Fine-Grained Shape-Appearance Mutual Learning for Cloth-Changing Person Re-Identification

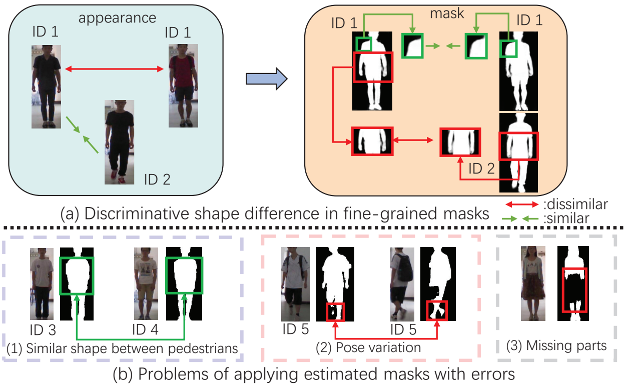
Peixian Hong[1,5#], Tao Wu[1,5#], Ancong Wu[1*], Xintong Han[4], Wei-Shi Zheng[1,2,3]
[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4]Huya Inc, China
[5]Pazhou Lab, Guangzhou, China
{hongpx,wutao63}@mail2.sysu.edu.cn,wuanc@mail.sysu.edu.cn,hanxintong@huya.com,wszheng@ieee.org

Abstract

Recently, person re-identification (Re-ID) has achieved great progress. However, current methods largely depend on color appearance, which is not reliable when a person changes the clothes. Cloth-changing Re-ID is challenging since pedestrian images with clothes change exhibit large intra-class variation and small inter-class variation. Some significant features for identification are embedded in unobvious body shape differences across pedestrians. To explore such body shape cues for cloth-changing Re-ID, we propose a ***F**ine-grained **S**hape-**A**ppearance **M**utual learning framework (FSAM), a two-stream framework that learns fine-grained discriminative body shape knowledge in a shape stream and transfers it to an appearance stream to complement the cloth-unrelated knowledge in the appearance features. Specifically, in the shape stream, FSAM learns fine-grained discriminative mask with the guidance of identities and extracts fine-grained body shape features by a pose-specific multi-branch network. To complement cloth-unrelated shape knowledge in the appearance stream, dense interactive mutual learning is performed across low-level and high-level features to transfer knowledge from shape stream to appearance stream, which enables the appearance stream to be deployed independently without extra computation for mask estimation. We evaluated our method on benchmark cloth-changing Re-ID datasets and achieved the start-of-the-art performance.*

1. Introduction

Person re-identification (Re-ID) aims at matching the same person across different cameras. Advanced methods have achieved high performance with deep learning [1, 34, 42, 45, 40, 53]. However, most current works largely rely on color appearance based on the assumption that the same person wears the same clothes in short term. Such limitation brings dramatic performance decrease in the situation that persons change their clothes, since different persons with similar clothes may be wrongly matched. To address this problem, we study the cloth-changing Re-ID problem [46, 29, 48, 13].



(a) Discriminative shape difference in fine-grained masks

(b) Problems of applying estimated masks with errors

Figure 1. Examples of images and masks in cloth-changing Re-ID. Color appearance under cloth-changing suffers from large intra-class variation and small inter-class variation, while body shape contains cloth-unrelated clues.

As the color appearance becomes unreliable in cloth-changing Re-ID, it is critical to learn cloth-unrelated features. Under moderate clothing change, the body shape does not change significantly for the same person and it is an important cue for identification. As shown in Figure 1 (a), the manually labeled fine-grained masks can capture detailed shape differences between pedestrians. In practical situations without manual label, human mask can be estimated by pretrained human parsing models. However,

# Equal contribution. Work done during the internship at Huya Inc.
* Corresponding author.

as shown in Figure 1 (b), error of estimated masks incurs difficulties for exploiting accurate body shape cues. First, the shapes of estimated coarse masks in corresponding body parts of different pedestrians may be highly similar and non-discriminative, since the human parsing models are not learned for identification (Figure 1 (b)(1)). Sometimes there are even missing parts in the mask because of domain gap caused by scene variations (Figure 1 (b)(3)). Second, the estimated mask of the same pedestrian suffers from large deformations when the poses changes, which causes large intra-class variation for utilizing body shape (Figure 1 (b)(2)).

To solve the above problems and mine discriminative body shape knowledge, we propose a **F**ine-grained **S**hape-**A**ppearance **M**utual learning framework (FSAM) that consists of a shape stream and an appearance stream as shown in Figure 2.

To extract discriminative fine-grained body shape features, in the shape stream, we learn fine-grained human masks under the guidance of identity while also preserving the prior knowledge of human parsing. To alleviate the impact of mask deformation brought by pose variation, we introduce a pose-specific multi-branch feature extractor to extract pose-specific fine-grained body shape features.

The cloth-unrelated appearance for cloth-changing Re-ID includes body shape, face, hair style, *etc*. Fine-grained body shape feature is important among the appearance cues but is hard to be mined from color image because of the dominant color-based appearance. In the appearance stream, to complement discriminative body shape knowledge for appearance feature learning, we propose dense interactive mutual learning to transfer the fine-grained body shape knowledge from shape stream to appearance stream in logit level and across different intermediate layer level. Meanwhile, mutual learning enables the appearance stream to be deployed independently for inference without extra computation of mask estimation and feature extraction in the shape stream.

Current works on cloth-changing Re-ID [46, 29] typically employ human poses or contour by off-the-shelf estimators, which can only capture limited discriminative shape knowledge and requires large estimation computation of poses or contours. Compared to them, our method instead learns the fine-grained masks with discriminative shape details and saves the mask estimation cost in inference.

In summary, our contributions are listed as follows:

(1) We learn fine-grained body shape features for cloth-changing Re-ID by estimating masks with discriminative shape details and extracting pose-specific features.

