Contrastive Clothing and Pose Generation for Cloth-Changing Person Re-Identification

Vuong D. Nguyen Pranav Mantini Shishir K. Shah
Quantitative Imaging Lab, Dept. of Computer Science, University of Houston
dnguy222@cougarnet.uh.edu pmantini@cs.uh.edu sshah@central.uh.edu

Abstract

*Cloth-Changing Person Re-Identification (CCRe-ID) aims at matching an individual across cameras after a long period of time, presenting variations in clothing compounded with changes in pose, viewpoint, etc. In this work, we propose **CCPG**: **C**ontrastive **C**lothing and **P**ose **G**eneration framework for CCRe-ID. Beyond appearance, CCPG captures cloth-invariant body shape information using a Relational Graph Attention Network. Training a robust CCRe-ID model requires a wide range of clothing variations and expensive cloth labeling, which is lacked in current CCRe-ID datasets. To address this, we propose a GAN-based model for clothing and pose transfer across identities to augment images of more clothing variations and of different persons wearing similar clothing. The augmented batch of images serve as inputs to our proposed Fine-grained Contrastive Losses, which not only supervise the Re-ID model to learn discriminative person embeddings under long-term scenarios but also ensure in-distribution data generation. Results on CCRe-ID datasets demonstrate the effectiveness of our CCPG framework. Code will be available here.*

1. Introduction

Person Re-Identification (Re-ID) involves matching the same person in a non-overlapping camera system. Since the emergence of deep learning, person Re-ID has been well advanced with plenty of efforts [36, 40, 54]. These works assume a simplistic Re-ID scenario where the target person reappears after a short span of time with the same clothing, pose, and viewpoint. Thus, they suffer severe performance degradation under long-term scenarios where clothing, pose, and viewpoint have changed, leading to unreliable appearance as illustrated in Fig. 1(a). This shortcoming opens a more practical Re-ID problem namely Cloth-Changing Person Re-ID (CCRe-ID).

Several methods have been proposed to tackle CCRe-ID, in which two main categories can be observed: single-modality and multi-modality. Single-modality methods [8, 12, 13, 18, 45] rely primarily on appearance, face, hairstyle, and cloth labels/templates from the RGB images to learn person representations. First, this approach fails under occlusion which makes appearance unobservable. Second, these texture-based models require large-scale cloth-changing data with explicit clothing labels/templates as auxiliary labels. However, current CCRe-ID datasets present a small range of clothing variations due to the difficulty in collecting and labeling such data. Moreover, clothing labels are ambiguous across identities, thus directly leveraging manually annotated labels for identity-relevant feature learning is not beneficial. Multi-modality methods aim to capture cloth-invariant modalities such as silhouettes [15, 21], contour sketches [7, 44], or skeleton-based pose [29, 33], which are coupled with appearance for distinguishing individuals. However, viewpoint variations make it challenging to sufficiently capture fine-grained feature from these 2D modalities for Re-ID as shown in Fig. 1(b).

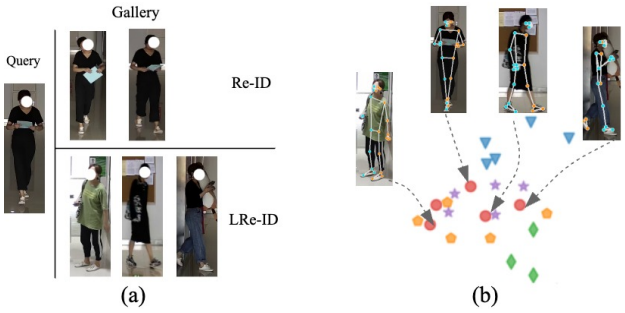

Figure 1. (a) Cloth-Changing Re-ID aims to re-identifies the same person under long-term scenarios with different clothing, viewpoint, illumination; (b) t-SNE visualization of distribution in the latent space of person embeddings output by the model trained without the proposed losses. Viewpoint variations cause ambiguity in learning pose-based shape, leading to large intra-class and small inter-class gap, shown by widespread red points. This forms the motivation of this paper.

To this end, we propose **C**ontrastive **C**lothing and **P**ose **A**ugmentation (CCPG) framework for CCRe-ID. Beyond appearance, CCPG extracts body shape information from skeleton-based pose using a Relational Graph Attention Network [2]. We then propose a GAN to perform clothing and pose transfer across identities leveraging appearance and body shape structure. CCPG addresses the lack of clothing variations in current CCRe-ID datasets by synthesizing images of different clothes for each person. Furthermore, CCPG augments training data by generating images of different identities wearing similar clothing, which is not presented in current CCRe-ID datasets. 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.

