Multi-Level Pose-Guidance with Cross-Modality Fusion for Long-Term Spatio-Temporal Person Re-Identification

Qingyuan Deng[1], Keyu Zhu[1], Jindan Wu[1], Xiaoning Li[1,*], Xinxin Li[2], Shihai He[3] and Lin Feng[1,*]

[1]*School of Computer Science, Sichuan Normal University, Chengdu 610066, China*

[2]*School of Computer and Software, Chengdu Jincheng College, , Chengdu, 611731, China*

[3]*Sichuan Mineral Electromechanic Technician College, Chengdu, 610503, China*

Abstract

Person re-identification (Re-ID) is an important visual task related to surveillance security, aimed at enhancing the tracking of the same individual across spatio-temporal regions. Traditional Re-ID methods predominantly depend on extracting garment-dominated texture features from global appearance representations. This inherent clothing bias leads to performance degradation in long-term spatio-temporal scenarios where appearance consistency cannot be guaranteed (e.g., clothing changes). In recent years, research on clothing changes in long-term scenarios has gained increasing attention. Although most existing methods for clothing changes Re-ID attempt to learn distinctive identity features of individuals (e.g., posture features), they are still subject to interference from clothing information. To mitigate this impact, this paper introduces a Multi-Level Pose-Guidance with Cross-Modality Fusion (MPCF) framework for clothing changes person re-identification. The framework consists of three main components: a Shape Embedding (SE) branch, a Cross-Modality Fusion (CMF) branch, and a Multi-Level Feature Guidance (MLFG) branch. The MLFG branch, in conjunction with the SE branch, helps the CMF branch learn more human pose information during the inference stage. We have demonstrated the effectiveness of this method through extensive experiments and achieved excellent performance in several clothing changes Re-ID benchmark tests.

Keywords

Person re-identification, Cross-temporal-spatial person tracking, Long-Term scenarios, Computer vision

1. Introduction

Person re-identification (Re-ID) is an important automated person retrieval technology in video surveillance systems. It aims to connect the movement trajectories of individuals across different spatio-temporal regions, facilitating person tracking across time, locations, and devices. This technology holds significant research value in the construction of public safety. Over the past decade, traditional person Re-ID has been extensively researched, but few models have been deployed in practical applications. This is because information in real-world spatio-temporal scenarios is complex and dynamic, and multiple factors constrain model performance. One of the key factors affecting re-identification performance is the change in person clothing.

In real-life scenarios, persons may change their clothes for various reasons, such as weather changes, personal preferences, or specific occasion requirements. These changes not only alter the appearance of persons but also increase the instability of their identity features, posing a significant challenge to traditional appearance-based Re-ID systems. Traditional Re-ID methods typically rely on shallow features such as color, texture, and shape; these features frequently exhibit instability and limited robustness in long-term scenarios.

As shown in Fig. 1, the same person wearing different clothes across different spatiotemporal scenarios exhibits significant appearance feature discrepancies. Conversely, different individuals dressed in similar clothing show excessively similar texture information. Therefore, solely relying on appearance information to address long-term problems is infeasible.

To address the challenge of clothing changes in long-term scenarios, recent research focuses on learning clothing-agn-
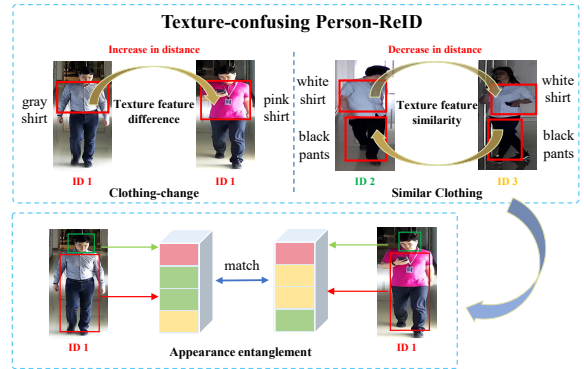
*Corresponding author.
✉ dqy@stu.sicnu.edu.cn (Q. Deng); fenglin@sicnu.edu.cn (L. Feng)

Figure 1: clothing changes degrade person recognition accuracy: different outfits in same-identity samples increase intra-class distance; similar outfits in different-identity samples reduce inter-class distance.

ostic identity features. While some methods [1, 2] directly decouple identity cues from raw images, this often results in incomplete feature learning due to the absence of multi-modal guidance. Others exploit biometric traits (e.g., body shape) as stable identity markers, yet their extraction from RGB images remains non-trivial. Consequently, auxiliary modalities are widely adopted: pose estimation [3, 4], gait recognition [2], and human keypoints/sketches [5] have been integrated to reduce clothing dependency. However, two critical issues persist: (1) clothing interference remains non-negligible even with multi-modal inputs, and (2) direct fusion of heterogeneous modalities risks information loss due to feature discrepancies. To mitigate these limitations, we propose MPCF, a multi-level pose-guided framework with cross-modal fusion for robust LT-ReID.

