-invariant information and couple it with the body shape feature to enhance the global representation [7]. This also helps enhance the model's ability to distinguish different identities wearing similar clothing. Specifically, predicted clothes class $\hat{y}_i^C$ of input image x_i is output by clothes classifier: $\hat{y}_i^C = \mathcal{C}_\phi^{clf}(f_i^a)$. Then, given clothes label $c = y_i^C$ and the total number of clothes classes N_C, clothes classification loss $\mathcal{L}_C$ based on cross entropy loss is then employed to optimize clothes classifier

$$\mathcal{L}_C^{clf} = -\sum_{i=1}^{N} \log \frac{e^{(f_i^a \cdot \phi_c/\tau)}}{\sum_{j=1}^{N_C} e^{(f_i^a \cdot \phi_j/\tau)}}, \tag{5}$$

where $\tau \in \mathbb{R}^+$ is a temperature parameter. Then, texture branch is able to highlight clothes-irrelevant features via clothes-based adversarial loss $\mathcal{L}_{CA}$, which penalizes the predictive power of the Re-ID model with respect to different clothes of the same identity, given as:

$$\mathcal{L}_{CA} = -\sum_{i=1}^{N}\sum_{c=1}^{N_C} s(c) \log \frac{e^{(f_i^a \cdot \phi_c/\tau)}}{e^{(f_i^a \cdot \phi_c/\tau)} + \sum_{j \in C_i^-} e^{(f_i^a \cdot \phi_j/\tau)}}, \tag{6}$$

where $s(c)$ denotes the cross entropy loss for c^{th} clothes class. Minimizing clothes-based loss $\mathcal{L}_C = \mathcal{L}_C^{clf} + \mathcal{L}_{CA}$ forces the backbone to output fine-grained global appearance by coupling clothes-irrelevant features such as face and hairstyle with human geometric cues.

3.4. Contrastive Viewpoint-aware Losses

Figure 1 demonstrates the effect of viewpoint changes and similar clothing on re-identification. Although body shape is a stable cue which represents the geometric characteristic of a person in the long term, different viewpoints result in highly dissimilar skeleton-based body shape, which leads to small inter-class variation and large intra-class variation. Texture branch faces a similar issue and struggles to identify different identities being captured from the same viewpoint and wearing similar clothing. To address these issues, we propose two contrastive viewpoint-aware losses, $\mathcal{L}_{CVL}^S$ and $\mathcal{L}_{CVL}^A$, to guide the training of the RSE branch and texture branch for more discriminative embeddings.

Specifically, inputs for $\mathcal{L}_{CVL}^S$ are contrastively sampled based on viewpoints as follows: for each shape feature vector $f_i^s \in F^S$ as anchor, if anchor is captured under frontal or back viewpoint, images of different identities but same viewpoint are chosen as positive samples, while images of the same identity but side viewpoint are considered negative samples. The reason is frontal and back viewpoints bring similar skeleton-based shape, while side viewpoint brings highly dissimilar shape. The proposed $\mathcal{L}_{CVL}^S$ loss is then able to pulls the shape feature vectors of the same identity under different viewpoints closer in the latent space. Denoting $y_i^V = \{1, 0, -1\}$ corresponding to frontal, side, and

back viewpoint of input image x_i, $\mathcal{L}_{CVL}^{S}$ is formulated as:

$$\mathcal{L}_{CVL}^{S} = -\sum_{i=1}^{N} \log \frac{\sum_{j \in S_i^+} e^{d\left(f_i^s, f_j^s\right)/\tau}}{\sum_{k \in S_i^-} e^{d\left(f_i^s, f_k^s\right)/\tau}}, \tag{7}$$

where $S_i^+ = \left\{j \in [1, ..., N] \,|\, y_j^{ID} = y_i^{ID}, y_j^V = y_i^V \pm 1\right\}$ and $S_i^- = \left\{k \in [1, ..., N] \,|\, y_k^{ID} \neq y_i^{ID}, y_k^V = y_i^V\right\}$ are the indices sets of positive and negative samples, and $d(\cdot, \cdot)$ denotes the cosine distance.