Current works that attempt to generate Re-ID data [10, 17, 23] construct the generation process separately from Re-ID learning, thus limiting the gain from generated data. Moreover, samples are generated randomly without identity-related optimization target, leading to out-of-distribution data which may not be useful for Re-ID. To exploit the augmentation efficiently, we propose novel losses which are designed specifically to serve the discriminative Re-ID learning to make better use of the generated data, while ensuring a robust generation module. The augmented batch of images is contrastively sampled as inputs for two proposed losses. First is the Fine-grained Contrastive Clothing-aware Loss (FCCL), which mitigates large intra-class gap and small inter-class across clothing variations. Second is the Fine-grained Contrastive Viewpoint-aware Loss (FCVL), which aims to minimize the ambiguity in body shapes of *different IDs under same viewpoint* and *same ID under different viewpoints* as shown in Fig. 1(b). FCCL helps to learn a robust appearance encoder under clothing confusion, while FCVL enhances discriminative power of shape representations under viewpoint changes.

Our main contributions in the paper can be summarized as follows:

1. We propose CCPG, a novel framework for CCRe-ID which can extract both identity-relevant and cloth-irrelevant features.
2. We address the lack of cloth-changing data by designing a GAN-based cross-identity clothing and pose generation process, which acts as an augmentation step to enhance CCRe-ID model's robustness under clothing changes and pose variations.
3. Leveraging the augmented training batch, we propose the fine-grained contrastive losses to further tackle large intra-class gap and small inter-class gap caused by clothing changes and viewpoint variations in CCRe-ID.
4. Extensive experiments on CCRe-ID datasets show that we achieve state-of-the-art Re-ID performance.

2. Related Work

Traditional Person Re-ID which assumes short-term scenarios has been well advanced using deep learning models for representation learning under both supervised [36, 40, 54] and unsupervised [5, 22, 24] setting. However, these appearance-based methods suffer performance degradation in long-term environment due to significant clothing changes.

2.1. Cloth-Changing Person Re-ID

With the recent release of CCRe-ID datasets [33, 41, 44], several CCRe-ID methods have been proposed which can be categorized into two approaches: single-modality and multi-modality methods. The single-modality methods only leverage the original RGB image to extract biometric features [4, 8, 13, 27]. Clothing labels and templates are mined as auxiliary labels for discriminative learning in [12, 45]. Relying solely on RGB modality suffers performance degradation when appearance is unobservable due to poor illumination conditions and occlusions. The multi-modality methods leverage clothes-irrelevant modalities that are more stable in long-term such as contour sketches [7, 44], silhouettes [15, 21, 43], 2D skeleton-based pose [31, 33, 35, 47], or 3D body structure [6, 30]. However, the ambiguity in 2D human geometric cues caused by viewpoint variations has not been well tackled.

2.2. Viewpoint-aware Person Re-ID

View-transformed feature extractors are designed in [34, 50] to attend to frontal viewpoint which is the most informative, which limits the applicability under clothing changes. Zhu *et al.* [55] proposed to use a viewpoint-aware hypersphere to cluster identities, while a viewpoint-aware feature fusion model is designed in [1]. However, these methods are not robust when different IDs wearing similar clothing under the same viewpoint. 3D shape is leveraged in [6, 30], however, capturing 3D human models requires heavy training with expensive 3D data.

2.3. Data Augmentation-based Person Re-ID

Data augmentation has been shown to help improve the feature learning ability of Re-ID models [20]. Besides traditional methods such as horizontal clipping or random erasing, GANs are widely used for Re-ID data generation [10, 17, 23, 42]. Several works [10, 22, 23, 26, 32] leverage body pose to condition the generation of pedestrian images in various poses, aiming to enhance robustness of models under pose variations. Pseudo-labels for generated images are estimated and iteratively refined based on knowledge from original data [17, 51]. Synthesizing data under different camera styles has also been exploited [52, 53]. However, these works have not considered cloth-changing scenarios which is crucial for real-world Re-ID.

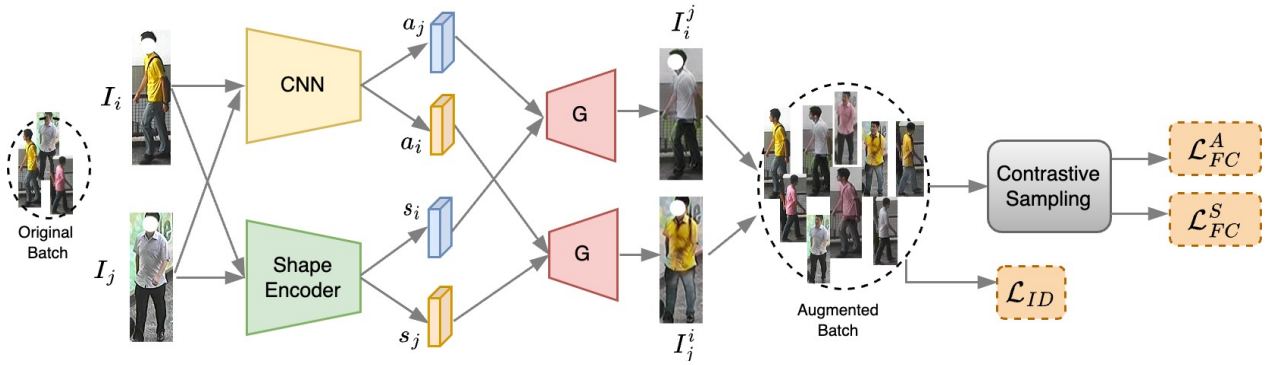

Figure 2. Overview of the proposed CCPG framework. From the original batch, for every pair of images, clothing and pose transfer is performed using an appearance encoder E^A (a CNN), a shape encoder E^S and a generator G to form the augmented batch. Contrastive sampling is performed on the augmented batch for inputs to the proposed fine-grained contrastive losses, which drive model training along with an identification loss. (Best viewed in color)