Specifically, the MPCF framework consists of three main branches: Shape Embedding (SE), Cross-Modality Fusion (CMF), and Multi-Level Feature Guidance (MLFG). In the first two branches, SE uses a weight-frozen pose extractor to extract body shape-related features, capturing structured in-

formation related to identity. CMF then reduces information differences between modalities by cross-modal aggregation of shape features and global appearance features, preserving more clothing-irrelevant identity cues. To further minimize interference from residual clothing information in the aggregated features, MLFG aligns the divergence between multi-level person appearance embeddings and SE's shape embeddings. This process not only helps extract pose information at different granularities from person appearances but also guides CMF to focus more on pose information during cross-modal aggregation, thereby better reducing the impact of clothing information. In summary, the main contributions of this paper are as follows:

- We obtain clothing-agnostic human shape embeddings through a frozen pose estimator and a shape encoder and interact these embeddings with pedestrian appearance in a cross-modal manner to generate more robust fused features. To further reduce clothing-related interference in appearance and highlight clothing-agnostic information in features, we use pose information as supervision to extract fine-grained pose details from raw images;
- We propose a MLGF branch that leverages biological information as supervision. This branch learns multi-granularity pose information from appearance features at three different levels, guiding the model to focus more on clothing-agnostic information during cross-modal feature aggregation and reducing clothing-related interference;
- The effectiveness of our method is demonstrated through extensive experiments on several cloth-changing datasets test benchmarks;

2. Related Work

2.1. Person Re-Identification

Traditional person re-identification methods primarily target scenarios with short-term appearance consistency, distinguishing individuals via visual feature extraction. These methods typically rely on the color, texture, and shape of clothing to characterize persons [6, 7, 8]. In recent years, with the advancement of deep learning, the field of person re-identification has made significant progress. Most methods now use deep neural networks to extract both global and local features for precise individual descriptions [9, 10, 11].

For example, Zheng et al. [10] employed a multi-class classification loss to learn discriminative global features by treating each identity as a unique category. However, the abstraction of global features weakens their sensitivity to subtle differences, particularly for visually similar individuals. To mitigate this, local feature-driven approaches have emerged, enhancing detail capture through localized regions or key points. For instance, Rigoll et al. [11] designed a multi-branch architecture that combines global features with local body region features, improving recognition performance from multiple aspects. Wang et al. [12] proposed a Multi-Granular Network (MGN) to integrate fine-grained local features with global features. Additionally, some studies have focused on optimizing similarity measurement functions [13, 14, 15] to reduce the distance between samples of the same class and increase the distance between different classes, thereby improving recognition accuracy. However,

since clothing often occupies a large portion of person images, these traditional appearance-based methods overly rely on extracting clothing information, resulting in significant performance degradation in scenarios involving long-term clothing changes. This has spurred the rise of research in Long-Term person re-identification (LT-ReID).

2.2. Long-Term Person Re-Identification

Unlike traditional person Re-ID, LT-ReID concentrates on scenarios where pedestrian appearances change over long-term spatio-temporal cycles. Clothing, which is the main part of pedestrian appearance, poses a significant challenge for identity recognition due to its variability. Many studies have attempted to address the problems caused by clothing changes. They have tried to bring in biometric attributes that are not related to clothing to enhance the representation of persons and minimize the interference from clothing. These biometric attributes include body shape, gait information, and facial features. By incorporating these attributes, they aim to provide a more comprehensive and stable representation of individuals, which can help improve the accuracy of identity recognition in LT-ReID scenarios.

Yang et al. [16] demonstrated the superior reliability of body contour curves over color-based appearance features under clothing variations. Their CC-ReID framework innovatively employs contour sketches as auxiliary biometric descriptors, translating anatomical silhouettes into identity-discriminative embeddings. Chen et al. [17] addressed clothing texture interference through 3D shape reconstruction, leveraging volumetric human models to capture anthropometric invariants like torso proportions and limb geometry. Wang et al. [18] developed a cross-modal fusion architecture that synergizes holistic appearance features with kinematic pose embeddings. By aligning spatiotemporal patterns of body joints with global representations, their method amplifies clothing-agnostic cues while suppressing transient apparel artifacts. Liu et al. [19] pioneered feature disentanglement via 3D human mesh estimation, isolating persistent identity markers (e.g., skeletal structure, joint topology) from transient non-identity variables like garment shape and dynamic postures. Their dual-path learning architecture enables parallel extraction of identity-sensitive features (from nude mesh models) and apparel-dependent features (from clothed RGB inputs). Through adversarial training, the model jointly optimizes both feature streams, achieving cross-apparel invariance by explicitly decoupling biological signatures from sartorial noise. This bidirectional learning paradigm not only enhances discrimination under clothing changes but also mitigates pose-induced feature distortions.