For texture branch, we first mitigate the influence of viewpoint variations by proposing $\mathcal{L}_V^A$ where we consider images of same identity and different viewpoint as positive samples and images of different identity and same viewpoint as negative samples. This helps pushing appearance feature vectors of different identities under the same viewpoint farther in the latent space. Given positive set $A_i^+ = \left\{j \in [1, ..., N] \,|\, y_j^{ID} = y_i^{ID}, y_i^V \neq y_j^V\right\}$ and negative set $A_i^- = \left\{k \in [1, ..., N] \,|\, y_k^{ID} \neq y_i^{ID}, y_k^V = y_j^V\right\}$, $\mathcal{L}_V^A$ has the formulation as:

$$\mathcal{L}_V^A = -\sum_{i=1}^{N} \log \frac{\sum_{j \in A_i^+} e^{d\left(f_i^a, f_j^a\right)/\tau}}{\sum_{k \in A_i^-} e^{d\left(f_i^a, f_k^a\right)/\tau}}. \tag{8}$$

In this work, we further deal with persons wearing similar clothes by proposing hard-mining triplet loss $\mathcal{L}_{sim}^A$. In each training step, for every $f_i^a \in F^A$, we compute the pairwise distance between f_i^a and every other feature vector in F^A. Then, by considering that the vector that is most similar to the anchor is likely to share similar clothing, triplet is chosen as follows: vector that has the largest (smallest) distance to f_i^a, with the same (different) identity is the positive sample $f_i^{a,p}$ and negative sample $f_i^{a,n}$, respectively. Given the sampled triplet, $\mathcal{L}_{sim}^A$ is formulated as:

$$\mathcal{L}_{sim}^A = \sum_{i=1}^{N} \left(d\left(f_i^a, f_i^{a,p}\right) + \max\left\{0, m - d\left(f_i^a, f_i^{a,n}\right)\right\}\right), \tag{9}$$

where m is margin parameter that controls the separation between the positive and negative pairs. In this work, m is set to 0.3. Finally, we have $\mathcal{L}_{CVL}^A$ that drives the training of texture branch as $\mathcal{L}_{CVL}^A = \mathcal{L}_V^A + \mathcal{L}_{sim}^A$.

3.5. Adaptive Fusion Module (AFM)

We propose to fuse shape feature f_i^s and appearance feature f_i^a by AFM as shown in Figure 2. Specifically, features are first transformed and projected onto a common latent space by a convolutional layer:

$$\mathbf{f}_i^s = \sigma(\text{Conv}(\|f_i^s\|_2), \mathbf{f}_i^a = \sigma(\text{Conv}(\|f_i^a\|_2), \tag{10}$$

where $\sigma(.)$ denotes sigmoid activation function. Then, to amplify the contribution of each feature to the final representation, we force them to optimize each other by concatenating two vectors and feeding them into the corresponding convolutional layers Conv_s and Conv_a which estimate corresponding weights $\mathbf{w}_i^s$ and $\mathbf{w}_i^a$. Global representation f_i of input image x_i is obtained by $f_i = \mathbf{w}_i^s \odot \mathbf{f}_i^s + \mathbf{w}_i^a \odot \mathbf{f}_i^a$. Batch feature set $F = \{f_i\}_{i=1}^N$ serves as input to identification loss $\mathcal{L}_{ID}$, which is the sum of a cross-entropy-based classification loss $\mathcal{L}_{ce}$ and pair-wise triplet loss $\mathcal{L}_{tri}$, i.e. $\mathcal{L}_{ID} = \mathcal{L}_{ce} + \mathcal{L}_{tri}$. Finally, the overall CVSL model is trained by the total loss:

$$\mathcal{L} = \mathcal{L}_{CVL}^S + \mathcal{L}_{CVL}^A + \mathcal{L}_C + \mathcal{L}_{ID}. \tag{11}$$

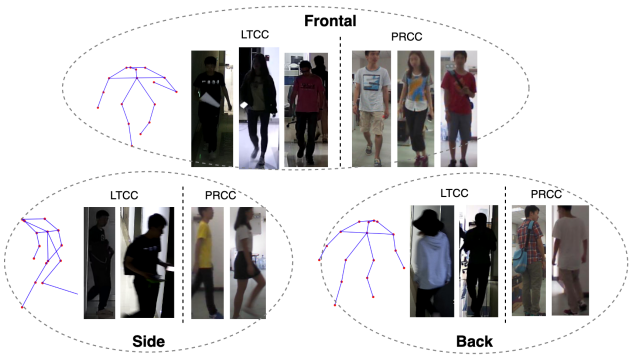

Figure 3. Samples images from LTCC [26] and PRCC [35] datasets. Viewpoints are clustered into three groups: frontal, side and back. PRCC is collected under good lighting condition and no occlusion, while LTCC poses severe challenges for Re-ID.

