Unsupervised Clothing Change Adaptive Person ReID

Ziyue Zhang [ID], Shuai Jiang [ID], Congzhentao Huang, and Richard Yi Da Xu [ID]

Abstract—**Clothing changes and lack of data labels are both crucial challenges in person ReID. For the former challenge, people may occur multiple times wearing different clothes. A majority of current research works for person ReID focuses on benchmarks in which a person's clothing is kept the same all the time. For the latter challenge, most researchers try to transfer information from a labeled source dataset to an unlabeled target dataset. Whereas purely unsupervised training is less used. In this letter, we aim to solve both problems simultaneously. We design a novel unsupervised model, Syn-Person-Cluster ReID, to solve the unlabeled clothing change person ReID problem. We develop a purely unsupervised pipeline equipped with two innovations: synthetic augmentation on person images and feature restriction for the same person. The synthetic augmentation is to supply additional information for the same person, which can then be used as partially supervised inputs to the feature restriction. Extensive experiments on clothing change ReID datasets show the out-performance of our methods.**

Index Terms—**Unsupervised learning, clothing change adaptive, person image data augmentation, person ReID.**

I. INTRODUCTION

PERSON re-identification (ReID) [1] is designed to match specific pedestrians in images or video sequences. Current ReID methods are mostly to learn discriminative features of person identity by a specifically designed backbone [2], [3]. There are also works focusing on problems of occlusions [4], different modalities [5], illumination changes [6] and resolution changes [7].

The clothing change plays a crucial character in person ReID. However, most researchers assume the same person wears the same clothes, which means the above works cannot handle the clothing change variations in person ReID. In daily life, people usually change their clothes and the same person captured by the camera at different times may wear different clothes. Existing conventional methods tend to fail in such a scenario because of the unreliability of clothing texture information and the

lack of identity information about the same person in different clothes.

Another key problem is that it is challenging to get person ReID labels. Because it is costly to record person images across multiple cameras, many previous works try to address person ReID problem by unsupervised learning [8]–[10]. There are two kinds of unsupervised ReID methods. One is the unsupervised domain adaptation, which first pretrains a model on the source labeled dataset and then fine-tunes it on the target unlabeled dataset. However, this kind of methods still needs at least one labeled dataset and suffers from the variation between the source and target domains. Another one is the purely unsupervised learning which eliminates the above shortcomings and is exactly what we use in our work.

In this letter, we propose a purely unsupervised model called synthetic person images and cluster based unsupervised ReID model (Syn-Person-Cluster) to solve the unlabeled clothing change person ReID problem. The whole pipeline follows the below procedure. First of all, the features of training person images are extracted by a CNN backbone. Secondly, a clustering algorithm is adopted to cluster image features and produce pseudo labels. At last, the backbone is trained with a contrastive loss using the person image features from the storage, which is updated iteratively. This pipeline is already capable of initially obtaining pedestrian features. However, it is not able to separate the noise information caused by different clothes and the obtained features are not robust enough. So to better shrink the gap among the features of the same person in different clothes and improve the unsupervised performance, we have added several innovations summarised as follows:

1) A clothing change augmentation module is proposed to generate synthetic person images in different clothes. A person image and several clothing templates are sent to the augmentation module to get multiple synthetic person images, which can be seemed as images from the same person and provide partially supervised information for further usage.

2) To constrain the synthetic features of the same person, we adopt a self identity constrained loss to reduce the distance among these features, enhancing the overall performance of our model, especially for the cluster contrast and update procedure.

