Proceedings of the 2nd International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/95/2024.21337

Enhancing cloth-changing person re-identification with Silhouette-Keypoint fusion

Xintong Li

School of Computer Science and Communication Engineering, JiangSu University, Jiangsu, China

3210812039@stmail.ujs.edu.cn

Abstract. In recent years, person re-identification (ReID) has experienced significant advancements due to its diverse real-world applications. However, traditional benchmarks often assume consistent attire across captured images, failing to reflect the reality of pedestrians frequently changing their clothing. This discrepancy has led to the emergence of the cloth-changing person re-identification (CC-ReID) problem and the development of relevant benchmarks. CC-ReID poses a substantial challenge, as pedestrians' primary visual cues, particularly their clothing, vary between query and gallery images, while non-attire-related features remain relatively weak. To address this gap and advance research in CC-ReID, this paper introduces a novel task termed Silhouette-Keypoint Fusion Re-Identification (SKF-ReID). This represents a dual-stream framework capable of extracting silhouette and keypoint details within the shape stream, subsequently transferring this data to the ReID stream to enrich appearance features with clothing-independent insights. Additionally, we employ the Maximum Mean Discrepancy (MMD) loss to ensure similarity between shape and appearance features, thereby enhancing the accuracy of cloth-changing person re-identification. Our proposed approach undergoes rigorous evaluation across benchmark cloth-changing person re-identification datasets, showcasing cutting-edge performance.

Keywords: Cloth-changing Person ReID, Clothing Attention Degradation, Human Semantic Attention

1. Introduction

Person Re-Identification (ReID) is a task in computer vision and machine learning aimed at matching and identifying people across disjointed cameras, locations, and time frames. Recent advancements in ReID methods have addressed challenges like diverse human poses [1] and variations in image styles and scales [2]. However, many existing methods assume that query and gallery images of the same person have identical clothing. Therefore, researchers utilize clothing information for ReID. However, in long-term ReID dataset testing, methods notably decline with clothing variations. As shown in figure 1, figure 1-(a) depicts the same person in identical clothes across five images, while figure 1-(b) shows people in different attire, complicating identification. This poses challenges to ReID, especially in low-quality images or when pedestrians wear masks. In real-world scenarios like long-term ReID and pedestrian tracking, clothing changes pose a significant challenge, demanding specialized methods for distinguishing people across different outfits.

DOI: 10.54254/2753-8818/95/2024.21337

Current research on Cloth-Changing Person Re-Identification (CC-ReID) aims to model identity information effectively while mitigating the impact of clothing variations. Existing methods often struggle to remove clothing details explicitly, leaving them vulnerable to clothing color effects. While some use silhouette-based techniques to overcome this, they sacrifice joint information. To address this, we propose a combined approach leveraging both silhouette and keypoint extraction, ensuring robust ReID by managing clothing variations while preserving crucial anatomical details.

Figure 1. Examples of cloth-changing person re-identification images. (a) presents three sets of images, with each set featuring the same person wearing identical clothes. (b) displays four sets of images, with each set depicting the same person wearing different clothes.

We introduce a solution named the Silhouette-Keypoint Fusion Re-Identification (SKF-ReID) to tackle the aforementioned issues and delve into discriminative body shape knowledge. The framework comprises two components, the ReID Stream and the Shape Stream, as illustrated in figure 2.

In the ReID Stream, we use ResNet [3] for extracting appearance features from RGB images, incorporating identity and triplet losses to ensure similar representations for individuals of the same identity while distinguishing between different ones. In the Shape Stream, we utilize DeepLabV3 [4] for silhouette information and OpenPose [5] for keypoint extraction. We generate heatmap features for each image and concatenate them with the extracted silhouette features. Additionally, we employ Maximum Mean Discrepancy (MMD) loss to ensure similarity between features from the Shape stream and ReID stream in the semantic space, enhancing the system's resilience and accuracy in pedestrian identification despite clothing changes.

Therefore, the contributions of this paper can be summarized as follows:

1)We tackle the clothing challenge to enhance practical usability by introducing the Silhouette-Keypoint Fusion Re-Identification (SKF-ReID) framework. It leverages silhouettes and keypoint data, allowing the removal of the shape stream during inference.

2)We investigated combining silhouettes and keypoints to improve CC-ReID performance. Our approach leverages this collaboration to capture people's distinctive characteristics more accurately.