4. Experimental Setup

4.1. Datasets and Evaluation Protocol

Two large-scale LRe-ID datasets including LTCC and PRCC are used for experiments in our work. **LTCC** [26] is an indoor cloth-changing person Re-ID dataset which has 152 identities. Each person has 2 to 14 outfits, in total 478 different outfits were captured from 12 camera views. **PRCC** [35] dataset consists of images from 221 identities. Each person in Cameras A and B is wearing the same clothes, but the images are captured in different rooms. For Camera C, the persons wear different clothes, and the images are captured on a different day. Samples from three viewpoints from the two datasets are visualized in Figure 3, which shows a large variation in image quality between the two datasets.

We utilize mean average precision (mAP) and rank-1 accuracy to evaluate the performance of our model. Following the evaluation procedures in [26, 35], we validate our model in **cloth-changing** setting where only cloth-changing samples are used for evaluation. We also report results in **standard** setting, where for PRCC, the test set consists of only cloth-consistent samples, while for LTCC, the test set includes both cloth-consistent and cloth-changing samples.

Methods	Modalities	LTCC				PRCC			
		Cloth-changing		Standard		Cloth-changing		Standard	
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
PCB [28]	RGB	23.5	10.0	65.1	30.6	41.8	38.7	99.8	97.0
RGA-SC [37]	RGB	31.4	14.0	65.0	27.5	42.3	-	98.4	-
RCSANet [12]	RGB	-	-	-	-	48.6	50.2	100	97.2
CAL [7]	RGB	40.1	18.1	74.2	40.8	55.2	55.8	**100**	**99.8**
PRCC-contour [35]	RGB + sketch	-	-	-	-	34.4	-	64.2	-
GI-ReID [15]	RGB + sil	23.7	10.4	63.2	29.4	33.3	-	80.0	-
CESD [26]	RGB + pose	26.1	12.4	71.4	34.3	-	-	-	-
FSAM [11]	RGB + pose + sil	38.5	16.2	73.2	35.4	54.5	-	98.8	-
CASE-Net [19]	RGB + pose	-	-	-	-	39.5	-	71.2	-
3DSL [4]	RGB + 3D pose + sil	31.2	14.8	-	-	51.3	-	-	-
CVSL (Ours)	RGB + pose	**44.5**	**21.3**	**76.4**	**41.9**	**57.5**	**56.9**	97.5	99.1

Table 1. Quantitative results comparison between CVSL and state-of-the-arts (SOTAs) on LTCC and PRCC datasets. Overall, our framework outperforms SOTAs on both datasets in cloth-changing setting, demonstrating the effectiveness of CVSL in real-world scenarios.

4.2. Implementation Details

Our work is implemented in PyTorch [25]. To estimate viewpoint, we leveraged MEBOW [33] and directly performed inference on training set using provided pretrained model. We choose ResNet-50 [9] initialized with weights pretrained on ImageNet [6] as CNN backbone for texture branch. For pose estimation, we employed OpenPose [3] to obtain 19 keypoints in COCO format. As we do not need specific features from nose, eyes, and ears, we averaged 5 keypoints from face as one point, leading to 14 keypoints in total. Refinement network consists of 3 linear layers of $[128, 512, 2048]$ neurons respectively, while GAT consists of two layers. During training, the images are resized to 256×128. Horizontal flipping and random erasing are applied for data augmentation. Batch size is 32 where 8 identities and 4 images per identity are sampled. Adam [17] optimizer is used with initial learning rate of $5e{-}4$, momentum of 0.9 and weight decay factor for L2 regularization of $1e-6$. Learning rate is reduced by a factor of 0.1 after every 30 epochs. τ is set to $1/16$. Our model was trained on a single NVIDIA GeForce GTX 1080 16GB RAM GPU for a total of 80 epochs, which took around 4 hours.