3) Both the average and batch-hard samplings are used on the features of query images (including the synthetic ones), which is more resource-saving and beneficial than using all image features. With the average sampling, the model can be trained by all the features but occupy less memory,
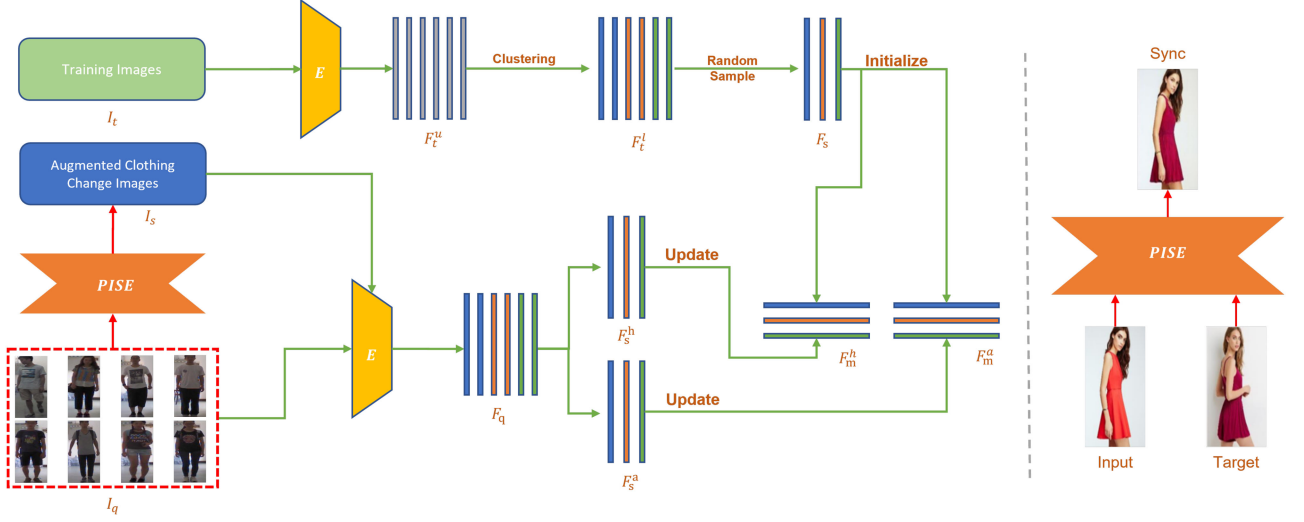
Fig. 1. The whole pipeline of our unsupervised clothing change ReID model (best viewed in color). The details of the augmentation module $PISE$ are showed separately on the right.

while with the batch-hard sampling, the model can be trained faster and more accurate.

II. PROPOSED METHOD

A. Overview of Pipeline

Inspired by [11], we propose a novel unsupervised pipeline for clothing change ReID as showed in Fig. 1.

Firstly, a modified ResNet-50 [12] is adopted as the backbone module, denoted by E, to extract the ReID features, which is pretrained on the ImageNet [13]. All training images I_t are sent to E and the features extracted are denoted by F_t^u:

$$F_t^u = E(I_t). \qquad (1)$$

Secondly, a clustering algorithm DBScan [14] is applied on F_t^u to generate pseudo labels for each feature of training images. DBScan is a density-based clustering method and can detect clusters of any shape with the effective discovery of noise points. Therefore, it is more suitable in person ReID problem than other clustering algorithms such as K-means [15] or agglomerative clustering [16].

DBScan requires two hyper-parameters: the maximal distance ε and the cluster neighbor density threshold M. The former represents the radius of neighbors, while the latter donates the maximal number of samples in a neighborhood for the core instances.

When the clustering finishes, a cluster ID is assigned to each feature of training images as its pseudo label:

$$F_t^l = C_\varepsilon^M(F_t^u), \qquad (2)$$

where C denotes the DBScan clustering.

Finally, a cluster contrast and update procedure is conducted, including computing a contrastive loss L_q using both batch-hard and average sampling. Moreover, to minimize the disruption brought by clothing information of an image, we add a clothing change augmentation module as well as a self identity constrained loss L_s. Thus the total loss L of the proposed pipeline can be represented as follows:

$$L = L_q + \alpha \cdot L_s, \qquad (3)$$

where α is the weight factor for L_s.

Details of the above mentioned modules are demonstrated in the following content.

B. Clothing Change Person Image Augmentation

To get the features of the same person in different clothes, we introduce a pretrained Person Image Synthesis Editing (PISE) [17] module to generate synthetic person images I_s from query images I_q and clothing style images I_y. As shown in the right of Fig. 1, it takes two inputs including an original query image and a targeting clothing style image, and generates a synthetic image with the content from the query image but in the style from the targeting image.

$$I_s = P(I_q, I_y). \qquad (4)$$

We use several different style images (the number is denoted by S) for each query image to get multiple synthetic images which can be seemed as images from the same person as the query image.