3. Methodology

3.1. The Proposed Framework

An overview of our proposed framework is given in Fig. 2. Given an original batch of N images $X = \{I_i\}_{i=1}^N$, denote the identity label set as $\{y_i\}_{i=1}^N$ (note that $y_i \neq y_j$ for $i, j = 1, ..., N$). The framework comprises an appearance encoder E^A which is a CNN and a shape encoder E^S. For every pair of images, we perform clothing and pose transfer using a generator G to obtain an augmented batch. The augmented batch X' containing N^2 images is then contrastively sampled as inputs for the proposed Fine-grained Contrastive Losses $\mathcal{L}_{FC}^A$ and $\mathcal{L}_{FC}^S$. An identification loss $\mathcal{L}_{ID}$ based on cross entropy loss is also employed as classification loss of the model. Input of $\mathcal{L}_{ID}$ is the person representation f_i obtained by concatenating appearance and shape embeddings, i.e. $f_i = [a_i, s_i]$, along with identity label y_i. The framework is trained by the total loss $\mathcal{L}$ formulated as:

$$\mathcal{L} = \mathcal{L}_{FC}^A + \mathcal{L}_{FC}^S + \mathcal{L}_{ID} \tag{1}$$

3.2. Cross-Identity Clothing and Pose Generation

To mimic the scenario of different identities wearing similar clothing and to enlarge clothing variations, we propose to transfer clothing from one person to another person while preserving the pose of the target person. Specifically, a pair of images $[I_i, I_j] \in X$ is sent to the CNN E^A and shape encoder E^S. E^A outputs appearance embeddings:

$$a_i = E^A(I_i), a_j = E^A(I_j) \tag{2}$$

which represent clothing style, while E^S outputs shape embeddings:

$$s_i = E^S(I_i), s_j = E^S(I_j) \tag{3}$$

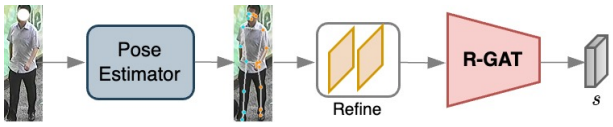

Figure 3. Architecture of Shape Encoder, which comprises of a pose estimator, a refinement network and a R-GAT.

which represent body structure. Details about shape encoding will be discussed later in Sec. 3.3. Then, a generator G takes in clothing style from one person and and body structure from the other person to generate new pedestrian images:

$$I_i^j = G(s_i, a_j), \; I_j^i = G(s_j, a_i) \tag{4}$$

where I_i^j is constructed from the body structure of person i and clothing style of person j, and vice versa. The identity label of the synthesized image corresponds to the label of the image that gives the body structure, i.e. $y_{I_i^j} = y_{I_i}$, since we apply new clothing from a_j on the old person. G consists of several convolutional layers followed by residual blocks [14], where every residual block contains two adaptive instance normalization layers [16]. A discriminator D is used to discriminate between real and fake images. Architecture of D is employed as in Patch-GAN [19]. Gradient punishment [28] is applied to stabilize D during training. Adversarial losses to train the generator G and discriminator D are similar to ones proposed in the original GAN paper [11]. G and D are pre-trained on Market-1501 dataset [49], and then are frozen during training of the Re-ID discriminative learning to ensure they do not lose their intrinsic generative ability.

3.3. Shape Representation Learning

To tackle long-term Re-ID scenarios, skeleton-based shape remains a competitive modality since it is invariant to clothing variations. The architecture of the Shape Encoder is illustrated in Fig. 3. Given an image $I \in X$, pose is estimated using OpenPose [3] which outputs a set of m joint nodes $\mathbf{J} = \{\mathbf{j}_i\}_{i=1}^{m}$. These nodes are then transformed into higher-dimensional feature vectors $\mathbf{J}' = \{\mathbf{j}'_i\}_{i=1}^{m}, \mathbf{j}'_i \in \mathbb{R}^F$ by a refinement network consisting of several fully-connected layers. Intuitively, to represent body shape, we need to capture the local and global relationships among nodes. To achieve this, we employ Relational Graph Attention Network (R-GAT) [2]. R-GAT differs from GCN by incorporating attention mechanism into graph operation to account for cross-node importance, while compared to GAT [37], R-GAT also amplifies different levels of relations among bones, which is crucial for capturing a discriminative shape when some body parts are invisible due to occlusion or extreme viewpoints.