(2) A dense interactive mutual learning framework is proposed to transfer the fine-grained body shape knowledge to learn robust cloth-unrelated appearance features in an end-to-end fashion.

(3) Our **F**ine-grained **S**hape-**A**ppearance **M**utual learning framework (FSAM) achieves state-of-the-arts results on several benchmark cloth-changing Re-ID datasets including PRCC [46], LTCC [29] and VC-Clothes [37].

2. Related Work

Person Re-Identification. Person Re-ID has witnessed fast development in recent years. Early works mainly focus on feature extraction [5, 20, 7, 25] or distance metric learning [24, 38, 41, 28, 17]. With the development of deep learning, current works are mainly based on convolutional neural networks to learn discriminative features [1, 34, 42, 45, 40, 53, 23]. Human pose and parsing have been used to facilitate local feature learning, align the features in semantic level or eliminate the impact of background clutters [22, 26, 55, 31, 32, 33, 52, 9, 14, 58]. For example, SPReid [14] utilizes human parsing to capture local features. PGFA [26] exploits human pose to disentangle the useful information from occlusion noise. MGCAM [32] generates contrastive attention maps under the guidance of masks to learn body-aware and background-aware features. P^2-Net [9] proposes dual part-aligned representation to learn from both the human part masks and the non-human parts. However, these works mainly focus on short-term Re-ID based on color-appearance-based features, which are not robust under cloth-changing.

Cloth-changing Person Re-ID. Currently, there are only a few works on cloth-changing Re-ID [43, 37, 46, 29, 13, 48], which mainly aim to learn cloth-unrelated features from body shape or faces. Yang *et al*. [46] introduce a spatial polar transformation on contour sketch to learn shape features. Qian *et al*. [29] utilize human keypoints to eliminate the impact of appearance. Yu *et al*. [48] propose a mask attention to focus on face and body shape. Wan *et al*. [37] detect the faces and extract face features. However, in existing works, body shape features extracted from human poses or contour estimated by off-the-shelf estimators contain limited discriminative shape knowledge due to estimation error. Moreover, extra mask, pose, or contour estimation increases computation costs for these methods. In our proposed FSAM, we propose to learn human masks with more discriminative body shape details by identity guidance for extracting fine-grained shape features and save extra computation for mask or contour estimation by mutual learning.

Knowledge Distillation and Mutual Learning. Knowledge distillation (KD) [11, 59, 15, 36, 49, 47] is proposed initially for model compression, which enables knowledge transfer from teacher network to student network. Deep Mutual learning (DML) [51] proposes a two-way knowledge transfer between two networks, which allows networks to learn from each other. DML and KD have been applied to Re-ID in different scenarios that require knowledge trans-

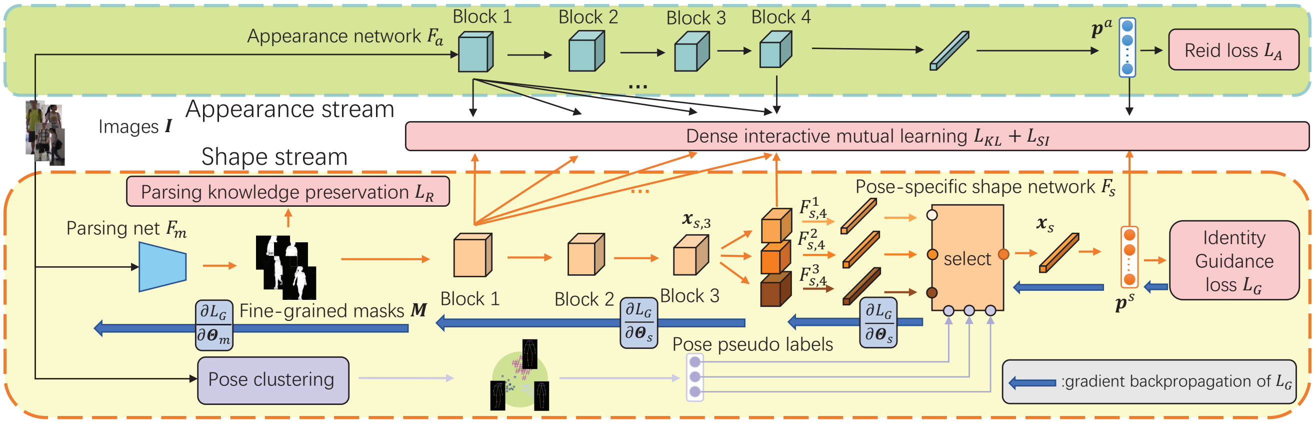

Appearance network F_a
Block 1
Block 2
Block 3
Block 4
...
p^a
Reid loss L_A
Images I
Appearance stream
Shape stream
Dense interactive mutual learning $L_{KL} + L_{SI}$
Parsing knowledge preservation L_R
Parsing net F_m
Fine-grained masks M
$\frac{\partial L_G}{\partial \Theta_m}$
Block 1
Block 2
Block 3
$\frac{\partial L_G}{\partial \Theta_s}$
$x_{s,3}$
$F^1_{s,4}$
$F^2_{s,4}$
$F^3_{s,4}$
Pose-specific shape network F_s
select
x_s
p^s
Identity Guidance loss L_G
$\frac{\partial L_G}{\partial \Theta_s}$
Pose pseudo labels
Pose clustering
:gradient backpropagation of L_G

Figure 2. Overview of our Fine-grained Shape-Appearance Mutual learning Framework. Our framework consists of two streams: an appearance stream and a shape stream. In the shape stream, the parsing net estimates fine-grained masks from input color images, and then the masks are fed into the shape feature extraction network to extract fine-grained body shape features. Then, dense interactive mutual learning transfers knowledge between appearance stream and shape stream to complement fine-grained body shape feature in appearance feature. During inference, only the appearance stream is required, which saves computation cost of mask estimation.

fer, including noisy labels refinement [6, 50, 54, 44], temporal knowledge transfer [8, 27] and scalable system learning [39]. Current works that use KD or DML for Re-ID ignore the knowledge embedded in intermediate layer. In contrast, our proposed dense interactive mutual learning allows dense knowledge interaction across features in different layers, which enables more effective knowledge transfer.