While existing multi-modal approaches have mitigated clothing dependency in traditional person re-identification (Re-ID), complete elimination of clothing bias remains a persistent challenge. To address this limitation, we propose a multi-level pose-guided feature learning framework that synergistically integrates pose estimation with Re-ID feature extraction. Beyond simply employing pose features as auxiliary inputs, our hierarchical design establishes explicit guidance mechanisms through progressively refined pose representations. This architecture compels the model to preserve discriminative non-appearance attributes including body geometry and motion patterns, thereby achieving enhanced robustness in long-term scenarios with clothing variations.
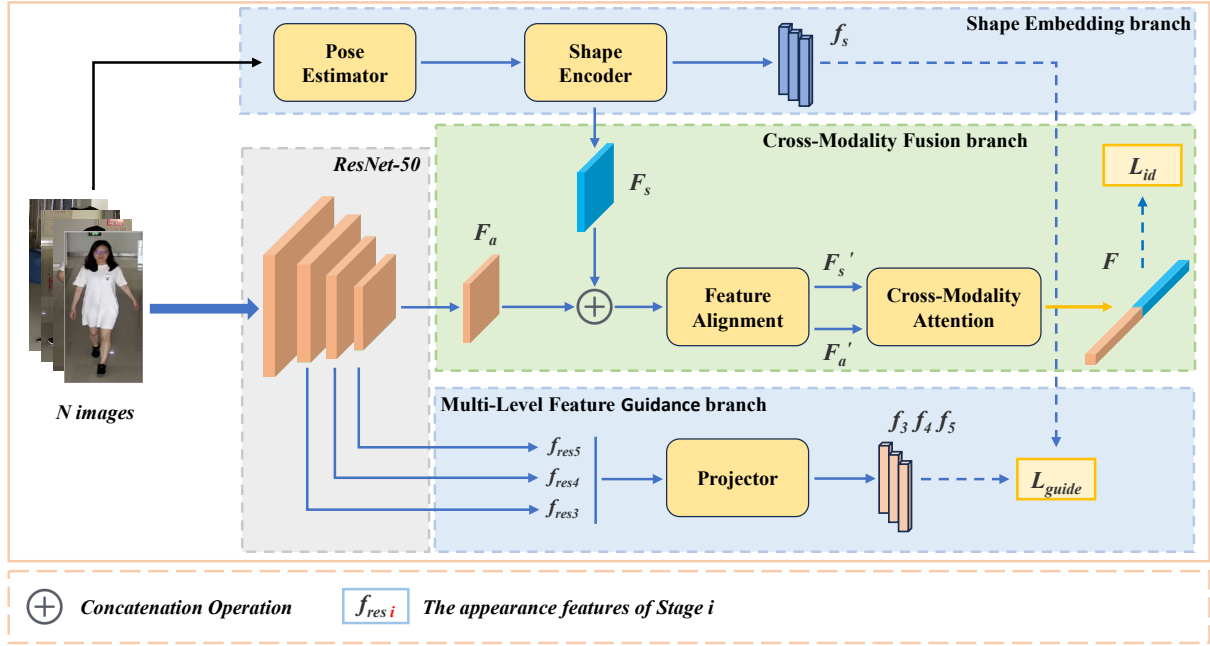
Figure 2: The overview of our MPCF framework. It consists of three branches: the SE branch, the CMF branch, and the MLFG branch. We use a frozen pose estimation model to introduce additional modal human shape information, and through multi-level learning of human pose information, we guide the model to retain more clothing-agnostic body shape semantic information when aggregating information across modalities, and the final output is a more robust fusion feature.

3. Methodology

3.1. Overview

In this section, we introduce our proposed MPCF framework in detail. The framework is mainly composed of three core branches: the Shape-Embedding (SE) branch, the Cross-Modality Fusion (CMF) branch, and the Multi-Level Feature Guidance (MLFG) branch, as shown in Fig. 2.

Specifically, given the person image $x \in \mathbb{R}^{H \times W \times 3}$, the SE branch extracts pose features from the original image and generates embedding information to supervise the MLFG branch. We use ResNet-50 [20] as the backbone to extract the person's global appearance features. These appearance features are then aligned and aggregated with the pose features from the SE branch via CMF, producing robust fused features. The MLFG branch extracts intermediate features from stages 3, 4, and 5 of the backbone network. Through a series of projection operations, it generates multi-level appearance embeddings, which are then aligned with the pose embeddings from SE. This alignment process helps guide the CMF branch during training to focus more on clothing-irrelevant identity information. The following sections will provide a detailed explanation of the specifics of each branch.

3.2. Shape-Embedding branch

To learn clothing-invariant discriminative features, we utilize the semantic information of human body shape, attributed to its stable manifestation across spatio-temporal scenarios and minimal impact from appearance changes. As shown in Fig. 2, the SE branch consists mainly of two modules: a pose estimator and a shape encoder. For the pose estimator, we adopt the well-established OpenPose [21] framework to extract pedestrian pose heatmaps.
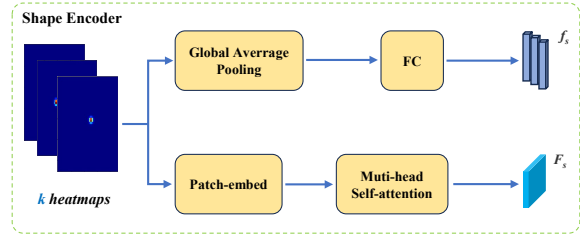

Figure 3: The Shape Encoder architecture outputs body shape embeddings f_s for multi-level pose-guided representations and pose features F_s for cross-modal information aggregation.

For a given input image x, OpenPose can generate k pose heatmaps, each heatmap highlights a key part of the human body (e.g., face, hands, feet). These heatmaps are then fed into the shape encoder to produce features related to overall body posture $f_s \in \mathbb{R}^{1 \times (H/8) \times (W/8)}$ and body semantic features $F_s \in \mathbb{R}^{h \times w \times D}$.