5. Results and Ablation Study

5.1. Results

Quantitative results of our proposed CVSL framework is reported in Table 1. We compared CVSL with existing methods, categorized by modalities utilized, including techniques based on RGB modalities only (i.e. PCB [28], RGA-SC [37], RCSANet [12] and CAL [7]), RGB with parsing (i.e. PRCC-contour [35] and GI-ReID [15]), and RGB with pose (i.e. CESD [26], FSAM [11], CASE-Net [19] and 3DSL [4]). Except for traditional methods PCB [28] and RGA-SC [37], the remaining are specifically designed for LRe-ID. Overall, CVSL outperforms current state-of-the-art methods in both cloth-changing and standard settings on LTCC and in cloth-changing setting on PRCC. Since PRCC contains images which are clear and captured under good lighting condition (Figure 3), results are saturated in same-clothes (standard) setting using texture-based models. On the other hand, on LTCC, compared to methods using RGB modalities only, our CVSL framework shows superiority by coupling human geometric features with texture information which can be severely affected by lighting condition and viewpoints in reality. Compared to other LRe-ID methods that also utilize auxiliary cues like pose and silhouette, our method significantly outperforms them by effectively capturing local relations between body parts, which is more stable in long-term and not affected by occluded viewpoints in certain cases like global shape. We also effectively mitigate the influence of viewpoint variations, resulting in a more robust model.

Methods	LTCC		PRCC	
	R-1	mAP	R-1	mAP
Texture, w/o CVL	38.9	16.5	45.7	41.1
Texture, w/ CVL	40.1	17.2	46.5	43.4
RSE, w/o CVL	32.1	14.2	42.2	39.2
RSE, w/ CVL	35.2	14.9	43.1	41.9
Joint, w/o CVL	42.1	20.1	53.6	51.1
Joint, w/ Triplet loss	42.3	20.5	54.1	51.4
CVSL (Ours)	**44.5**	**21.3**	**57.5**	**56.9**

Table 2. Ablation studies of (1) the RSE branch and (2) the contrastive viewpoint-aware losses (CVL) on LTCC and PRCC datasets in cloth-changing setting only.

5.2. Ablation Study

We carry out ablation study on the two key components of our proposed CVSL framework: (1) the effectiveness of

shape features produced by RSE branch when being coupled with appearance features from texture branch; and (2) the effectiveness of the proposed contrastive viewpoint-aware losses $\mathcal{L}^S_{CVL}$ and $\mathcal{L}^A_{CVL}$. Overall, RSE branch effectively extracts clothing-invariant body shape cues that can be coupled with appearance for long-term person representations. CVL also significantly boosts the accuracy of the Re-ID framework.

5.2.1 RSE branch

Table 2 reports quantitative comparison results of appearance, shape, and joint representation. It can be seen that models with texture branch achieve higher performance than models with RSE branch on both datasets. The reasons are two-fold. First, in cases of slight cloth-changing or cloth-consistence, appearance remains a competitive features for re-identification. In those cases, body shape becomes less competitive than exploiting visual similarities from persons. Second, discriminability of shape from RSE branch relies on the accuracy of pose estimator which may struggle to produce accurate poses from low quality and occluded images. Overall, the joint representations outperform appearance and shape by a large margin, showing that RSE branch effectively extracts discriminative body shape cue that can be coupled with appearance and other clothes-irrelevant features to boost the model's performance.

Our CVSL framework surpasses CAL [7], which uses RGB modalities only, by a margin of 4.4% in R-1 and 3.2% in mAP on LTCC. The reason is that LTCC contains a large number of low-quality and occluded images (Figure 3), which limits the model's ability to mine texture information. Our method effectively utilizes body shape cue as complementary to appearance for a more discriminative global representation.

Our CVSL framework also achieves superior performance over 3DSL [4], which incorporates information from 3D shape and silhouette in addition to appearance. These modalities require complicated estimation and additional 3D ground truths, leading to heavy multi-stage training. CVSL is an end-to-end framework with lightweight RSE branch. Most importantly, results reveal that 2D shape features extracted by our method can effectively enhance robustness of model against clothing changes.

5.2.2 Contrastive Viewpoint-aware Losses (CVL)

Table 2 reports the comparison of models trained with and without the proposed contrastive viewpoint-aware losses $\mathcal{L}^S_{CVL}$ and $\mathcal{L}^A_{CVL}$, while a comparison in viewpoint variations between LTCC and PRCC is shown in Figure 4. Overall, in all model settings, it can be observed that CVL significantly improves the re-identification accuracy. Furthermore, Figure 5 shows that features produced by CVSL are more separable on latent space than model trained without CVL. The reasons are two-fold: (1) $\mathcal{L}^S_{CVL}$ guides the models to minimize the confusion caused by dissimilarity of shape from the same person under different viewpoints; and (2) side and back viewpoints cause severe occlusion, which makes captured clothes-irrelevant features not discriminative enough. In this case, $\mathcal{L}^A_V$ and $\mathcal{L}^A_{sim}$ effectively complement the missing texture information by explicitly pulling features of the same identity closer while pushing those of different identities away on latent space. Our proposed framework also shows superiority over model trained with the widely used Triplet loss [10].