3. Fine-Grained Body Shape Features

In this work, we address the cloth-changing Re-ID problem, in which the same pedestrian wears different clothes. Knowledge embedded in body shape is essential as it is more robust against cloth changes than color appearance. Thus we aim to learn fine-grained masks and mine discriminative shape features by a parsing net and a shape feature extraction network, as shown in the lower part of Figure 2.

3.1. Identity-guided Fine-grained Mask Learning

Identity-guided Discriminative Mask Learning. To represent body shape of a pedestrian, we utilize the mask obtained by human parsing. Given an image $\mathbf{I}$, we feed it into a human parsing network F_m and obtain the human mask $\mathbf{M} = F_m(\mathbf{I}; \mathbf{\Theta}_m)$, where $\mathbf{\Theta}_m$ denotes the parameters of the network. We initialize F_m by training SCHP [18] on PASCAL-Person-Part [2]. As discussed in Section 1 and shown in Figure 1(b), such coarse estimated human masks are not accurate enough to tell the differences between cloth-changing pedestrians.

To solve these problems, we propose to use identity to guide the learning of human masks from coarse-grained level to fine-grained level. Specifically, the masks generated from human parsing network are fed into a shape feature extraction network F_s parameterized by $\mathbf{\Theta}_s$ to extract the shape features $\mathbf{x}_s = F_s(\mathbf{M}; \mathbf{\Theta}_s)$. Then, a fully-connected layer is applied for the features $\mathbf{x}_s$ for identity classification.

To guide the learning of both the parsing network F_m and the shape feature extraction network F_s for joint fine-grained mask estimation and body shape feature extraction, we introduce an identity guidance loss as follow:

$$L_G(\mathbf{\Theta}_m, \mathbf{\Theta}_s) = L^s_C + L^s_T, \tag{1}$$

where L^s_C denotes the cross entropy loss and L^s_T denotes the triplet loss, which are commonly used for discriminative feature learning in Re-ID [58, 53, 45, 1]. The cross entropy loss is formulated as

$$L^s_C(\mathbf{\Theta}_m, \mathbf{\Theta}_s) = -\sum_{i=1}^{N} y_i \mathrm{log} p^s_i, \tag{2}$$

where p^s_i is the probability of the i-th class for feature $\mathbf{x}_s$ and y_i is the one-hot identity label. N is the class number. Triplet loss can be written as

$$L^s_T(\mathbf{\Theta}_m, \mathbf{\Theta}_s) = [d_{ap} - d_{an} + m]_+ \,, \tag{3}$$

where $[\cdot]_+ = \max(\cdot, 0)$ and m denotes the margin. d_{ap} denotes the Euclidean distance between the anchor and positive sample in a triplet, and d_{an} denotes the distance between the anchor and negative sample.

Parsing Knowledge Preservation. With the identity guidance, the estimated coarse masks become more discriminative. However, pretrained parsing model contains prior parsing knowledge, which is of benefit to learn shape features but can be lost under identity guidance. As shown in Figure 3 (c), in the masks learned only by identity guidance loss L_G, there are some missing parts and shape-unrelated

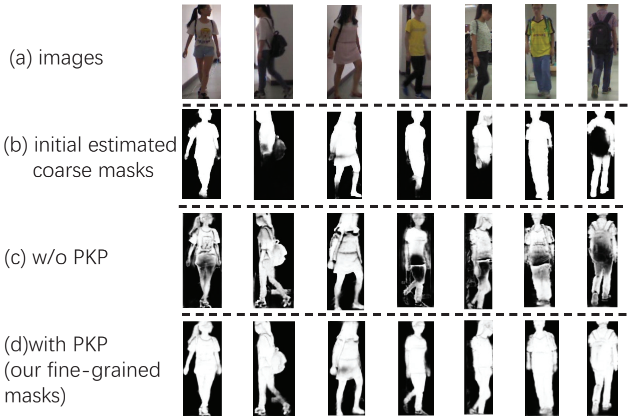

(a) images
(b) initial estimated coarse masks
(c) w/o PKP
(d)with PKP (our fine-grained masks)

Figure 3. Visualization of coarse masks generated by off-the-shelf human parsing model and fine-grained masks estimated by our parsing net F_m learned by identity guidance. "PKP" denotes parsing knowledge preservation.

textures (*e.g.* texture of clothes). To solve the problem, we preserve the parsing knowledge in the pretrained parsing model. To achieve this, we force the mask estimated by F_m to be close to the initially estimated coarse mask with a parsing knowledge preservation loss L_R:

$$L_R(\mathbf{\Theta}_m) = \frac{1}{P}\|\mathbf{M} - \hat{\mathbf{M}}\|_F^2, \tag{4}$$

where $\mathbf{M}$ and $\hat{\mathbf{M}}$ denote fine-grained mask and the initial coarse mask, respectively. $\|\cdot\|_F^2$ denotes the Frobenius norm, and P is the number of pixels in the mask image.

Objective Function. The objective function L_M for joint fine-grained mask estimation and feature extraction is formulated as:

$$L_M(\mathbf{\Theta}_m, \mathbf{\Theta}_s) = L_G + \lambda_R L_R, \tag{5}$$

where λ_R denotes the weight of mask parsing knowledge preservation loss L_R for regularization. The parsing net F_p and the shape feature extractor F_m are end-to-end trained by minimizing L_M, so as to jointly estimate fine-grained masks and extract fine-grained body shape features.