C. Cluster Contrast and Update Procedure

At the beginning of each training epoch, we randomly sample one single image feature from each clustered features F_t^l using uniform sampling U:

$$F_s(i) = U(F_t^l(i)), \qquad (5)$$

where $F_s(i)$ is the feature sampled from the i_{th} cluster $F_t^l(i)$. These sampled features are then used as the initial memory features for both of the average memory feature list F_m^a and the batch-hard memory feature list F_m^h, which will be updated respectively. The number of query images I_q is set $N \times K$, in which N is the number of person identities and K is the number of instances for each person identity. For each person identity, we send its K images to the PISE module to generate $S * K$

corresponding synthetic images. All of the images from I_q and I_s are sent to the ReID backbone E to get the query features F_q.

$$F_q = E(I_q \cup I_s). \tag{6}$$

The ith feature $F_m^h(i)$ in the batch-hard memory feature list is updated by part of the feature from queries related to the ith person identity, which has the maximal KL-Divergence to the current $F_m^h(i)$. The update rules are as follows:

$$F_m^h(i) = m \cdot F_m^h(i) + (1 - m) \cdot \underset{f_q \in F_q(i)}{\arg\max} D(f_q, F_m^h(i)), \tag{7}$$

where m denotes the momentum updating factor, $F_q(i)$ denotes all the query features marked with person identity i in the current mini batch and $D(\cdot, \cdot)$ is the KL-Divergence.

Similarly, for the ith feature $F_m^h(i)$ in the average memory feature list, it is updated by part of the average of features from queries related to the ith person identity:

$$F_m^a(i) = m \cdot F_m^a(i) + (1 - m) \cdot \frac{\sum_{f_q \in F_q(i)} f_q}{K \cdot (S + 1)}. \tag{8}$$

The query image features (including the original input and synthetic image features) are compared to all cluster features (including average memory and hardest memory) with InfoNCE loss [18]. The InfoNCE loss can be formulated as follows:

$$
\begin{aligned}
&L_q(F_q, F_m) \\
&= -\sum_i \sum_{\substack{f_q \in F_q(i) \\ f_m \in (F_m^h \cup F_m^a)}} \log \frac{\exp(f_q \cdot f_m(i)/\tau)}{\sum_{j=0}^K \exp(f_q \cdot f_m(j)/\tau)}, \tag{9}
\end{aligned}
$$

where f_m^i is the so-called "positive" memory feature to the query instance f_q because of their same identities and τ is a temperature hyper-parameter [19]. The InfoNCE loss tries to make f_q as similar as f_m^i, i.e., to make f_q similar to the features from its positive cluster and dissimilar to the features from other clusters.

D. Self Identity Constraint

Since the synthetic clothing change images can seem as images from the same person, we use a self identity constrained loss to reduce the variances among all image features related to the same identity. The loss is defined as

$$L_s = \frac{1}{N \cdot K \cdot (S + 1)} \sum_i \sum_{f_q \in F_q(i)} \sum_{a,b=0}^{S+1} D(f_q^a, f_q^b). \tag{10}$$

The self identity constrained loss tries to make the features of each original image and its corresponding synthetic images as similar as possible by reducing the sum of KL-Divergence between each pair of them.

III. EXPERIMENTS

A. Implementation Details

The whole model is based on the unsupervised learning baseline [11] and implemented in PyTorch [27]. The backbone is ResNet-50 [12] with the modification that the last layer stride is set 1, which makes the final output features have more dimensions and carry more information.

Each input image is resized to 256×128. Only the random cropping and horizontal flipping are performed for data augmentation. The random erasing and image padding are not used because the content of input images is required to keep unchanged for generating synthetic person images with clothing changes.

As mentioned previously, $N \times K$ is the number of query images, which is also the batch size. In our implementation, the number of images from each person identity K is set 4 and the number of pseudo identities (clusters) N is set 8. The number of synthetic images S for each input image is set 4. So each mini-batch has 160 images. The weight factor α in the total loss is set 0.3 and the momentum m for updating memory features is set 0.3 as well.

We use Adam optimizer and set both the weight decay and bias factors 0.0005. The base learning rate is set 0.00035. A linear warming up strategy is adopted to dynamically adjust the learning rate at the training stage. As for the hyper-parameters of DBScan, the maximal distance ε is set 0.4 and the cluster neighbor density threshold M is set 4.

When it comes down to testing, a 2048 d feature is extracted for each image from the Global Average Pooling (GAP) layer in the backbone, which is used to calculate the ReID distance.

B. Datasets and Evaluation Protocols

We evaluated our method on two cloth changing ReID datasets: Person ReID under the moderate Clothing Change (PRCC) [28] and VC-Clothes [29].

PRCC was collected exactly for the task of clothing change person ReID. It consists of 221 identities captured by three camera views. VC-Clothes is a synthetic dataset rendered by the GTA5 game engine, containing 19060 images of 512 identities captured by four cameras.

We followed the evaluation protocols of the above two datasets respectively. Specifically, for the PRCC dataset, we used the single-shot matching, i.e., randomly choosing one image from each identity to form the gallery. For the VC-Clothes dataset, we used the multi-shot matching, i.e., randomly choosing multiple images from each identity to form the gallery. To compare with existing methods comprehensively, we conducted experiments under both settings of "clothing change" and "same clothing," which determines whether the test set contains clothing change samples or not.