Compared to FSAM [11], which utilizes human geometric cues but overlook the effect of viewpoints, our framework outperforms FSAM by 6.0% in R-1 and 5.1% in mAP on LTCC, while only a slight performance gap can be noted on PRCC. This shows effectiveness of CVSL in real-world scenarios since unlike PRCC, LTCC poses large variations in occluded viewpoints (Figure 4).

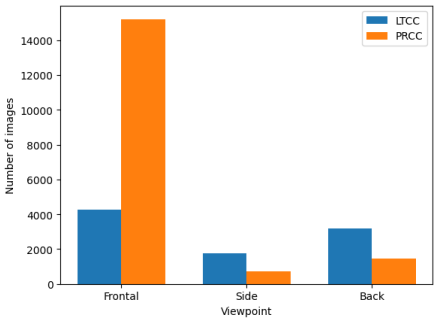

Figure 4. Comparison in viewpoint variations between LTCC and PRCC. Most images in PRCC are of frontal viewpoint, which is less challenging than LTCC.

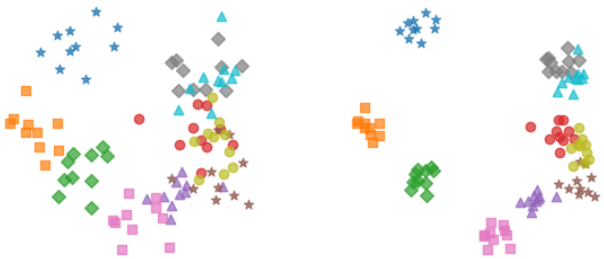

Figure 5. Distribution on feature space of global representations produced by model trained without CVL (left) and our CVSL (right); Visualized by t-NSE from 10 identities and 10 images per identity under different viewpoints randomly selected from LTCC. It can be seen that CVL is able to enhance simultaneously the intra-class similarity and inter-class diveristy, leading to more separable feature embeddings.

5.3. Further Analysis

Graph Attention Network (GAT). In Table 3, we validate the effectiveness of the proposed GAT in capturing local and global shape features. Compared to the model using the widely used Graph Convolutional Network (GCN), the proposed CVSL framework achieves higher performance on both datasets. Different from GCN, by incorporating an attention mechanism, GAT is capable of: (1) capturing finer local shape features and (2) amplifying the contribution of each keypoint feature for a discriminative aggregated global shape embedding.

Methods	LTCC		PRCC	
	R-1	mAP	R-1	mAP
Model using GCN	43.1	20.4	56.3	55.2
CVSL (Model using GAT)	**44.5**	**21.3**	**57.5**	**56.9**

Table 3. Comparison between the model using GCN and the proposed CVSL using GAT on LTCC and PRCC in cloth-changing setting only.

Adaptive Fusion Module (AFM). In Table 4, we validate the effectiveness of AFM in aggregating shape and appearance embeddings for final representations of identities. Concatenating the two embeddings degrades the model's performance on both datasets, showing the superiority of AFM which first projects the embeddings onta a latent space then amplifies the importance of each embedding to the final representation. This mitigates the influence of shape or appearance under severely occluded viewpoints.

Methods	LTCC		PRCC	
	R-1	mAP	R-1	mAP
Concatenation	43.6	20.8	56.7	55.4
AFM	**44.5**	**21.3**	**57.5**	**56.9**

Table 4. Comparison between the two methods for aggregating shape and appearance embeddings: concatenation and the proposed AFM, on LTCC and PRCC in cloth-changing setting only.