The structure of the Shape Encoder is depicted in Fig. 3 and includes two branches for processing the input pose heatmaps. The upper branch transforms the human pose heatmap into a body shape feature embedding $f_s \in \mathbb{R}^{1 \times 1152}$ through a global average pooling layer followed by a fully connected layer. To enable body shape information to interact more effectively with appearance information in the cross-modal fusion branch, the lower branch employs a method similar to CAMC [18] for shape embedding. This branch consists of an image patch embedding module and a multi-head attention module based on ViT [22]. The goal is to capture the relationships between different key points of the human body. The image patch embedding module processes the heatmap of size $h \times w$ by cutting it into a series of overlapping patches using a sliding window. The stride is denoted as S, and the patch size as P (e.g., 4), resulting in an

overlap between adjacent patches of shape $(P - S) \times P$. In this way, the entire heatmap is divided into N such patches.

$$N = N_h \times N_w = \lfloor \frac{H + S - P}{S} \rfloor \times \lfloor \frac{W + S - P}{S} \rfloor \quad (1)$$

Afterwards, through the self-attention mechanism, the patches are correlated with each other and thus learn to obtain more robust semantic features of human shapes $F_s \in \mathbb{R}^{288 \times 2048}$.

3.3. Cross-Modality Fusion branch

In our approach, we utilize ResNet-50 as the backbone network and set the stride of its fifth convolutional layer to 1. We extract the intermediate outputs from the third, fourth, and final layers to obtain multi-scale feature representations. Within this branch, we flatten the output features from the fifth layer to obtain the texture feature representation $F_a \in \mathbb{R}^{HW \times D}$. To prevent information loss when aggregating texture features F_a and body shape features F_a from different modalities, we first use a feature alignment module to concatenate the features from both modalities along the channel dimension, resulting in $F_{channel} = [F_a, F_s] \in \mathbb{R}^{HW \times 2D}$. Based on the channel attention mechanism [23], this module, which consists of two fully connected layers forming a bottleneck structure, is used to model the inter-channel relationships within $F_{channel}$ and outputs weights of the same quantity as the input features. We first reduce the feature dimension to one-fourth of the input (e.g., *D/2*), then pass it through a ReLU activation, and then through a fully connected layer to restore the original dimension, followed by a sigmoid to obtain normalized weight scores *s*. These weights *s* are then added to the channels of both modal features and summed with their original features to obtain the aligned features $F_a' \in \mathbb{R}^{HW \times D}$ and $F_s' \in \mathbb{R}^{HW \times D}$. The overall process can be represented as follows:

$$s = Sigmoid(W_2 ReLU(W_1 F_{channel} + b_1) + b_2) \quad (2)$$

$$F_a' = s[:, 0 : N] \otimes F_a + F_a \quad (3)$$

$$F_s' = s[:, N : 2N] \otimes F_s + F_s \quad (4)$$

where W_1 is the weight matrix of the first fully connected layer with dimensions $\mathbb{R}^{2D \times D/2}$, b_1 is its bias vector with dimensions $\mathbb{R}^{D/2}$. $W_2 \in \mathbb{R}^{D/2 \times 2D}$, and $b_2 \in \mathbb{R}^{2D}$. After aligning features from both modalities, we use a multi-head cross-modal attention module for adaptive fusion of texture and morphological semantic features. The queries, keys, and values in the attention block are represented as:

$$Q/K/V = Transpose_{(1,2)}(Reshape_{3D}(F)) \quad (5)$$

where *F* represents the features from both modalities, and $Reshape_{3D}$ indicates reshaping *F* into a three-dimensional feature map. To integrate information across different modalities, we use texture features and body shape information as queries, with corresponding body shape features and appearance features serving as keys and values:

$$F_{s \to a} = F_a' + Reshape_{2D}(MHA(Q_a, K_s, V_s)) \quad (6)$$

$$F_{a \to s} = F_s' + Reshape_{2D}(MHA(Q_s, K_a, V_a)) \quad (7)$$

This bidirectional access helps texture features to enhance shape features that are clothing-independent, while body shape features incorporate necessary identity traits, minimizing the information gap between modalities. The concatenated features F will be utilized to compute the ID recognition loss.

3.4. Multi-Level Feature Guidance branch

Furthermore, to fully leverage the body semantic information embedded in a person's appearance and reduce the interference of clothing information, we opt to use pose information as guidance on top of cross-modal aggregated appearance features and body semantic features. This approach aims to steer the model's focus towards regions closely related to posture, thereby enhancing recognition accuracy. Specifically, we align the body shape embeddings f_s obtained from the shape embedding branch with person feature embeddings. Without compromising other essential information, this highlights the pose information within person representations, allowing for the retention of more posture-related details during cross-modal feature aggregation. To capture richer original body shape information from images, we extract three levels of person appearance feature maps $f_{res3}, f_{res4}, f_{res5}$ from intermediate layers of the backbone network. These feature maps are then passed through a feature projection layer, which maps them into a feature space identical to the body shape embeddings f_s without significantly harming the original information, forming implicit multi-level person feature embeddings f_3, f_4, f_5. The projection layer is designed with linear projection, Transformer encoder, global pooling, and a normalization layer to ensure effective feature transformation and integration.