3)We rigorously assessed SKF-ReID on the PRCC [6] dataset, highlighting its superior performance in CC-ReID. Our findings indicate that SKF-ReID generates strong and distinctive feature representations, surpassing state-of-the-art methods in metrics such as mAP and rank-1.

2. Related work of cloth-changing person ReID

The key problem in CC-ReID lies in extracting clothing-agnostic features to establish a more robust and adaptive feature representation. Several methods address this challenge. Gu et al. [7] achieved accurate recognition using only RGB images. Jin et al. [8] combined gait features and pose estimation. Gao et al. [9] leveraged identity information for collaborative learning, while Yang et al. [10] modeled and mitigated clothing variations using causal relationships. Li et al. [11] introduced a large-scale CC-ReID method based on clothing templates, and Hong et al. [12] learned the mutual relationship between shape and appearance information. Additionally, recent efforts by Zhang et al. [13] incorporated spatial-temporal attention mechanisms to improve performance in dynamic environments. Our proposed SKF-

DOI: 10.54254/2753-8818/95/2024.21337

ReID framework learns RGB image features through silhouette and keypoint information, establishing a robust representation. By applying Maximum Mean Discrepancy (MMD) loss, clothing-agnostic features are encouraged to align with features from the ReID stream, improving re-identification performance under clothing changes.

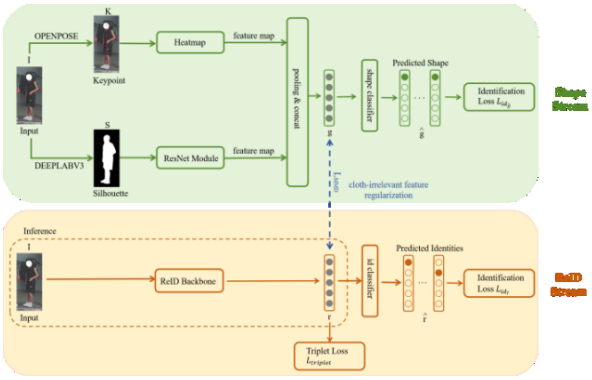

Figure 2. Structure of SKF-ReID involves the ReID Stream and Shape Stream, jointly trained with cloth-irrelevant feature regularization. The Shape Stream guides the ReID Stream to obtain clothing-agnostic representations, with the Shape Stream excluded during inference for efficiency.

3. Method

The SKF-ReID framework utilizes human silhouette and keypoint features to handle clothing change challenges in ReID. Figure 2 depicts the framework's structure. From a single-person image, we extract silhouette using DeepLabV3 and keypoints using OpenPose, serving as inputs to the shape stream. ResNet extracts features from the silhouettes, and heatmap features are concatenated. By employing Maximum Mean Discrepancy (MMD) as a constraint, the clothing-agnostic shape stream acts as a regulator, encouraging the main ReID stream to capture clothing-agnostic features from the single RGB image. Further details on each component will be discussed in the following sections.

3.1. The Auxiliary Shape Stream

In ReID tasks, clothing variations challenge recognition accuracy. To overcome this problem, we propose an auxiliary shape stream for extracting clothing-agnostic shape features. This stream includes silhouette extraction using DeepLabV3 and keypoint information extraction via OpenPose.

Silhouette Extraction. We utilize DeepLabV3 to extract pedestrian contour shapes for silhouette extraction. Specifically, DeepLabV3 can be represented as:

$$S = DeepLabV3(I), \tag{1}$$

Here, I represents the input RGB pedestrian image, and S denotes the extracted silhouette.

Keypoint information extraction. We employ OpenPose for keypoint information extraction in our auxiliary shape stream. In the shape stream, following [13], we generate heatmap features for each image. Specifically, OpenPose and heatmap features can be represented as:

$$K = OpenPose(I), X_{heatmap} = Heatmap(I, K), \tag{2}$$

Here, K denotes the extracted coordinates of keypoints. $X_{heatmap}$ represents the heatmap features.

Feature fusion. In the feature fusion of the shape stream, we merged silhouette and heatmap feature using MaxPooling to extract key information. These fused features were then concatenated to form the final shape feature representation, defined as follows:

$$SF = MaxPool\left(X_{silhouette}\right), HF = MaxPool\left(X_{heatmap}\right), g = [SF, HF], \tag{3}$$

DOI: 10.54254/2753-8818/95/2024.21337

Here, SFrepresents the silhouette feature obtained by MaxPooling on the silhouette representation $X_{silhouette}$. HF represents the heatmap feature. g is the concatenated feature vector.