Visualization of Masks. To visually understand the effect of identity-guided fine-grained mask learning, we show some examples of estimated masks in Figure 3.

The guidance of identity can benefit the human masks in the following ways. First, human masks are encouraged to improve its ability to distinguish different persons. Some ID-related details of shape can be mined to refine the coarse masks to become fine-grained, as we compare the masks estimated by pretrained parsing model (Figure 3(b)) with those estimated by our model (Figure 3(d)). Second, the effect of domain gap between Re-ID data and training data for parsing is reduced under identity guidance, and thus some missing body part and prediction errors can be corrected.

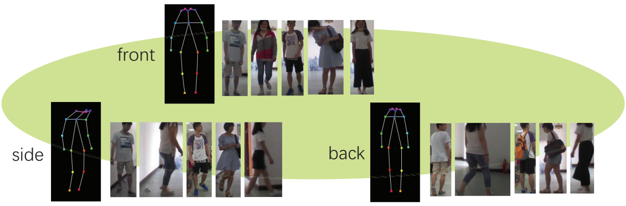


Figure 4. Poses are clustered into three groups. We show the clustering centers on PRCC [46]. We find the clustering centers are representative for poses of different views.

To show the effect of parsing knowledge preservation, we compare the cases without and with parsing knowledge preservation in Figure 3 (c) and (d). When only identity guidance is used for mask learning, there are missing parts and shape-unrelated textures in (c). With parsing knowledge preservation, these flaws are alleviated to obtain high-quality fine-grained masks in (d).

As a result, with fine-grained masks for more effective body shape representation, the shape feature extraction network F_s can learn more discriminative features.

3.2. Pose-Specific Fine-Grained Feature Extractor

As shown in Figure 1, estimated masks suffer from the problem of pose variation that the same pedestrian may have highly dissimilar shapes of the same body part, which leads to large intra-class variation for identification. In other words, the discriminative patterns of masks are not shared across different poses. To tackle this issue, we develop a pose-specific multi-branch network to extract fine-grained shape features under pose variations.

We first use AlphaPose [4] to extract human keypoints. Then, K-means clustering algorithm [21] is applied on the keypoint coordinates to cluster the images into three groups, as shown in Figure 4. Each mask is assigned with the corresponding pose clustering label. As shown in Figure 2, the 4^{th} convolution block of shape feature extraction network is divided into three branches with specific parameters, which are denoted as $\{F^i_{s,4}(\mathbf{x}_{s,3})\}^b_{i=1}$. b denotes the branch number and $\mathbf{x}_{s,3}$ denotes the features of the 3^{rd} convolution block. During training, the shape feature extraction network selects the corresponding branch according to the pose pseudo labels of the masks, while the other two branches are neglected. The mask feature $\mathbf{x}_s$ extracted by the pose-specific multi-branch network network is

$$\mathbf{x}_s = \sum_{i=1}^{b} \mathbb{1}_i F^i_{s,4}(\mathbf{x}_{s,3}), \tag{6}$$

where the value of $\mathbb{1}_i$ is one when the pose pseudo label of input mask is i, and it is zero otherwise. With specific feature extractor for specific pose, the body shape features are more fine-grained and robust to pose change.

4. Fine-Grained Shape-Appearance Mutual Learning Framework

The fine-grained body shape feature extracted in Section 3 is a significant appearance feature for cloth-changing Re-ID. Besides body shape, the pedestrian appearance also contains other cloth-unrelated information that is robust against cloth changes, such as face and hair styles, which are complementary to body shape. Mask image is suitable for body shape feature extraction, but it does not contain rich enough information of face and hair styles. In color image, the face and hair styles are more easily to be extracted, but the dominant color-based appearance makes it hard to focus on learning body shape. Therefore, we expect to extract comprehensive cloth-unrelated appearance features from both mask images and color images. We achieve this goal by introducing a two-stream framework that consists of a shape stream as introduced in Section 3 and an appearance stream.

4.1. Appearance Feature Learning

Similar to the feature extraction in the shape stream, a deep appearance network, F_a parameterized by $\mathbf{\Theta}_a$, takes color image as input to extract the appearance feature, as shown in the upper part of Figure 2. For training, cross entropy loss and triplet loss are applied to form the appearance feature learning loss L_A as

$$L_A(\mathbf{\Theta}_a) = L_C^a + L_T^a, \tag{7}$$

where L_C^a is the cross entropy loss and L_T^a is the triplet loss. Note that we do not apply pose-specific multi-branch structure to appearance stream, in order to save the extra cost for keypoints estimation during inference.

4.2. Dense Interactive Mutual Learning

To learn comprehensive cloth-unrelated appearance features from color images, we take advantage of the cloth-unrelated knowledge in both the shape stream and the appearance stream. To fuse the complementary appearance knowledge, we transfer the shape knowledge learned by the shape stream to complement the appearance stream.

To this end, we propose dense interactive mutual learning to transfer knowledge between the two streams in both *intermediate layer level* and *logit level*, which allows them to be trained collaboratively by mutual teaching.

Intermediate Layer Level Interaction. Intermediate layers learn multi-level knowledge for identification. On the one hand, low-level feature maps contain texture information such as corners or edges, which can enrich the semantic knowledge at high-level intermediate layers. On the other hand, semantic knowledge in high-level feature maps can guide the low-level intermediate layers to extract more discriminative texture features. Therefore, we propose a dense similarity loss for densely interactive knowledge transfer across low-level and high-level layers between appearance stream and shape stream, as shown in Figure 2.