Ultimately, the person feature embeddings f_i (where $i = 3, 4, 5$) obtained will be combined with the body shape embeddings f_s to jointly compute the guidance loss. To ensure the alignment of information between the two and to emphasize the pose information within the person feature embeddings, we use the Kullback-Leible (KL) divergence as the guidance loss L_{guide} to consistently measure the similarity between the appearance embeddings f_i and the body shape embeddings f_s. The lower the value of L_{guide}, the more semantically consistent information the model has learned, meaning it can better capture features related to posture. The specific formulation of the overall loss function is as follows:

$$L_{total} = (1 - \lambda)L_{ID} + \lambda L_{guide} \quad (8)$$

where L_{ID} represents the identification loss function based on cross-entropy, with inputs being the cross-modal aggregated features F and the identity labels y_i, and λ is a fixed value. The L_{guide} function can be specifically expressed as:

$$KL(p|q) = \sum_k p_k log(p_k/q_k) \quad (9)$$

$$L_{gi} = \frac{1}{2} KL(softmax(f_s), softmax(f_i)), i \in 3, 4, 5 \quad (10)$$

$$L_{guide} = L_{g3} + L_{g4} + L_{g5} \quad (11)$$

Table 1

Statistics of the LT-ReID datasets.

Dataset	#Identities	#Images	#Cams
LTCC [5]	152	17,119	12
PRCC [16]	221	33,698	3
Celeb-reID [24]	1,052	34,186	-

Table 2

Our MPCF's performance is compared with other competitors on the clothing-change dataset LTCC, with CAMC as the baseline model. Red bold denotes the best performance, and black bold denotes the second-best performance.

Methods	Cloth-Changing		Standard	
	Rank-1	mAP	Rank-1	mAP
PCB [25]	23.5	10.0	61.8	27.5
CESD [5]	25.2	12.4	71.4	34.4
FSAM [26]	38.5	**16.2**	73.2	35.4
GI-ReID [2]	23.7	10.4	63.2	29.4
MBUNet [27]	**40.3**	15.0	67.6	34.8
ACID [28]	29.1	14.5	65.1	30.6
LDF [29]	32.9	15.4	**73.4**	**36.9**
CPC [30]	21.9	12.8	-	-
CAMC [18]	35.9	15.4	73.2	35.3
Ours	**40.5**	**15.7**	**75.4**	**37.3**

In the calculation of KL divergence, p and q represent two probability distributions, where p_k and q_k are the probabilities of these distributions for the k-th category, respectively. We obtain probability vectors for the person feature embeddings and body shape embeddings through normalization, and then compute the difference between them. The divergence value is divided by 2 to balance the scale of the loss function.

4. Experiment

4.1. Experimental Setup

Datasets. As shown in Table 1. To evaluate the effectiveness of our proposed MPCF framework, we conducted assessments primarily on three widely-used long-term clothing-change person Re-ID datasets: LTCC [5], PRCC [16], and Celeb-reID [24]. **LTCC** comprises 17,119 images of different individuals, covering 152 distinct identities and 416 different outfits, with an average of 5 varying outfits per person, and the number of outfit changes ranging from 2 to 14. **PRCC** includes 33,698 images of 221 individuals captured from three camera views. The training set consists of 150 individuals, while the test set comprises the remaining 71. During training, 25% of the images from the training set are used as a validation set. **Celeb-reID** utilizes street photos of celebrities to address long-term clothing changes. The dataset contains 34,186 images of 1,052 identities, each with unique clothing, thus presenting a greater challenge in clothing changes scenarios compared to the previous two datasets.

Implementation details. Our model is constructed on the PyTorch framework. We utilized a pre-trained ResNet-50 from ImageNet [31] as the backbone network to extract texture features of persons. The dimensions of the multi-level features extracted by the backbone network are 512, 1024, and 2048, respectively. All training was conducted on

Table 3

We evaluate MPCF against state-of-the-art methods on the PRCC clothing-change dataset, using CAMC as the baseline. Top and second-best results are highlighted in red and black bold, respectively. Results marked with † denote those reproduced from original implementations.

Methods	Cloth-Changing		Standard	
	Rank-1	mAP	Rank-1	mAP
PCB [25]	41.8	38.7	**99.8**	97.0
IANet [32]	46.3	45.9	99.4	98.3
SAGE [16]	34.4	-	64.2	-
RCSANet [33]	**50.2**	**48.6**	97.2	**100**
AFD-Net [34]	42.8	-	95.7	-
GI-ReID [2]	33.3	-	86.0	-
UCAD [35]	45.3	-	96.5	-
CPC [30]	40.8	-	-	-
CAMC [18]	47.0†	46.7†	99.7†	97.3†
Ours	**51.1**	**50.4**	**99.9**	**99.0**

a single NVIDIA RTX 3090 GPU. During both training and testing phases, images were resized to a uniform size of 384x192. Data augmentation included color jittering, random horizontal flipping, padding, random cropping, and random erasing [36]. We employed the Adam optimizer [37] for model optimization and performed 150 training epochs, with a warm-up strategy applied in the first 10 epochs, gradually increasing the learning rate from 3e-5 to 3e-4. The learning rate was reduced by 1/10 at epochs 40 and 80. For the PRCC and Celeb-reID datasets, the batch size was set to 48, while for the LTCC dataset, it was set to 32, with each identity ID having 4 images. For pose estimation, we used the OpenPose model pre-trained on the COCO dataset [38], generating 18 heatmaps, and we froze its weights during the training process.