6. Conclusion

In this paper, we have presented CVSL, a framework for Long-term Person Re-ID which is robust to clothing changes and viewpoint variations. We address the challenge of clothing-confusion in Re-ID by exploiting body shape which is a stable cue in long-term for re-identification. Shape features are extracted both locally and globally using the proposed Relational Shape Embedding branch, and then effectively coupled with appearance and clothes-irrelevant features using the Adaptive Fusion module. We further improve the performance of the model by proposing contrastive viewpoint-aware losses. We mine viewpoint information to guide the model in learning a latent space where features from the same identity under different viewpoints are pulled closer, while those from different identities that share similar geometric shape or clothing are pushed away. Extensive experiments have shown the superiority of our proposed framework over state-of-the-art methods.

References

[1] Mingjing Ai, Guozhi Shan, Bo Liu, and Tianyang Liu. Rethinking reid: Multi-feature fusion person re-identification based on orientation constraints. In *ICPR*, pages 1904–1911, 2021. 2

[2] Vaibhav Bansal, Gian Luca Foresti, and Niki Martinel. Cloth-changing person re-identification with self-attention. In *WACVW*, pages 602–610, 2022. 2

[3] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019. 3, 6

[4] 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. 6, 7

[5] Yeong-Jun Cho and Kuk-Jin Yoon. Pamm: Pose-aware multi-shot matching for improving person re-identification. *IEEE TIP*, 27(8):3739–3752, 2018. 2

[6] 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. 6

[7] 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, 4, 6, 7

[8] Ke Han, Shaogang Gong, Yan Huang, Liang Wang, and Tieniu Tan. Clothing-change feature augmentation for person re-identification. In *CVPR*, pages 22066–22075, 2023. 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 6

[10] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network, 2018. 7

[11] 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, 2, 6, 7

[12] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *ICCV*, pages 11875–11884, 2021. 6

[13] Xuemei Jia, Xian Zhong, Mang Ye, Wenxuan Liu, and Wenxin Huang. Complementary data augmentation for cloth-changing person re-identification. *IEEE TIP*, 31:4227–4239, 2022. 2

[14] Na Jiang, Junqi Liu, Chenxin Sun, Yuehua Wang, Zhong Zhou, and Wei Wu. Orientation-guided similarity learning for person re-identification. In *ICPR*, pages 2056–2061, 2018. 2

[15] Xin Jin, Tianyu He, Kecheng Zheng, Zhiheng Yin, Xu Shen, Zhen Huang, Ruoyu Feng, Jianqiang Huang, Zhibo Chen, and Xian-Sheng Hua. Cloth-changing person re-identification from a single image with gait prediction and regularization. In *CVPR*, pages 14258–14267, 2022. 1, 2, 6

[16] Khadija Khaldi and Shishir Shah. Cupr: Contrastive unsupervised learning for person re-identification. In *VISIGRAPP*, 2021. 2

[17] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 6

[18] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 1

[19] 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, 6

[20] 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. 1, 2

[21] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, pages 1487–1495, 2019. 1

[22] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE TIP*, 23(8):3656–3670, 2014. 1, 2

[23] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, pages 1363–1372, 2016. 1, 2

[24] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, pages 542–551, 2019. 2

[25] 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

[26] 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, 5, 6

[27] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, pages 420–429, 2017. 2

[28] 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*, page 501–518, 2018. 1, 2, 6

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, volume 30, 2017. 4

[30] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017. 3

[31] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPRW*, pages 3620–3628, 2020. 2

[32] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *ACM MM*, 2018. 1, 2

[33] Chenyan Wu, Yukun Chen, Jiajia Luo, Che-Chun Su, Anuja Dawane, Bikramjot Hanzr, Zhuo Deng, Bilan Liu, James Z. Wang, and Cheng-Hao Kuo. Mebow: Monocular estimation of body orientation in the wild. In *CVPR*, 2020. 6

[34] Peng Xu and Xiatian Zhu. Deepchange: A long-term person re-identification benchmark with clothes change. In *ICCV*, pages 11196–11205, 2023. 2

[35] 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, 5, 6

[36] Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *CVPR*, pages 1472–1481, 2023. 2

[37] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification, 2020. 6

[38] 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. 1

[39] Ruochen Zheng, Changxin Gao, and Nong Sang. Viewpoint transform matching model for person re-identification. *Neurocomputing*, 433:19–27, 2021. 2

[40] Zhu Zhihui, Xinyang Jiang, Feng Zheng, Xiaowei Guo, Feiyue Huang, Xing Sun, and Weishi Zheng. Viewpoint-aware loss with angular regularization for person re-identification. *AAAI*, 34:13114–13121, 2020. 2

[41] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. Occluded person re-identification, 2018. 2