As Re-ID is intrinsically a retrieval task, similarity between different persons represents the their relationship in the embedding space. So, we employ the feature similarity matrix of pedestrians in a batch to represent the knowledge of each layer. Take the feature maps of the d-th convolution block $\mathbf{x}_{s,d}$ of the shape stream as example, the similarity between the i-th and j-th sample is computed as

$$S_d^s(i,j) = \phi(\mathbf{x}_{s,d}^i)^T \phi(\mathbf{x}_{s,d}^j), \tag{8}$$

where ϕ denotes the operation of global average pooling and $\ell 2$ normalization. Then the feature map similarity matrix of the d-th convolution block within a batch can be denoted as $\mathbf{S}_d^s$ for shape stream and $\mathbf{S}_d^a$ for appearance stream. Based on this, we introduce our dense similarity loss to perform densely knowledge transfer between shape and appearance stream. Specifically, we minimize the distance between similarity matrices across different layers between two streams. As shown in Figure 2, the dense similarity loss can be formulated as

$$L_{SI}(\mathbf{\Theta}_a, \mathbf{\Theta}_m, \mathbf{\Theta}_s) = \sum_{i=1}^{l} \sum_{j=1}^{l} \|\mathbf{S}_i^a - \mathbf{S}_j^s\|_F, \tag{9}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm and $l = 4$ denotes the number of blocks used for mutual learning. By optimizing the dense similarity loss, low- and high-level features are encouraged to enhance each other across two streams, which enables more efficient knowledge fusion.

Logit Level Interaction. In logit level, each stream learns from both the ground truth labels and soft labels which are provided by the output class probabilities of the other stream. To utilize soft labels for knowledge transfer, we use Kullback Leibler (KL) Divergence as in deep mutual learning (DML) [51], which can be formulated as

$$D_{KL}(\mathbf{p}^a \| \mathbf{p}^s) = \sum_{i=1}^{N} p_i^a \log \frac{p_i^a}{p_i^s}, \tag{10}$$

where $\mathbf{p}^s$ denotes output class probabilities of shape stream and $\mathbf{p}^a$ denotes the output class probability of the appearance stream. As KL Divergence is asymmetric, we also compute $D_{KL}(\mathbf{p}^s \| \mathbf{p}^a)$ and the KL Divergence loss for appearance and shape stream is computed as

$$L_{KL}(\mathbf{\Theta}_a, \mathbf{\Theta}_m, \mathbf{\Theta}_s) = D_{KL}(\mathbf{p}^a \| \mathbf{p}^s) + D_{KL}(\mathbf{p}^s \| \mathbf{p}^a). \tag{11}$$

4.3. Overview of the Mutual Learning Framework

By mutual learning with dense similarity loss L_{SI} and KL divergence loss L_{KL}, the fine-grained body shape knowledge from the shape stream is complemented to

Table 1. Comparison on cloth-changing datasets. “Cloth-changing” and “Standard” denote two evaluation protocols illustrated in Section 5.1. “R-k” denotes rank-k accuracy (%). “mAP” denotes mean average precision (%). “-” denotes not reported.

Methods	LTCC				PRCC				VC-Clothes			
	Cloth-changing		Standard		Cloth-changing		Standard		Cloth-changing		Standard	
	R-1	mAP	R-1	mAP	R-1	R-10	R-1	R-10	R-1	mAP	R-1	mAP
LOMO [20] + KISSME [17]	10.75	5.25	26.57	9.11	18.55	49.81	47.40	81.42	-	-	-	-
LOMO [20] + XQDA [20]	10.95	6.29	25.35	9.54	14.53	43.63	29.41	67.24	34.5	30.9	86.2	83.3
PCB [34]	23.52	10.03	61.86	27.52	22.86	61.24	86.88	98.79	62.0	62.2	94.7	94.3
HACNN [19]	21.59	9.25	60.24	26.71	21.81	59.47	82.45	98.12	-	-	-	-
RGA-SC [53]	31.4	14.0	65.0	27.5	42.3	79.4	98.4	100	71.1	67.4	**95.4**	94.8
ISP [58]	27.8	11.9	66.3	29.6	36.6	66.5	92.8	98.4	72.0	72.1	94.5	94.7
Yang *et al*. [46]	-	-	-	-	34.38	77.3	64.2	92.62	-	-	-	-
Qian *et al*. [29]	25.15	12.4	71.39	34.41	-	-	-	-	-	-	-	-
baseline (ResNet50)	29.8	11.8	65.5	29.3	43.7	73.7	98.3	100	68.4	68.5	94.3	94.6
FSAM (ours)	**38.5**	**16.2**	**73.2**	**35.4**	**54.5**	**86.4**	**98.8**	**100**	**78.6**	**78.9**	94.7	**94.8**

the appearance stream to learn more comprehensive cloth-unrelated appearance features. Besides, appearance stream also provides complementary features for shape feature extraction in the shape stream.

Loss Function. We illustrate the loss function of our full framework. In shape stream, we generate fine-grained mask and learn pose-specific shape features by minimizing L_M. To utilize fine-grained body shape to assist appearance feature learning, we densely and mutually transfer knowledge between two stream with L_{SI} and L_{KL}. The two streams are trained in an end-to-end manner by optimizing

$$L = L_M + L_A + \lambda_{SI} L_{SI} + \lambda_{KL} L_{KL}, \tag{12}$$

where λ_{SI} and λ_{KL} denote the weights of dense similarity loss L_{SI} and KL Divergence loss L_{KL}, respectively.

Inference. For inference, we only use the appearance stream and the shape stream is discarded, so that the computation costs of estimation of masks and keypoints in the shape stream can be saved for fast inference.

5. Experiments

5.1. Datasets and Evaluation Protocols

Datasets. We mainly evaluated our method on three cloth-changing Re-ID datasets: PRCC [46], LTCC [29] and VC-Clothes [37]. PRCC is a large indoor cloth-changing Re-ID dataset, which contains totally 33,698 images of 221 identities captured from 3 cameras. LTCC is another indoor cloth-changing Re-ID dataset, which has 17,138 images of 152 identities with 478 different outfits captured from 12 camera views. LTCC is challenging as it contains various illumination, diverse human poses and occlusion. VC-Clothes is a synthetic dataset rendered by GTA5 game engine, which contains 19,060 images of 512 identities captured from 4 cameras. We also additionally evaluated our method on DukeMTMC [30] and Market-1501 [56], which are benchmark datasets for standard Re-ID without cloth changing.