Evaluation metrics. We employed the two standard metrics commonly used in most clothing-change Re-ID literature: mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC). To ensure a fair comparison with existing studies, we evaluated LTCC and PRCC under both standard and clothing-change settings. Under standard settings, the test set included both consistent and varied clothing samples. In clothing-change settings, the test set exclusively contained samples with varied outfits.

4.2. Performance Comparisons

Performance on the LTCC dataset. We evaluated our proposed method on the LTCC dataset and compared it with baseline models and other state-of-the-art clothing-change person re-identification approaches, as shown in Table 2. Compared to the baseline model, our model achieved improvements of +2.0% in mAP and +2.2% in R1 under standard settings. In clothing-change settings, compared to FSAM [26], although our method slightly underperformed in the mAP metric, it achieved a significant +2.0% improvement in the R1 metric. Moreover, compared to the second-best performing method LDF [29], our approach outperformed in both mAP and R1 metrics under both settings. It also surpassed the MBUNet [27] method, which had the second-best R1 performance in the clothing-change setting.

Performance on the PRCC dataset. We also assessed our proposed method on the PRCC dataset, with results shown in Table 3. It is noteworthy that the original baseline

Table 4
In the performance comparison on the clothing-change dataset Celeb-reID, our MPCF was tested against other competitors, with CAMC serving as the baseline model. Red bold indicates the best performance, while black bold denotes the second-best performance.

Methods	Celeb-reID		
	Rank-1	Rank-5	mAP
CESD [5]	50.9	66.3	9.8
RCSANet [33]	55.6	-	11.9
SirNet [39]	56.0	70.3	14.2
CT-Net [40]	60.2	**74.2**	13.7
ACID [28]	52.5	-	11.4
3DInvarReID [19]	**61.2**	-	**15.2**
MCSC [41]	57.8	-	13.0
CAMC [18]	57.5	71.5	12.3
Ours	**62.7**	**77.3**	**16.1**

Table 5
Ablation studies in the standard setting of the LTCC dataset. The effect of different levels of pose guidance pairings on model performance, with the best performance marked by red bolding and black bolding indicating the second best performance.

Methods	Rank-1	Rank-5	mAP
CAMC(baseline)	73.2	81.9	35.3
Ours+res3	73.2(+0.0)	82.1(+0.2)	36.0(+0.7)
Ours+res4	72.8(-0.4)	81.3(-0.6)	36.1(+0.8)
Ours+res5	**75.0(+1.8)**	**83.7(+1.8)**	**36.4(+1.1)**
Ours+res3+res4	72.4(-0.8)	81.7(-0.2)	**36.4(+1.1)**
Ours+res3+res5	72.0(-1.2)	80.7(-1.2)	35.8(+0.5)
Ours+res4+res5	73.6(+0.4)	82.3(+0.4)	**36.4(+1.1)**
backbone(ResNet-50)	69.1(-4.1)	-	33.1(-2.2)
Ours w/o CMF	72.2(-1.0)	81.3(-0.6)	34.5(-0.8)
MPCF	**75.4(+2.2)**	**83.5(+1.6)**	**37.3(+2.0)**

model was not evaluated on this dataset. We faithfully reproduced the experimental results by strictly adhering to the implementation protocols outlined in the original paper. It can be observed that under the clothing changes setting, our method significantly outperforms the baseline model on both the R1 and mAP metrics, with improvements of +4.1% and +3.7% respectively. Although the baseline model integrates clothing-agnostic pose information into person identity representation and minimizes the information discrepancy between appearance texture and pose features as much as possible, it is still inevitably affected by clothing information. Our method, however, with multi-level pose information supervision, can further reduce clothing noise. Other comparative results indicate that our method achieves comparable results with other advanced approaches.

Performance on the Celeb-reID dataset. Compared to the first two datasets, Celeb-reID is larger and more challenging, with images captured from uncontrolled street snapshots without any clothing annotations.

As shown in Table 4, all advanced methods perform relatively poorly. Competitors such as FSAM [26] and MBUNet [27] have not reported results in this area. Our method, MPCF, achieved notable performance improvements of 62.7%, 77.3%, and 16.1% in R1, R5, and mAP metrics, respectively. Compared to the baseline model, our method significantly improved by +5.2% in R1 and +3.8% in mAP. When compared to the second-best performing method, 3DInvarReID [19], our method improved by +0.9% in mAP and +1.5% in R1.

The performance results across the three datasets demonstrate that our approach helps person re-identification models prioritize pose information over clothing during training, effectively addressing the challenge of clothing changes.

4.3. Ablation Study

Component Analysis. To demonstrate the effectiveness of our approach, we evaluated the multi-level pose guidance and the effectiveness of the two branches, SE and CMF, on the LTCC dataset under the standard Settings and compared them with the baseline model. The results are shown in Table 5.

In single-level guidance, the pose guidance at stage 5 showed the most significant improvement over the baseline model. The guidance at stages 3 and 4 resulted in slight

increases in mAP, but there was no noticeable improvement in the Rank metrics, and even a decrease was observed. When combining two levels of guidance, the joint pose guidance at stages 4 and 5 performed the best, while the other two methods improved mAP but did not perform well on the Rank metrics. Ultimately, our method integrated guidance across three levels, and after experimentation, the optimal weight ratio for the three levels in the guidance loss was found to be 5:3:2, achieving the best overall performance. Compared to other methods, our approach achieved the best results in both R1 and mAP metrics. This also confirms the effectiveness of using multi-level guidance for pose information.