For the evaluation metrics, we selected the mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC) Rank-K accuracy.

C. Comparison With the State-of-The-Art

Since there are no existing unsupervised methods for clothing change person ReID, we compare our method with the state-of-the-art supervised methods on this topic.

As shown in Table I, we can see that our method does not have the highest score over all methods. The reason is that, although we have used the augmented information brought by the synthetic generation and a well-performed unsupervised deep ReID pipeline, our model still lacks the supervised identity information for better convergence of the ReID backbone. The pseudo labels are unreliable and the number of person identities

TABLE I
COMPARISON ON PRCC AND VC-CLOTHES. THE R1 IS RANK-1 CMC ACCURACIES (%). THE MAP IS MEAN AVERAGE PRECISION SCORE (%)

Method	Venue	PRCC				VC-Clothes			
		Clothing Change		Same Clothing		Clothing Change		Same Clothing	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP
LOMO [20] + KISSME [21]	CVPR15	18.55	-	47.40	-	-	-	-	-
LOMO [20] + XQDA [20]	CVPR15	14.53	-	29.41	-	34.5	30.9	86.2	83.3
PCB [22]	ECCV18	41.8	38.7	86.88	-	62.0	62.2	94.7	94.3
RGA-SC [23]	CVPR20	42.3	-	98.4	-	71.1	67.4	95.4	94.8
MGN [24]	ACMMM18	33.8	35.9	99.5	98.4	-	-	-	-
FSAM [25]	CVPR21	54.5	-	98.8	-	78.6	78.9	94.7	94.8
SGPS [26]	SPL21	65.8	61.2	99.5	96.7	-	-	-	-
Syn-Person-Cluster		43.7	39.8	87.4	82.1	67.4	62.5	91.9	89.3

TABLE II
SELF COMPARISON ON PRCC DATASET

Components			PRCC	
T_a	T_c	T_s	R1	mAP
			29.6	27.4
✓			30.3	27.9
		✓	32.4	30.8
✓		✓	35.1	33.3
✓	✓		37.2	35.3
✓	✓	✓	43.7	39.8

TABLE III
ANALYSIS OF ε AND M VALUE ON PRCC DATASET

ε Value	PRCC		m Value	PRCC	
	R1	mAP		R1	mAP
0.2	38.4	33.1	2	39.1	34.6
0.3	41.2	36.5	3	43.4	39.2
0.4	43.7	39.8	4	43.7	39.8
0.5	40.8	35.9	5	42.8	38.5
0.6	40.2	35.8	-	-	-
0.7	40.4	34.1	-	-	-

keeps changing during training, making it impossible to use the classification loss. Besides, the synthetic images for persons in different clothes are generated from a pretrained PISE module, whose image information comes from other texture style transfer datasets; thus, the quality of this kind of supervised information may not be satisfactory.

However, our method can achieve as good performance as some supervised person ReID methods. As shown in Table I, our method has higher R1 accuracy and mAP than PCB [22], RGA-SC [23] and LOMO [20]. This group of experiments demonstrates that our method is effective for unsupervised clothing change ReID.

D. Ablation Study

1) Component Effectiveness: In this section, we study the effectiveness of key components of our proposed method. We used the PRCC dataset under the "clothing change" setting in this part. The baseline is a purely unsupervised person ReID pipeline. The components include the clothing change person image augmentation module T_a, the self identity constrained loss T_c and the average and batch-hard sampling operations T_s. The results of different components' effectiveness are shown in Table II.

As shown in Table II, each proposed component can improve the performance of the baseline model, which obtains only 29.6% R1 accuracy and 27.4% mAP. Adding T_a can provide much more information from the augmented clothing change person images. However, the improvement of the model's performance is not significant (0.7% R1 accuracy and 0.5% mAp). The reason for such low improvement is that the sole use of T_a has no constraints on the person identities and the extra image information makes the model hard to converge. Solely using T_s can achieve an improvement of 2.8% R1 accuracy and 3.4% mAP. The average and batch-hard samplings can help the model with better and faster convergence. Combining T_a and T_s together gives 35.1% R1 accuracy and 33.3% mAP. T_c constrains the identities of both original and synthetic person images. It can

only be used when T_a exists and their combination achieves an improvement of 7.6% R1 accuracy and 7.9% mAP compared to the baseline. This significant improvement shows that the cooperation of the augmentation and constraints is able to disentangle people's clothing and identity information, helping our model to adapt the scenario of clothing change person ReID. Finally, our method with all components can reach the highest R1 accuracy (43.7%) and mAP (39.8%). This demonstrates that the three components T_a, T_c and T_s can work well collaboratively.