After generating the concatenated feature g, we pass it through the shape classifier to obtain $\hat{g} = Shape\ \ Classifier.(g)$. This maps g to $\hat{g}$, both having the same feature dimensions, introduced into a common interaction space. Then, $\hat{g}$ is utilized to compute the id loss for the shape stream, commonly used in classification tasks, defined as:

$$L_{id_{\hat{g}}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C} y_{i,k}\, log(p_{i,k}^{g}), \tag{4}$$

Here, $N$and C represent the number of samples and the number of classes, respectively. $y_{i,k}$ denotes the label of class k for sample i, and $p_{i,k}^{g}$ is the predicted probability of class k for sample i.

3.2. The Main ReID Stream

The ReID stream employs a ResNet50 [3] as its backbone architecture and is trained with the ReID feature r using classification loss and triplet loss with batch hard mining as primary objectives. Specifically, we utilize triplet loss $L_{triplet}$ to optimize the r feature.

After generating the ReID feature vector r, we transform it using the id classifier to obtain $\hat{r} = Id\ \ Classifier.(r)$. This maps r to $\hat{r}$. Classification loss is utilized, defined as follows:

$$L_{id_{\hat{r}}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j}\, log(p_{i,j}^{r}), \tag{5}$$

Following $L_{id_{\hat{g}}}$.

Training the ReID stream with these loss functions aims to learn discriminative features invariant to clothing variations, enabling accurate pedestrian matching across different conditions.

3.3. Cloth-irrelevant feature regularization

Cloth-irrelevant feature regularization uniquely preserves shared features between clothing-agnostic features and RGB images. Figure 2 illustrates this using Minimum-Maximum Mean Discrepancy (MMD) to align ReID feature r and concatenated feature g. This drives the ReID stream to focus on silhouette and keypoint features independently of clothing. The original form of the MMD loss is rather complex, here we use a simple and empirical format, defined as follows:

$$L_{MMD}(r,g) = \left\| \frac{1}{n}\sum_{i=1}^{n}\varphi(r_i) - \frac{1}{m}\sum_{j=1}^{m}\varphi(g_j) \right\|_{H}^{2}, \tag{6}$$

Here, r and g represent the ReID feature and the concatenated feature, respectively. r_i and g_j are samples from these feature distributions, and n and m are the sample sizes in each distribution. The function φ maps the features to a common and interactive space.

Overall Loss Function. To train our SKF Module effectively, we employ a composite loss function consisting of four components:

$$L = L_{id_{\hat{g}}} + L_{id_{\hat{r}}} + L_{triplet} + L_{MMD}, \tag{7}$$

Inference. During inference in SKF-ReID, we only use the main stream, combining the shape stream and the ReID stream, for computational efficiency.

4. Experiment

4.1. Datasets and protocols

Dataset. We evaluated the proposed method on the PRCC dataset for CC-ReID, which includes 33,698 images from three cameras commonly used in ReID research. These images depict 221 different people in diverse indoor scenes. In images from Camera A and Camera B, people wear the same attire, while in images from Camera A and Camera C, they wear different clothing.

DOI: 10.54254/2753-8818/95/2024.21337

Evaluation Protocol. We utilize rank-1 accuracy and mAP as evaluation metrics, with two test settings. 1) Clothing Change (CC): Only samples with clothing changes are used to calculate accuracy, and 2) Same Clothing (SC): Only samples with consistent clothing are used to calculate accuracy.

4.2. Implementation details

We utilized ResNet50 [3] as the backbone for our person ReID model. To boost CC-ReID accuracy, we omitted the final downsampling layer of ResNet50. For the PRCC dataset, we adopted the feature map aggregation method from [14]. Global average pooling and max pooling operations were applied to the feature maps, followed by concatenation. Batch normalization [15] was utilized to normalize image features. Input images were resized to 384×192 following [16]. Data augmentation included random horizontal flipping, cropping, and erasing. The batch size was set to 64, comprising 8 pedestrians with 8 images per person. Training employed the Adam optimizer [17] for 60 epochs.