Evaluation Protocols. For evaluation, we adopted the mean average precision (mAP) and rank-k accuracy.

For cloth-changing datasets PRCC [46], LTCC [29] and VC-Clothes [37], we followed their evaluation protocols and evaluated the performance in both cloth-changing setting and standard setting. For PRCC, we used single-shot matching by randomly choosing one image of each identity as gallery, which was repeated 10 times. The cloth-changing setting in PRCC means there are all cloth-changing samples in test set, while in the standard setting, there are all cloth-consistent samples in test set. As for LTCC, we used multi-shot matching by choosing all the images of each identity as gallery. The cloth-changing setting is the same as that of PRCC. Unlike PRCC, in standard setting, there are both cloth-consistent and cloth-changing samples in test set. For VC-Clothes, we also used multi-shot matching and the cloth-changing and standard setting are the same as that of PRCC. For standard Re-ID datasets Market-1501 [56] and DukeMTMC [30], we followed their standard evaluation protocols.

5.2. Implementation Details

We adopted ResNet50 [10] initialized by ImageNet [3] as backbone for both shape stream and appearance stream. The input images were resized to 256×128. For data augmentation, we adopted horizontal flipping and random erasing [57]. We used Adam optimizer [16] with the warm-up strategy that linearly increased the learning rate from 3×10^{-5} to 3×10^{-4} in the first 10 epochs. We then decreased the learning rate by a factor of 10 at epoch 40 and 70, and the training was stopped at epoch 150. Each batch contained 64 images of 16 identities. For PRCC [46], we set $\lambda_{KL} = 5$ and $\lambda_{SI} = 5$. For LTCC[29], we set $\lambda_{KL} = 1$ and $\lambda_{SI} = 0.5$. For VC-Clothes[37], we set $\lambda_{KL} = 1$ and $\lambda_{SI} = 1$. We set $m = 0.3$ for margin in the triplet loss and $\lambda_R = 10$ to control the regularization effect of parsing knowledge preservation. The values of λ_{KL}, λ_{SI} and λ_R were determined by cross validation.

5.3. Comparison with the State-of-the-Art Methods

We compared our method with the state-of-the-art methods on cloth-changing Re-ID datasets in Table 1. We can see that our method outperformed all compared methods by a large margin on cloth-changing datasets, with 13.2%/7.1% absolute improvement in rank-1 accuracy on

Table 2. Ablation study in cloth-changing setting."POSE" denotes pose-specific multi-branch structure. "3B" denotes a plain multi-branch structure without specific handing for poses. The 1^{st} row represents the result of baseline using only appearance stream. The last row represents the result of our final framework FSAM. For all experiments, we report the results of the appearance stream.

Methods	L_{KL}	L_{SI}	L_R	3B	POSE	LTCC		PRCC	
						R-1	mAP	R-1	R-5
1 (baseline)	-	-	-	-	-	29.8	11.8	43.7	63.3
2	✓	-	-	-	-	31.6	13.6	47.6	68.5
3	✓	✓	-	-	-	34.4	14.9	51.3	74.0
4	✓	✓	✓	-	-	35.5	15.8	53.1	73.9
5	✓	✓	✓	✓	-	33.9	14.6	52.3	75.2
6 (full)	✓	✓	✓	-	✓	**38.5**	**16.2**	**54.5**	**77.6**

PRCC [46]/LTCC [29] in cloth-changing setting, when compared with current state-of-the-art method RGA-SC [53]. As for VC-Clothes [37], we outperformed the state-of-the-art method ISP [58] by 6.6% in rank-1 accuracy.

Among the compared methods, Yang *et al.* [46] and Qian *et al.* [29] are also designed for cloth-changing Re-ID. Yang *et al.* [46] utilize contour sketch to capture shape features while Qian *et al.* [29] utilize human pose to distill the shape knowledge. In contrast to these works that model shape knowledge from relatively coarse estimations by pretrained contour or keypoints extractor, we learn fine-grained mask with identity guidance, which enables the model to learn detailed shape differences. Moreover, as our shape-appearance mutual learning complements the appearance features with body shape knowledge from fine-grained masks, we achieved better performance using only the appearance stream without extra computation for extracting poses or contour in inference as compared with them.

We also compared our method with current state-of-art standard Re-ID methods. Our method significantly outperformed them, because these methods assume that pedestrians wear the same clothes so that they do not consider learning fine-grained cloth-unrelated features.

5.4. Ablation Study

In this section, we study the effectiveness of key components of our proposed method. As shown in Table 2, our proposed FSAM significantly outperformed the baseline model by 10.8% and 8.7% on PRCC [46] and LTCC [29] respectively in cloth-changing setting on R1-accuracy.

Analysis of Shape Stream and Appearance Stream. We first evaluated the performance of shape stream and appearance stream. The details of notations can be referred in the caption of Table 3. We observe that although shape stream S achieves relatively low performance compared with appearance stream A, the concatenating of the features of the two streams $A + S$ still obtain improvement, which shows that shape features can complement to appearance features and therefore validates the potential of knowledge fusion.

Identity Guidance. We evaluated the effectiveness of identity guidance in the shape stream. As shown in Table 4, compared S^* to S, we observe that identity guidance brings significant improvement of performance of shape stream.

Table 3. Analysis of the performance of each stream with and without our dense interactive mutual learning. "A" denotes the appearance stream trained only by color images. "S" denotes the shape stream trained by the updating fine-grained masks under identity guidance. "$A+S$" denotes concatenating features of two streams. "**MU**" denotes the results with dense interactive mutual learning. "A (**MU**)" is the results of our final framework.