Input of R-GAT is a graph constructed from $\mathbf{J}'$ with m nodes and k relations (k edges/bones). Let $\mathbf{J}' = [\mathbf{j}'_1, ..., \mathbf{j}'_m] \in \mathbb{R}^{m \times F}$ be the input feature matrix, the intermediate representation matrix under a relation $\mathbf{r}$ is computed as:

$$\mathbf{H}^{(\mathbf{r})} = \mathbf{J}' W^{(\mathbf{r})} \in \mathbb{R}^{m \times F'}, \tag{5}$$

where $\mathbf{H}^{(\mathbf{r})} = [\mathbf{h}_1^{(\mathbf{r})}, ..., \mathbf{h}_m^{(\mathbf{r})}]$ and $W^{(\mathbf{r})} \in \mathbb{R}^{F \times F'}$ are learnable parameters. For each node $\mathbf{h}_i^{(\mathbf{r})} \in \mathbf{H}$, denote the indices set of its neighbors as $\mathcal{N}_i^{(\mathbf{r})}$, $\mathbf{h}_i^{(\mathbf{r})}$ is updated after a R-GAT layer based on a weighted sum over the nodes in its neighborhood and over all the relations as follows:

$$\hat{\mathbf{h}}_i = \sigma\Big(\sum_{\mathbf{r}} \sum_{j \in \mathcal{N}_i^{(\mathbf{r})}} \alpha_{i,j}^{(\mathbf{r})} \hat{\mathbf{h}}_j^{(\mathbf{r})}\Big), \;\; \alpha_{i,j}^{(\mathbf{r})} = \mathrm{SM}\left(\mathbf{h}_i^{(\mathbf{r})} \mathbf{h}_j^{(\mathbf{r})}\right), \tag{6}$$

where $\alpha_{i,j}^{(\mathbf{r})}$ is attention score between the i^{th} and j^{th} nodes, σ is sigmoid activation function, and SM denotes softmax. The graph is finally aggregated to obtain shape embedding s by computing the mean of node representations as:

$$s = \frac{1}{m} \sum_{i=1}^{m} \hat{\mathbf{h}}_i. \tag{7}$$

3.4. Fine-grained Contrastive Losses

Given the augmented batch X', denote the identity label set as $\{y_i\}_{i=1}^{N^2}$, clothing label set as $\{c_i\}_{i=1}^{N^2}$, and viewpoint label set as $\{v_i\}_{i=1}^{N^2}$. To enhance the robustness of appearance encoder under cloth-confusing scenarios, we propose the **Fine-grained Contrastive Clothing-aware Loss** (FCCL), whose inputs are sampled as illustrated in Fig. 4(a): for an appearance embedding a_i of $I_i \in X'$ as anchor, the set of positive samples (denoted as P_i^{A+}) consists of images of the same identity, while the set of negative samples (denoted as P_i^{A-}) consists of images of different identities. Since in the original batch we sample two images per identity, *positive samples in P_i^{A+} have a wide range of clothing variations transferred from other identities*, while *negative samples include different identities in similar clothing as anchor*. Therefore, the proposed FCCL helps enforce a large penalization on small inter-class and large intra-class difference caused by ambiguity in clothing variations. FCCL is formulated as:

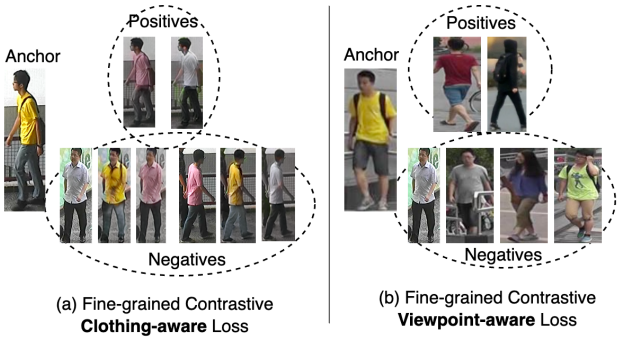

Figure 4. Illustration of Contrastive Sampling for the proposed Fine-grained Contrastive Losses: (a) FCCL, and (b) FCVL.

$$\mathcal{L}_{FC}^{A} = -\sum_{i=1}^{N^2} \gamma_1 \log \frac{\sum_{j \in P_i^{A+}} \eta(a_i, a_j)}{\sum_{k \in P_i^{A-}} \eta(a_i, a_k)}, \tag{8}$$

where $P_i^{A+} = \{j \,|\, I_j \in X' \,|\, y_j = y_i, c_j \neq c_i\}$, $P_i^{A-} = \{k \,|\, I_k \in X' \,|\, y_k \neq y_i\}$, $\eta(\cdot, \cdot)$ is a distance-to-distribution function, and γ_1 is a scalar controlling the scale of $\mathcal{L}_{FC}^{A}$.