To show our framework is effective, we did ablation studies on its branches. Since all branches use pedestrian pose features, removing the SE branch leaves only the ResNet-50 backbone working. This leads to much worse performance on the LTCC dataset, as shown in Table 5. If we remove the CMF branch, the model loses key info due to the difference between pose and appearance features, harming performance. The final MPCF results prove the CMF branch's cross-modal fusion is necessary.

Computational Complexity Analysis. We systematically evaluated the impact of adding three levels of pose guidance components on the model under the PRCC dataset's cloth-changing setting, focusing on changes in computational cost and performance improvements. The results are shown in Table 6.

Experiments show that the introduction of a single level pose guidance component leads to a significant increase in the training parameters (Params) of the model, but the increase in the computational time complexity (FLOPs) of the model is small. This is mainly because the projection module in the MLFG branch uses fully connected layers and Transformer encoders, which add parameters but have relatively low computational complexity. Furthermore, our MPCF framework integrates all three levels of pose guidance components. Compared to the baseline model, while Params increased by 25%, the performance metrics showed significant improvements: Rank-1 improved by +4.1%, and mAP improved by +3.7%. Meanwhile, the increase in FLOPs remained small, indicating that the computational complexity did not rise significantly.

This design shows that our method can achieve significant performance improvements with limited computational cost,

Table 6

Compared to the baseline under the cloth-changing setting of the PRCC dataset, we assessed the impact of gradually adding three levels of guidance components on the model's computational cost and performance. MPCF represents the method that includes all levels of guidance components. Results marked with "†" are reproduced by us based on the original implementation.

Methods	Computational Cost		Metrics	
	Params(M)↓	FLOPs(G)↓	Rank-1↑	mAP↑
CAMC [18]	62.33	20.2	47.0†	46.7†
Our+res3	68.39(+6.06)	20.3(+0.1)	47.6(+0.6)	49.2(+2.5)
Our+res3+res4	73.47(+11.14)	20.39(+0.19)	48.7(+1.7)	48.7(+2.0)
MPCF	**78.26(+15.63)**	**20.48(+0.28)**	**51.1(+4.1)**	**50.4(+3.7)**

Figure 4: Visualization of retrieval results. The left side of (a) and (b) is the input query image. For the right side, the first and the second row are the ordered matching results obtained by using the baseline model and MPCF, respectively. Images with green borders and red borders indicate correct and error matching results, respectively.

proving that these additions are worthwhile.

4.4. Visualization of retrieval results

Our proposed method integrates multi-modal feature fusion and multi-level pose guidance to better address the challenges of person re-identification in long-term clothing changes scenarios. To visually demonstrate this conclusion, we visualized the top-10 retrieval results of the baseline model CAMC and our method on the LTCC dataset under clothing changes settings, as shown in Fig. 4.

Our proposed model significantly reduces the dependency on clothing information during the identification process. As shown in the first row of Fig. 4(a), the baseline model's matching results mostly display persons with similar clothing but different identities compared to the query image. In contrast, as depicted in the second row of Fig. 4(a), our method's matching results can still effectively identify the correct person identities even in clothing changes scenarios where there may be similarities between samples of different categories. Additionally, as demonstrated in the results

of Fig. 4(b), due to the interference of clothing information, the top retrieval results in the first row are images with similar clothing textures and colors. However, thanks to the multi-level pose guidance in our approach, the model focuses more on body shape information that is independent of clothing. Consequently, in the second row of results shows that even when the queried person is wearing different clothing, our model can still achieve more robust person identity representations.

5. Conclusion

To mitigate information interference caused by long-term and cross-scenario appearance variations in persons, this paper proposes a Multi-Level Pose-Guidance with Cross-Modality Fusion for Long-Term Spatio-Temporal Re-ID (MPCF). Specifically, we introduce additional modality human pose feature embeddings through a SE branch, supplementing identity information independent of clothing. Then, a CMF branch reduces the modality gap between person appearance features and pose features, preventing the

loss of key information across modalities when aggregating clothing-independent features. Furthermore, to further reduce the model's focus on clothing information during the aggregation process, we propose a MLFG branch that uses multi-level person pose embeddings as guidance, compelling the model to concentrate attention on clothing-independent information areas, ensuring that aggregated features include more clothing-independent, distinctive identity information. Our proposed method has been extensively tested on multiple datasets, validating its effectiveness.

Acknowledgments

References

[1] X. Jia, X. Zhong, M. Ye, W. Liu, W. Huang, S. Zhao, Patching your clothes: Semantic-aware learning for cloth-changed person re-identification, in: International Conference on Multimedia Modeling, Springer, 2022, pp. 121–133.

[2] X. Jin, T. He, K. Zheng, Z. Yin, X. Shen, Z. Huang, R. Feng, J. Huang, Z. Chen, X.-S. Hua, Cloth-changing person re-identification from a single image with gait prediction and regularization, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 14278–14287.

[3] M. Liu, Z. Ma, T. Li, Y. Jiang, K. Wang, Long-term person re-identification with dramatic appearance change: Algorithm and benchmark, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 6406–6415.

[4] Y. Xian, J. Yang, F. Yu, J. Zhang, X. Sun, Graph-based self-learning for robust person re-identification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 4789–4798.