Methods	LTCC		PRCC	
	R-1	mAP	R-1	R-5
A	29.8	11.8	43.7	63.3
S	16.3	6.8	40.1	65.6
$A + S$	30.9	12.1	50.3	71.5
A (**MU**)	**38.5**	**16.2**	**54.5**	77.6
S (**MU**)	17.6	7.9	43.1	67.5
$A + S$ (**MU**)	35.7	15.2	52.3	**77.8**

Table 4. Effectiveness of the identity guidance. "S^*" denotes the shape stream trained only with the initial masks estimated by off-the-shelf parsing model. "$\leftrightarrow$" denotes two-way knowledge transfer. The other notations can be referred in Table 3. FSAM ($A \leftrightarrow S$) is our final framework.

Methods	LTCC		PRCC	
	R-1	mAP	R-1	R-5
S^*	9.7	3.9	31.2	56.8
S	16.3	6.8	40.1	65.6
FSAM ($A \leftrightarrow S^*$)	33.9	14.7	48.3	73.3
FSAM ($A \leftrightarrow S$)	**38.5**	**16.2**	**54.5**	**77.6**

We also evaluated its effectiveness within our full framework by changing the input of shape stream from the updating fine-grained masks to the initial estimated masks. The results of FSAM ($A \leftrightarrow S$) and FSAM ($A \leftrightarrow S^*$) in Table 4 indicate that, for our framework, the performance drops significantly by 6.2%/4.6% on PRCC [46]/LTCC [29] in rank-1 accuracy without identity guidance.

With identity guidance, we also successfully learned the masks from coarse-grained to fine-grained level, which can be seen by comparing Figure 3(b) and Figure 3(d). The performance improvement and visualization validated the effectiveness of identity guidance.

Dense Interactive Mutual Learning. Aiming for mutual knowledge transfer, the dense interactive mutual learning consists of KL divergence loss L_{KL} and dense similarity loss L_{SI}. As we can see in Table 2, rank-1 accuracy is improved with KL Divergence loss by 3.9% and 1.8% and can be further boosted with dense similarity loss by 3.7% and 2.8% on PRCC [46] and LTCC [29] respectively, verifying the effectiveness of our dense interactive mutual learning.

As shown in Table 3, we observe that performance of both two streams can be improved with the dense interactive mutual learning by comparing A/S with A (**MU**)/ S (**MU**), as they provide complementary features to each other.

Parsing Knowledge Preservation. In Table 2, the results show that parsing knowledge preservation (L_R) improves rank-1 accuracy by 1.8%/1.1% on PRCC [46]/LTCC [29], as it keeps the prior shape knowledge from human pars-

Table 5. Comparison between two-way and one-way knowledge transfer and analysis on changes of input modalities. "DML" denotes Deep Mutual Learning [51]. "↔" denotes two-way knowledge transfer while "←" denotes one-way knowledge transfer from shape stream to appearance stream. The other notations can be referred in the caption of Table 3. FSAM ($A \leftrightarrow S$) denotes our final framework.

Methods	LTCC		PRCC	
	R-1	mAP	R-1	R-5
DML [51] ($A \leftrightarrow A$)	30.9	12.6	46.5	66.8
DML [51] ($A \leftrightarrow S$)	31.6	13.6	47.6	68.5
FSAM ($A \leftrightarrow A$)	33.9	14.4	51.8	75.3
FSAM ($A \leftarrow S$)	34.4	14.8	52.0	72.3
FSAM ($A \leftrightarrow S$)	**38.5**	**16.2**	**54.5**	**77.6**

ing. The visualization also shows that with parsing knowledge preservation, parsing model avoids to generate shape-unrelated clothes texture and cause errors of part missing, as comparing (c) and (d) in Figure 3.

Pose-specific Feature Learning Structure. The results are shown in Table 2. The difference between "3B" and "POSE" is that there is only a simple multi-branch structure without specific handing of different poses in the "3B" setting, where all features in three branches are directly added for identification. It can be observed that, the pose-specific multi-branch structure brings improvement compared with both single-branch in row 4 and multi-branch structure in row 5 in Table 2, which shows that it can alleviate the effect of pose variation on mask shape deformation.

5.5. Further Analysis

Modalities for Mutual Learning. To validate the effectiveness of our shape stream, we evaluated different input modalities in our framework. Specially, we replaced the fine-grained masks with RGB color image, and performed dense interactive mutual learning between appearance and appearance, denoted as FSAM ($A \leftrightarrow A$) in Table 5. Comparing it with our full framework FSAM ($A \leftrightarrow S$), we observe that with the input of masks, we can achieve a much higher performance in cloth-changing setting, which is mainly because we can capture the shape knowledge much easier with mask input while it is hard to mine such knowledge implicitly with color image input. Comparison with other knowledge transfer method DML [51] also shows the effectiveness of the input masks as in Table 5.

Two-way *vs*. One-way Knowledge Transfer. In Table 5, we compared two-way knowledge transfer that allows mutual knowledge interaction between two streams with one-way knowledge transfer that only allows knowledge transfer from shape stream to appearance stream.

The results show that two-way knowledge transfer is better than one-way knowledge transfer by comparing FSAM ($A \leftrightarrow S$) and FSAM ($A \leftarrow S$). This is because mutual knowledge transfer can improve both streams by enhancing appearance stream to mine robust shape knowledge while also providing complementary features for shape stream.

Table 6. Comparison on standard datasets without cloth-changing.