To mitigate the ambiguity in shape caused by viewpoint variations as shown in Fig. 1(b), we propose **Fine-grained Contrastive Viewpoint-aware Loss** (FCVL) $\mathcal{L}_{FC}^{S}$, whose inputs are sampled as illustrated in Fig. 4(b): for a shape embedding s_i of $I_i \in X'$ as anchor, images of the *same identity* but *different viewpoints* are chosen as positive samples (denoted as P_i^{S+}), while images of *different identities* under the *same viewpoint* are negative samples (denoted as P_i^{S-}). Specifically, denote viewpoint labels as v, FCVL is formulated as:

$$\mathcal{L}_{FC}^{S} = -\sum_{i=1}^{N^2} \gamma_2 \log \frac{\sum_{j \in P_i^{S+}} \eta(s_i, s_j)}{\sum_{k \in P_i^{S-}} \eta(s_i, s_k)}, \tag{9}$$

where $P_i^{S+} = \{j \,|\, I_j \in X' \,|\, y_j = y_i, v_j \neq v_i\}$, $P_i^{S-} = \{k \,|\, I_k \in X' \,|\, y_k \neq y_i, v_k = v_i\}$, and γ_2 is a scalar controlling the scale of $\mathcal{L}_{FC}^{S}$. Viewpoint labels can be estimated using off-the-shelf body orientation estimators.

The distance-to-distribution function η is a key component in Eq. 8 and 9. In this work, we formulate η as an

Methods	Modality	LTCC				PRCC				DeepChange	
		CC		Standard		CC		Standard		R1	mAP
		R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP		
PCB [36]†	RGB	23.5	10.0	65.1	30.6	41.8	38.7	99.8	97.0	7.6	6.6
RCSANet [18]	RGB	-	-	-	-	50.2	48.6	**100**	97.2	-	-
CAL [12]†	RGB	40.1	18.1	74.2	40.8	55.2	55.8	**100**	**99.8**	8.1	7.4
ACID [46]	RGB	29.1	14.5	65.1	30.6	55.4	56.1	99.1	99.0	-	-
IGEP [48]	RGB	43.4	18.2	-	-	57.3	55.8	99.7	99.8	-	-
AFL [27]	RGB	42.1	18.4	74.4	39.1	57.4	56.5	100	99.7	-	-
LDF [4]	RGB	32.9	15.4	73.4	36.9	54.6	-	99.6	-	-	-
CCFA [13]	RGB	45.3	22.1	75.8	42.5	61.2	**58.4**	99.6	98.7	-	-
SAGE [44]†	RGB + sketch	-	-	-	-	34.4	-	64.2	-	8.0	7.2
CESD [33]	RGB + pose	26.2	12.4	71.4	34.3	-	-	-	-	-	-
GI-ReID [21]†	RGB + sil	23.7	10.4	63.2	29.4	33.3	-	80.0	-	7.9	7.0
3DSL [6]	RGB + pose + sil	31.2	14.8	-	-	51.3	-	-	-	-	-
DCR-ReID [8]	RGB + sil	41.1	20.4	76.1	42.3	57.2	57.4	100	99.7	-	-
FSAM [15]	RGB + pose + sil	38.5	16.2	73.2	35.4	54.5	-	98.8	-	-	-
CAMC [38]	RGB + pose	36.0	15.4	73.2	35.3	-	-	-	-	-	-
IRANet [35]	RGB + pose	-	-	-	-	54.9	53.0	99.7	99.8	-	-
UCAD [43]	RGB + sil	23.7	10.4	63.2	29.4	45.3	-	96.5	-	-	-
MBUNet [47]	RGB + pose	40.3	15.0	67.6	34.8	58.7	55.2	99.8	99.6	-	-
AIM [45]†	RGB + gray	40.6	19.1	76.3	41.1	57.9	58.3	**100**	**99.8**	8.3	7.9
CVSL [29]	RGB + pose	44.5	21.3	76.4	41.9	57.5	56.9	97.5	99.1	-	-
CCPG (Ours)	RGB + sil	**46.2**	**22.9**	**77.2**	**42.9**	**61.8**	58.3	**100**	99.6	**12.2**	**10.7**

Table 1. Quantitative comparison with previous CCRe-ID methods on LTCC, PRCC, and DeepChange datasets. Best results are shown in **bold**, while second-best results are underlined. "†" denotes we reproduced results on DeepChange using the provided open-source code. "-" denotes results are not available.

exponential function as follows: $\eta(\cdot,\cdot) = e^{-d(\cdot,\cdot)}$ where $d(\cdot,\cdot)$ denotes euclidean distance. The exponentially sensitive penalization enforces the model to pull embeddings of the same identities closer while pushing that of different identities farther, resulting in a well-separable latent space. Simultaneously, it ensures that the distribution of the generated samples is close to original images.

4. Experiments

4.1. Experimental Setup

Datasets. To validate the performance of our framework, we use three CCRe-ID datasets: LTCC [33], PRCC [44], and DeepChange [41]. LTCC consists of 17k images from 152 identities captured by 12 camera during two months. PRCC contains 33k images of 221 identities captured by 3 cameras. Both LTCC and PRCC are collected under indoor environments. DeepChange is a much larger-scale dataset with 178k images from 1,121 identities captured by 17 cameras during a longer span of time (twelve months). Persons in PRCC are mostly captured in frontal viewpoint with good lighting condition, and each identity only has from 2 to 5 clothing variations. On the other hand, LTCC data presents large variations in illumination and occlusion, which mimics a more challenging Re-ID environment. For DeepChange, it was collected under various outdoor scenes and weather conditions with large variations in clothing, viewpoint, pose, etc. Moreover, there is a significant distribution gap in data between DeepChange and LTCC/PRCC. Thus, it can be used to validate generalization ability of the Re-ID model.