[5] X. Qian, W. Wang, L. Zhang, F. Zhu, Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, Long-term cloth-changing person re-identification, in: Proceedings of the Asian Conference on Computer Vision, 2020.

[6] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 152–159.

[7] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, S. Z. Li, Salient color names for person re-identification, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, Springer, 2014, pp. 536–551.

[8] O. Oreifej, R. Mehran, M. Shah, Human identity recognition in aerial images, in: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE, 2010, pp. 709–716.

[9] R. R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, A siamese long short-term memory architecture for human re-identification, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14, Springer, 2016, pp. 135–153.

[10] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, Q. Tian, Person re-identification in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1367–1376.

[11] F. Herzog, X. Ji, T. Teepe, S. Hörmann, J. Gilg, G. Rigoll, Lightweight multi-branch network for person re-identification, in: 2021 IEEE international conference on image processing (ICIP), IEEE, 2021, pp. 1129–1133.

[12] G. Wang, Y. Yuan, X. Chen, J. Li, X. Zhou, Learning discriminative features with multiple granularities for person re-identification, in: Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 274–282.

[13] Y. Suh, J. Wang, S. Tang, T. Mei, K. M. Lee, Part-aligned bilinear representations for person re-identification, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 402–419.

[14] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, W. Gao, Attention driven person re-identification, Pattern Recognition 86 (2019) 143–155.

[15] L. Zhao, X. Li, Y. Zhuang, J. Wang, Deeply-learned part-aligned representations for person re-identification, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3219–3228.

[16] Q. Yang, A. Wu, W.-S. Zheng, Person re-identification by contour sketch under moderate clothing change, IEEE transactions on pattern analysis and machine intelligence 43 (2019) 2029–2046.

[17] J. Chen, X. Jiang, F. Wang, J. Zhang, F. Zheng, X. Sun, W.-S. Zheng, Learning 3d shape feature for texture-insensitive person re-identification, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 8146–8155.

[18] Q. Wang, X. Qian, Y. Fu, X. Xue, Co-attention aligned mutual cross-attention for cloth-changing person re-identification, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 2270–2288.

[19] F. Liu, M. Kim, Z. Gu, A. Jain, X. Liu, Learning clothing and pose invariant 3d shape representation for long-term person re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 19617–19626.

[20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[21] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7291–7299.

[22] D. Alexey, An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv: 2010.11929 (2020).

[23] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[24] Y. Huang, Q. Wu, J. Xu, Y. Zhong, Celebrities-reid: A benchmark for clothes variation in long-term person re-identification, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.

[25] Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond

part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 480–496.

[26] P. Hong, T. Wu, A. Wu, X. Han, W.-S. Zheng, Fine-grained shape-appearance mutual learning for cloth-changing person re-identification, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 10513–10522.

[27] G. Zhang, J. Liu, Y. Chen, Y. Zheng, H. Zhang, Multi-biometric unified network for cloth-changing person re-identification, IEEE Transactions on Image Processing 32 (2023) 4555–4566.

[28] Z. Yang, X. Zhong, Z. Zhong, H. Liu, Z. Wang, S. Satoh, Win-win by competition: Auxiliary-free cloth-changing person re-identification, IEEE Transactions on Image Processing 32 (2023) 2985–2999.

[29] P. P. Chan, X. Hu, H. Song, P. Peng, K. Chen, Learning disentangled features for person re-identification under clothes changing, ACM Transactions on Multimedia Computing, Communications and Applications 19 (2023) 1–21.

[30] M. Li, S. Cheng, P. Xu, X. Zhu, C.-G. Li, J. Guo, Unsupervised long-term person re-identification with clothes change, in: 2023 8th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC), IEEE, 2023, pp. 167–171.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (2015) 211–252.

[32] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, X. Chen, Interaction-and-aggregation network for person re-identification, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9317–9326.

[33] C. Yan, G. Pang, J. Jiao, X. Bai, X. Feng, C. Shen, Occluded person re-identification with single-scale global representations, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 11875–11884.

[34] W. Xu, H. Liu, W. Shi, Z. Miao, Z. Lu, F. Chen, Adversarial feature disentanglement for long-term person re-identification., in: IJCAI, 2021, pp. 1201–1207.

[35] Y. Yan, H. Yu, S. Li, Z. Lu, J. He, H. Zhang, R. Wang, Weakening the influence of clothing: Universal clothing attribute disentanglement for person re-identification., in: IJCAI, 2022, pp. 1523–1529.

[36] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: Proceedings of the AAAI conference on artificial intelligence, volume 34, 2020, pp. 13001–13008.

[37] P. K. Diederik, Adam: A method for stochastic optimization, (No Title) (2014).

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[39] S. Yang, B. Kang, Y. Lee, Sampling agnostic feature representation for long-term person re-identification, IEEE Transactions on Image Processing 31 (2022) 6412–6423.

[40] J. Wu, Y. Huang, M. Gao, Z. Gao, J. Zhao, H. Zhang, A. Zhang, A two-stream hybrid convolution-transformer network architecture for clothing-change person re-identification, IEEE Transactions on Multimedia (2023).

[41] Y. Huang, Q. Wu, Z. Zhang, C. Shan, Y. Zhong, L. Wang, Meta clothing status calibration for long-term person re-identification, IEEE Transactions on Image Processing (2024).