Table 1. Performance evaluation and comparison of 15 ReID methods on the PRCC dataset. Specifically, the bolded data in the table has the best performance, while the underlined data ranks second in performance

Method	Modality	PRCC			
		Standard		Cloth-changing	
		Rank-1	mAP	Rank-1	mAP
MGN [18]	RGB	99.5	98.4	33.8	35.9
PCB [19]	RGB	99.8	97.0	41.8	38.7
CAL [7]	RGB	**100**	<u>99.8</u>	55.2	55.8
GI-ReID [8]	Multi-modality	-	-	37.6	-
FSAM [12]	Multi-modality	98.8	-	54.5	-
RCSANet [14]	RGB	**100**	97.2	48.6	50.22
AIM [10]	RGB	**100**	**99.9**	<u>57.9</u>	**58.3**
SKF	RGB+silhoutette	<u>99.9</u>	<u>99.8</u>	**58.1**	<u>57.2</u>

4.3. Comparison with the state-of-the-art methods

We compared SKF-ReID with the CC-ReID model on the PRCC dataset, as shown in table 1. SKF-ReID exhibits significant advantages on the cloth-changing dataset. In the cloth-changing mode, it achieves comparable accuracy to the top-ranked model AIM. In the standard mode, SKF-ReID achieves near-saturation performance with 99.8% mAP and 99.9% rank-1 accuracy. However, compared to CAL, SKF-ReID has a slight disadvantage as it aims to learn clothing-independent features. Yet, in this mode, only ground truth samples consistent with the clothing are available. Our approach integrates multiple information sources and prioritizes efficiency in both training and inference, offering advantages over existing methods proposed for cloth-changing scenarios.

4.4. Ablation study

Results of Real Cloth-Changing Image ReID. Baseline represents the model that uses only RGB images. Ablation experiments on PRCC dataset, as detailed in table 2, reveal that: 1) Silhouette-based schemes outperform baseline by over 3.9% in mAP, addressing cloth-changing issues effectively. 2) Pose-based schemes also surpass baseline by over 7.0% in mAP, indicating keypoint information's efficacy. 3) Integrating silhouette and keypoint information via SKF-ReID module achieves 10.0% mAP increase over baseline+silhouette and 6.9% increase over baseline+pose, enhancing recognition accuracy for cloth-changing pedestrians. 4)Cloth-irrelevant feature regularization improves performance while reducing computational cost during inference by discarding shape stream.

DOI: 10.54254/2753-8818/95/2024.21337

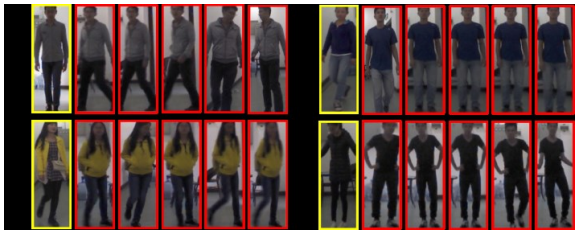

Figure 3. Failure case: People in similar clothes are wrongly thought to be the same. In each set, those outlined in yellow and red aren't the same but are mistakenly identified.

Table 2. ablation study of SKF on the PRCC dataset.

Method	PRCC			
	Standard		Cloth-changing	
	Rank-1	mAP	Rank-1	mAP
baseline	99.8	97.9	45.6	43.3
baseline+silhouette	99.8	98.1	53.5	47.2
baseline+pose	99.8	99.2	54.1	50.3
baseline+silhouette&pose	**99.9**	**99.8**	**58.1**	**57.2**

4.5. Limitation

Our network sometimes misclassifies people due to visually similar clothing, as our method relies on an auxiliary stream for regulating design choices, providing a more subtle approach to clothing issues. We propose the SKF framework, which integrates an auxiliary shape stream with the main stream. While the main stream extracts appearance features, the auxiliary stream focuses on non-clothing features, reducing errors. However, limitations are apparent. In some cases, visually indistinguishable people with similar clothing were misclassified as the same person, as shown in figure 3. These errors highlight challenges in our approach, particularly when appearances are similar but identities differ. Despite using the auxiliary shape stream to extract clothing-irrelevant features and addressing clothing change issues, challenges remain. Further research is necessary to refine the SKF model for accurately distinguishing people with similar appearances but different identities.