Evaluation Protocols. CMC at rank k (rank-k accuracy) and mean Average Precision (mAP) are used as evaluation metrics. Two evaluation settings are set up. First is *Cloth-Changing (CC)* where only cloth-changing samples are used for testing. Second is *Standard* where for LTCC, both cloth-consistent and cloth-changing samples form the query and gallery sets, while for PRCC, due to its nature, the test sets only contain cloth-consistent samples. For DeepChange, no cloth labels are provided to set up cloth-changing setting.

Implementation Details. Viewpoint is estimated using MEBOW [39]. For appearance encoder E^A, we adopt ResNet-50 [14] with pretrained weights on ImageNet [9]. For shape encoding, pose is estimated in COCO format using OpenPose [3]. The refinement network consists of 2 fully-connected layers of size [512, 2048], while R-GAT consists of 2 layers. Generator G has four residual blocks and four convolutional layers. The architecture of discriminator D is identical to one proposed in [19], with the last

Method	CPG	$\mathcal{L}_{FC}^{A}$	$\mathcal{L}_{FC}^{S}$	LTCC		PRCC	
				R-1	mAP	R-1	mAP
1	✓	-	-	42.4	19.2	56.1	54.3
2	✓	✓	-	44.6	20.9	59.1	56.8
3	✓	-	✓	44.0	20.6	58.8	56.2
4	-	✓	✓	44.3	20.8	59.2	57.0
CCPG	✓	✓	✓	**46.1**	**22.9**	**61.7**	**58.4**

Table 2. Ablation study of Cross-Identity Clothing and Pose Generation (CPG) and the proposed Fine-grained Contrastive Losses on LTCC and PRCC in cloth-changing setting.

layer of one neuron to output fake/real probability of the generate image. We set $\gamma_1 = 0.7$ and $\gamma_2 = 0.3$. Input images are resized to 256×128 for training. Each original batch is sampled with 8 images from 8 identities, giving an augmented batch size of 64. The framework is trained for 90 epochs using Adam optimizer with an initial learning rate $lr = 0.0005$, momentum of 0.9, and weight decay of 0.01. lr is reduced by a factor of 0.1 after every 30 epochs. Implementation is in PyTorch.

4.2. Results

In Tab. 1, we provide a quantitative comparison in results between CCPG and previous CCRe-ID methods on LTCC, PRCC, and DeepChange.

It can be seen that clothing changes cause relatively inferior performance to the standard Re-ID method since appearance is no longer reliable. Among CCRe-ID methods, a majority of them [6, 15, 21, 25, 29, 33, 38, 44, 45] resort to cloth-invariant modalities such as sketch, silhouette, or pose, with results showing that this approach is effective in mitigating the influence of clothing changes. Although recent texture-based methods [12, 13] achieve comparable performance, they rely heavily on the availability of large clothing variations with expensive manual cloth labeling to be aware of clothing changes. The influence of viewpoint changes has also been ignored. Moreover, the discriminative fine-grained information identity information under similar-clothing scenarios has not been adequately mined. Our CCPG framework achieves 46.1/22.9% Rank-1/mAP accuracy on LTCC and 61.7/58.4% Rank-1/mAP accuracy on PRCC. By addressing these issues, overall, we outperform previous methods in cloth-changing environment on both LTCC and PRCC. **Cross-dataset Generalizability.** On the other hand, cross-dataset generalizability of the model is demonstrated via results on DeepChange, since there is a large distribution gap between this dataset and LTCC or PRCC. It can be seen that CCPG remarkably outperforms the second-best method AIM [45] by 3.9/2.8% in Rank-1/mAP. This shows the effectivenes of our method in handling highly challenging Re-ID scenarios of pose variations, viewpoint changes, and especially clothing changes in a large-scale real-world Re-ID dataset.

Figure 5. Visualization of samples generated by Cross-Identities Clothing and Pose Generation. Clothes from the target is transferred to the upper image for the output image of the same identity in different clothing.

4.3. Ablation Study

To further validate the effectiveness of our proposed method, we perform comprehensive ablation studies on LTCC and PRCC of: Cross-Identity Clothing and Pose Generation (CPG), the proposed Fine-grained Contrastive Losses, and the Relational Graph Attention Network.

Cross-Identity Clothing and Pose Generation. From Tab. 2, it can be seen that CPG significantly improves Re-ID performance on both datasets (method 4 vs CCPG). For example, Rank-1 accuracy is boosted by 1.8/2.5% on LTCC/PRCC. This demonstrates the rationality and effectiveness of our proposed approach to address the lack of clothing variations and labels in current CCRe-ID datasets. CPG leads to better generalizability by exposing the model to similar-clothing situations, shown by an improvement of 1.1/1.4% in mAP on LTCC/PRCC. A visualization of augmented images generated by our framework is shown in Fig. 5.