2) DBScan Hyper-Parameters: In the DBScan clustering algorithm, the values of ε and M affect the clustering performance and decide the final number of clusters (the number of person pseudo identities in our case). When ε is too small, the cluster with a small density will be divided into multiple clusters with similar properties. However, when ε is too large, clusters with a closer distance and a larger density will be merged into one cluster. The value of M affects the number of clusters in a similar way.

We analyzed the influence of ε and M respectively on the PRCC dataset under the "clothing change" setting using our complete method. The results are showed in Table III. From the table, we can see that our method works best when ε equals 0.4 and M equals 4.

IV. CONCLUSION

In this letter, we propose an unsupervised clothing change person ReID method called Syn-Person-Cluster. Its highlights are three-fold: a clothing change augmentation module, a self identity constrained loss and the average and batch-hard samplings. Extensive experiments have demonstrated that the proposed unsupervised method is capable of solving both problems of clothing changes and lack of data labels in person ReID simultaneously. To the best of our knowledge, we are the first to adopt unsupervised learning in the clothing change ReID problem and the method's performance is comparable to supervised person ReID methods.

REFERENCES

[1] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, 2021, doi: 10.1109/TPAMI.2021.3054775.

[2] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 2109–2118.

[3] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 8933–8940.

[4] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "VRSTC: Occlusion-free video person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 7183–7192.

[5] H. Liu, Y. Chai, X. Tan, D. Li, and X. Zhou, "Strong but simple baseline with dual-granularity triplet loss for visible-thermal person re-identification," *IEEE Signal Process. Lett.*, vol. 28, pp. 653–657, 2021, doi: 10.1109/LSP.2021.3065903.

[6] Z. Zhang, R. Y. Da Xu, S. Jiang, Y. Li, C. Huang, and C. Deng, "Illumination adaptive person reid based on teacher-student model and adversarial training," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 2321–2325.

[7] Y.-J. Li, Y.-C. Chen, Y.-Y. Lin, X. Du, and Y.-C. F. Wang, "Recover and identify: A generative dual model for cross-resolution person re-identification," in *Proc. Int. Conf. Comput. Vision Workshop*, 2019, pp. 8090–8099.

[8] S. Chen, Z. Fan, and J. Yin, "Pseudo label based on multiple clustering for unsupervised cross-domain person re-identification," *IEEE Signal Process. Lett.*, vol. 27, pp. 1460–1464, 2020, doi: 10.1109/LSP.2020.3016528.

[9] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised person re-identification by soft multilabel learning," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 2148–2157.

[10] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 2275–2284.

[11] Z. Dai, G. Wang, W. Yuan, S. Zhu, and P. Tan, "Cluster contrast for unsupervised person re-identification," 2021, *arXiv:2103.11568*.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996, vol. 96, no. 34, pp. 226–231.

[15] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probability*, Oakland, CA, USA, 1967, vol. 1, no. 14, pp. 281–297.

[16] W. H. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *J. Classification*, vol. 1, no. 1, pp. 7–24, 1984.

[17] J. Zhang, K. Li, Y.-K. Lai, and J. Yang, "PISE: Person image synthesis and editing with decoupled gan," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2021, pp. 7982–7990.

[18] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[19] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 3733–3742.

[20] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 2197–2206.

[21] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2012, pp. 2288–2295.

[22] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 480–496.

[23] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 3186–3195.

[24] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 274–282.

[25] P. Hong, T. Wu, A. Wu, X. Han, and W.-S. Zheng, "Fine-grained shape-appearance mutual learning for cloth-changing person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2021, pp. 10513–10522.

[26] X. Shu, G. Li, X. Wang, W. Ruan, and Q. Tian, "Semantic-guided pixel sampling for cloth-changing person re-identification," *IEEE Signal Process. Lett.*, vol. 28, pp. 1365–1369, 2021.

[27] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.

[28] Q. Yang, A. Wu, and W.-S. Zheng, "Person re-identification by contour sketch under moderate clothing change," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2029–2046, Jun. 2021, doi: 10.1109/TPAMI.2019.2960509.

[29] F. Wan, Y. Wu, X. Qian, and Y. Fu, "When person re-identification meets changing clothes," *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. Workshops*, pp. 830–831, 2020.