5. Conclusion

We tackled cloth-changing person re-identification, where attire variations and personal belongings may obscure a pedestrian's silhouette and keypoint features. We proposed Silhouette-Keypoint Fusion ReIdentification (SKF-ReID), comprising ReID stream and Shape stream. The Shape stream extracted silhouette and keypoint information from RGB images, leveraging ResNet for silhouette features and heatmap transformation for keypoints. Concatenating these with silhouette features yielded clothing-agnostic characteristics. Utilizing Maximum Mean Discrepancy (MMD) loss bridged the semantic gap between the ReID stream and clothing-agnostic features, enhancing overall discriminative capability. Our method excelled on cloth-changing ReID datasets, positioning itself at the field's forefront.

References

[1] Zhao, H., Tian, M., Sun, S. Shao, J., & Tang, X. (2017). Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4929-4938). IEEE.

[2] Jin, X., Lan, C., Zeng, W., Wei, G., & Chen, Z. (2019). Semantics-Aligned Representation Learning for Person Re-identification. arXiv preprint arXiv:1905.13143. DOI: 10.48550/ arXiv.1905.13143.

[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition, in CVPR, 2016, pp. 2, 4, 5.

[4] Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. . (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. (2017). DOI: 10.48550/arXiv.1706.05587.

[5] Cao, Z., Hidalgo, G., Simon, T., Wei, S. E., & Sheikh, Y. (2018). Openpose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(12), 2893-2906.

[6] Q. Yang, A. Wu, W.-S. Zheng, (2019). Person reidentification by contour sketch under moderate clothing change, in TPAMI, 2019, pp. 1, 2, 5, 8.

[7] X. Gu, H. Chang, B. Ma, S. Bai, S. Shan, & X. Chen. (2022). Clothes-changing person re-identification with RGB modality only, in CVPR, 2022, pp. 1, 3, 6, 7.

[8] Jin, X., He, T., Zheng, K., Yin, Z., Shen, X., Huang, Z., Feng, R., Huang, J., Hua, X., & Chen, Z. (2021). Cloth-Changing Person Re-identification from A Single Image with Gait Prediction and Regularization. arXiv preprint arXiv:2103.15537. DOI: 10.48550/arXiv.2103.15537.

[9] Gao, Z., Wei, S., Guan, W., Zhu, L., Wang, M., & Chen, S. (2023). Identity-Guided Collaborative Learning for Cloth-Changing Person Reidentification. arXiv preprint arXiv:2304.04400.

[10] Yang, Z., Lin, M., Zhong, X., Wu, Y., & Wang, Z. (2023). Good is Bad: Causality Inspired Cloth-debiasing for Cloth-changing Person Re-identification. Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[11] Li, S., Chen, H., Yu, S. He, Z., Zhu, F., Zhao, R., Chen, J., & Qiao, Y. (2017). COCAS+: Large-Scale Clothes-Changing Person Re-Identification With Clothes Templates. In IEEE, 2023. 4

[12] Hong, P., Wu, T., Wu, A., Han, X., & Zheng, W. S. (2021). Fine-Grained Shape-Appearance Mutual Learning for Cloth-Changing Person Re-Identification. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), DOI: 10.1109/CVPR46437. 2021.01037.

[13] Miao, J., Wu, Y., Liu, P., Ding, Y., & Yang, Y. . (2019). Pose-Guided Feature Alignment for Occluded Person Re-Identification. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 3996-4005). IEEE.

[14] Huang. Y, Wu. Q, Xu. J, Zhong. Y, & Zhang. Z. (2021). Clothing status awareness for long-term person re-identification, in ICCV, 2021, pp. 1, 3, 7.

[15] Ioffe, S., Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Journal of Machine Learning Research, 2015. doi: 10. 48550/arXiv.1502.03167.

[16] Qian, X. , Wang, W. , Zhang, L. , Zhu, F. , & Xue, X. (2021). Long-term cloth-changing person re-identification, in ACCV, 2020, pp. 1-3, 6, 7.

[17] Kingma, D., Ba, J. (2014). Adam: A Method for Stochastic Optimization. Computer Science, 2014. doi: 10.48550/arXiv.1412.6980.

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

[19] Sun, Y., Zheng, L., Yang, Y., Tian, Q., & Wang, S. (2017). Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In European Conference on Computer Vision (pp. 418-434). Springer, Cham.