Fine-grained Contrastive Losses. Tab. 2 demonstrates that overall, with the proposed Fine-grained Contrastive Losses, Re-ID performance is remarkably improved on both datasets (method 1 vs CCPG) with an increase of 3.7/3.7% in Rank-1/mAP on LTCC and 5.6/4.1% in Rank-1/mAP on PRCC. We also run experiments with adding one loss and excluding the other to explore the contribution of each loss (method 2 and method 3). By leveraging contrastive learning, the clothing-aware loss $\mathcal{L}_{FC}^{A}$ effectively guides the model to distinguish identities under clothing variations and similar clothing, shown by a boost of 2.2/1.7% in Rank-1/mAP on LTCC. The viewpoint-aware loss $\mathcal{L}_{FC}^{S}$ enhances the model's performance by learning shape embeddings with high discriminative power under viewpoint changes.

Relational Graph Attention Network. In Tab. 3, we show the effectiveness of using R-GAT in encoding shape representation from the skeleton-based graph. R-GAT clearly

Method	LTCC		PRCC	
	R-1	mAP	R-1	mAP
GCN	44.9	22.0	59.8	56.1
GAT	45.3	22.3	60.4	57.1
R-GAT	**46.1**	**22.9**	**61.7**	**58.4**

Table 3. Ablation study of R-GAT on LTCC and PRCC in cloth-changing setting.

Component	Training	Inference	Num. params
CNN	✓	✓	28M
Shape Encoder	✓	-	4M
Gen & Disc	✓	-	6M

Table 4. Summary of model architecture size and training/inference complexity.

shows superiority over GAT and GCN. For example, R-GAT outperforms GAT by 0.8/0.6% and GCN by 1.2/0.9% in Rank-1/mAP on LTCC. This indicates that besides capturing relationship among joints, it is beneficial to capture the relationship among bones for a robust shape embedding.

4.4. Further Analysis

Model Complexity Analysis. A summary of model architecture size and training/inference complexity is presented in Tab. 4. Our entire framework is not of large size with only around 38M parameters, where the components like shape encoder or generator and discriminator do not significantly increase model size. In terms of computational efficiency, note that during testing, *only the original DCN backbone is used for inference* to obtain the person representation from an input RGB image. Shape encoder or generator and discriminator are used *only during training* for clothing and pose augmentation to assist the learning ability of CNN backbone. This helps save computational cost, ensuring efficiency in deployment of real-time applications.

Limitation. A limitation can be noticed from the last sample provided in Fig. 5. Our generator tends to learn the overall clothing texture information, while some rare patterns (e.g., logos on shirts) are ignored and can not be transferred to the new image. The reason may be two folds. First, it is the bias in training data where very limited patterns like logos on shirts are presented. Second, we are using a CNN to extract the overall clothing texture information for transferring and have not paid attention to details in outfit. To improve this, in future work, we intend to segment the cloth first to be able to attend more to clothing regions and extract more fine-grained clothing details.

5. Conclusion

In this work, we propose CCPG, a novel framework for CCRe-ID. CCPG extracts simultaneously appearance and cloth-invariant pose-based shape, then leverages these modalities to transfer clothing and pose across identities for cloth-changing data generation. The augmented data serves as inputs to the Fine-grained Contrastive Losses, which guides the model to learn a well separable latent space of person embeddings under cloth-confusing and viewpoint-changing scenarios. Extensive experiments and ablation studies on CCRe-ID datasets validate the effectiveness of our proposed framework.

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] Dan Busbridge, Dane Sherburn, Pietro Cavallo, and Nils Y Hammerla. Relational graph attention networks. *arXiv preprint arXiv:1904.05811*, 2019. 2, 4

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017. 4, 5

[4] Patrick P. K. Chan, Xiaoman Hu, Haorui Song, Peng Peng, and Keke Chen. Learning disentangled features for person re-identification under clothes changing. *ACM MM*, 19(6), 2023. 2, 5

[5] Hao Chen, Yaohui Wang, Benoit Lagadec, Antitza Dantcheva, and Francois Bremond. Joint generative and contrastive learning for unsupervised person re-identification. In *CVPR*, pages 2004–2013, 2021. 2

[6] Jiaxing Chen, Xinyang Jiang, Fudong Wang, Jun Zhang, Feng Zheng, Xing Sun, and Wei-Shi Zheng. Learning 3d shape feature for texture-insensitive person re-identification. In *CVPR*, pages 8142–8151, 2021. 2, 5, 6

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

[8] Zhenyu Cui, Jiahuan Zhou, Yuxin Peng, Shiliang Zhang, and Yaowei Wang. Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. *IEEE TCSVT*, 2023. 1, 2, 5

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

[10] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *NeurIPS*, 2018. 2

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeuRIPS*, 27, 2014. 3

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

[13] 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. 1, 2, 5, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqin Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5

[15] 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, 5, 6

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *CVPR*, pages 1501–1510, 2017. 3

[17] Yan Huang, Jingsong Xu, Qiang Wu, Zhedong Zheng, Zhaoxiang Zhang, and Jian Zhang. Multi-pseudo regularized label for generated data in person re-identification. *IEEE TIP*, 28(3):1391–1403, 2018. 2

[18] 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. 1, 5

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 1125–1134, 2017. 3, 5

[20] Yiqi Jiang, Weihua Chen, Xiuyu Sun, Xiaoyu Shi, Fan Wang, and Hao Li. Exploring the quality of gan generated images for person re-identification. In *ACM MM*, pages 4146–4155, 2021. 2

[21] 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, 5, 6

[22] Khadija Khaldi, Vuong D. Nguyen, Pranav Mantini, and Shishir Shah. Unsupervised person re-identification in aerial imagery. In *WACV Workshops*, pages 260–269, 2024. 2

[23] Amena Khatun, Simon Denman, Sridha Sridharan, and Clinton Fookes. Pose-driven attention-guided image generation for person re-identification. *Pattern Recognition*, 137: 109246, 2023. 2

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

[25] Yu-Jhe Li, Xinshuo Weng, and Kris M. Kitani. Learning shape representations for person re-identification under clothing change. In *WACV*, pages 2431–2440, 2021. 6

[26] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *CVPR*, pages 4099–4108, 2018. 2

[27] Yuxuan Liu, Hongwei Ge, Zhen Wang, Yaqing Hou, and Mingde Zhao. Clothes-changing person re-identification via universal framework with association and forgetting learning. *IEEE TMM*, 26:4294–4307, 2024. 2, 5

[28] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 3

[29] Vuong D. Nguyen, Khadija Khaldi, Dung Nguyen, Pranav Mantini, and Shishir Shah. Contrastive viewpoint-aware shape learning for long-term person re-identification. In *WACV*, pages 1041–1049, 2024. 1, 5, 6

[30] Vuong D. Nguyen, Pranav Mantini, and Shishir Shah. Temporal 3d shape modeling for video-based cloth-changing person re-identification. In *WACV Workshops*, pages 173–182, 2024. 2

[31] Vuong D. Nguyen, Samiha Mirza, Pranav Mantini, and Shishir K. Shah. Attention-based shape and gait representations learning for video-based cloth-changing person re-identification. In *VISIGRAPP (2: VISAPP)*, pages 80–89, 2024. 2

[32] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *ECCV*, pages 650–667, 2018. 2

[33] 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

[34] 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

[35] Wei Shi, Hong Liu, and Mengyuan Liu. Iranet: Identity-relevance aware representation for cloth-changing person re-identification. *Image and Vision Computing*, 117:104335, 2022. 2, 5

[36] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, page 501–518, 2018. 1, 2, 5

[37] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 4

[38] Qizao Wang, Xuelin Qian, Yanwei Fu, and Xiangyang Xue. Co-attention aligned mutual cross-attention for cloth-changing person re-identification. In *ACCV*, pages 2270–2288, 2022. 5, 6

[39] 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. 5

[40] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *CVPR*, pages 2119–2128, 2018. 1, 2

[41] Peng Xu and Xiatian Zhu. Deepchange: A long-term person re-identification benchmark with clothes change. In *ICCV*, pages 11196–11205, 2023. 2, 5

[42] Wanlu Xu, Hong Liu, Wei Shi, Ziling Miao, Zhisheng Lu, and Feihu Chen. Adversarial feature disentanglement for

long-term person re-identification. In *IJCAI*, pages 1201–1207, 2021. 2

[43] Yuming Yan, Huimin Yu, Shuzhao Li, Zhaohui Lu, Jianfeng He, Haozhuo Zhang, and Runfa Wang. Weakening the influence of clothing: Universal clothing attribute disentanglement for person re-identification. In *IJCAI*, pages 1523–1529, 2022. 2, 5

[44] 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

[45] 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. 1, 2, 5, 6

[46] Zhengwei Yang, Xian Zhong, Zhun Zhong, Hong Liu, Zheng Wang, and Shin'Ichi Satoh. Win-win by competition: Auxiliary-free cloth-changing person re-identification. *IEEE TIP*, 32:2985–2999, 2023. 5

[47] Guoqing Zhang, Jie Liu, Yuhao Chen, Yuhui Zheng, and Hongwei Zhang. Multi-biometric unified network for cloth-changing person re-identification. *IEEE TIP*, 32:4555–4566, 2023. 2, 5

[48] Zhiwei Zhao, Bin Liu, Yan Lu, Qi Chu, Nenghai Yu, and Chang Wen Chen. Joint identity-aware mixstyle and graph-enhanced prototype for clothes-changing person re-identification. *IEEE TMM*, 26:3457–3468, 2024. 5

[49] 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. 3

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

[51] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3754–3762, 2017. 2

[52] Zhun Zhong, Liang Zheng, Shaozi Li, and Yi Yang. Generalizing a person retrieval model hetero-and homogeneously. In *ECCV*, pages 172–188, 2018. 2

[53] Zhun Zhong, Liang Zheng, Zhiming Luo, Shaozi Li, and Yi Yang. Invariance matters: Exemplar memory for domain adaptive person re-identification. In *CVPR*, 2019. 2

[54] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *ICCV*, pages 3701–3711, 2019. 1, 2

[55] Zhihui Zhu, Xinyang Jiang, Feng Zheng, Xiaowei Guo, Feiyue Huang, Xing Sun, and Weishi Zheng. Aware loss with angular regularization for person re-identification. In *AAAI*, pages 13114–13121, 2020. 2