Methods	DukeMTMC		Market1501	
	R-1	mAP	R-1	mAP
PCB [34]	83.3	69.2	93.8	81.6
IANet [12]	87.1	73.4	94.9	83.1
AANet [35]	87.7	74.3	93.9	83.4
DSA-reID [52]	86.2	74.3	95.7	87.6
RGA-SC [53]	-	-	96.1	88.4
ISP [58]	89.6	80.0	95.3	88.6
Baseline	85.5	75.3	94.1	84.9
FSAM (ours)	86.4	75.7	94.6	85.6

Results in Standard Re-ID Setting. To show the feasibility of our method for the cases without clothes change, we additionally evaluated our method on standard benchmark Re-ID datasets. As shown in Table 6, the performance of our method is comparable with the state-of-the-art methods on datasets DukeMTMC [30] and Market-1501 [56] without cloth-changing in short term. Specifically, we adopted the appearance stream as the baseline and find that our FSAM still achieves improvement, which shows that without clothes change our framework can still learn discriminative features from human body shape.

6. Conclusion

We study the challenging cloth-changing Re-ID problem. Body shape contains cloth-unrelated clues while human mask estimated by off-the-shelf human parsing model causes error, which makes it difficult to exploit accurate body shape. Therefore we propose a novel **F**ine-grained **S**hape-**A**ppearance **M**utual learning framework (FSAM), which consists of two streams: an appearance stream and a shape stream. In shape stream we learn the fine-grained masks and extract discriminative shape features under identity guidance with parsing knowledge preservation by the pose-specific multi-branch network. To complement body shape features in appearance features, we propose dense interactive mutual learning to transfer shape knowledge from shape stream to appearance stream, which allows appearance stream to be deployed independently in inference. The experiments show that our method achieves the state-of-the-art performance on cloth-changing Re-ID datasets.

Acknowledgement

This work was partially supported by the NSFC (U1911401,U1811461), China National Postdoctoral Program for Innovative Talents (BX20200395), Guangdong NSF Project (No. 2020B1515120085, 2018B030312002), Guangzhou Research Project (201902010037), Research Projects of Zhejiang Lab (No. 2019KD0AB03), and the Key-Area Research and Development Program of Guangzhou (202007030004).

References

[1] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *ICCV*, 2019. 1, 2, 3

[2] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 3

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[4] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 4

[5] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2

[6] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In *ICLR*, 2020. 3

[7] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008. 2

[8] Xinqian Gu, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Temporal knowledge propagation for image-to-video person re-identification. In *ICCV*, 2019. 3

[9] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. Beyond human parts: Dual part-aligned representations for person re-identification. In *ICCV*, 2019. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[12] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *CVPR*, 2019. 8

[13] Yan Huang, Jingsong Xu, Qiang Wu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. Beyond scalar neuron: Adopting vector-neuron capsules for long-term person re-identification. *IEEE TCSVT*, 2019. 1, 2

[14] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *CVPR*, 2018. 2

[15] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. In *NeurIPS*, 2018. 2

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 2, 6

[18] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *arXiv preprint arXiv:1910.09777*, 2019. 3

[19] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *CVPR*, 2018. 6

[20] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 2, 6

[21] Yoseph Linde, Andres Buzo, and Robert Gray. An algorithm for vector quantizer design. *IEEE Trans Commun*, 1980. 4

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

[23] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *CVPRW*, 2019. 2

[24] Lianyang Ma, Xiaokang Yang, and Dacheng Tao. Person re-identification over camera networks using multi-task distance metric learning. *IEEE TIP*, 2014. 2

[25] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016. 2

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

[27] Angelo Porrello, Luca Bergamini, and Simone Calderara. Robust re-identification by multiple views knowledge distillation. In *ECCV*, 2020. 3

[28] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary. Person re-identification by support vector ranking. In *BMVC*, 2010. 2

[29] Xuelin Qian, Wenxuan Wang, Li Zhang, Fangrui Zhu, Yanwei Fu, Tao Xiang, Yu-Gang Jiang, and Xiangyang Xue. Long-term cloth-changing person re-identification. *arXiv preprint arXiv:2005.12633*, 2020. 1, 2, 6, 7

[30] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016. 6, 8

[31] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *CVPR*, 2018. 2

[32] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 2018. 2

[33] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *ECCV*, 2018. 2

[34] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 2018. 1, 2, 6, 8

[35] Chiat-Pin Tay, Sharmili Roy, and Kim-Hui Yap. Aanet: Attribute attention network for person re-identifications. In *CVPR*, 2019. 8

[36] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*, 2020. 2

[37] Fangbin Wan, Yang Wu, Xuelin Qian, Yixiong Chen, and Yanwei Fu. When person re-identification meets changing clothes. In *CVPRW*, 2020. 2, 6, 7

[38] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *NeurIP*, 2009. 2

[39] Ancong Wu, Wei-Shi Zheng, Xiaowei Guo, and Jian-Huang Lai. Distilled person re-identification: Towards a more scalable system. In *CVPR*, 2019. 3

[40] Tong Xiao, Hongsheng Li, Wanli Ouyang, and Xiaogang Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, 2016. 1, 2

[41] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, 2014. 2

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

[43] Jia Xue, Zibo Meng, Karthik Katipally, Haibo Wang, and Kees van Zon. Clothing change aware person identification. In *CVPRW*, 2018. 2

[44] Fengxiang Yang, Ke Li, Zhun Zhong, Zhiming Luo, Xing Sun, Hao Cheng, Xiaowei Guo, Feiyue Huang, Rongrong Ji, and Shaozi Li. Asymmetric co-teaching for unsupervised cross-domain person re-identification. In *AAAI*, 2020. 3

[45] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. Spatial-temporal graph convolutional network for video-based person re-identification. In *CVPR*, 2020. 1, 2, 3

[46] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE TPAMI*, 2019. 1, 2, 4, 6, 7

[47] Anbang Yao and Dawei Sun. Knowledge transfer via dense cross-layer mutual-distillation. In *ECCV*, 2020. 2

[48] Shijie Yu, Shihua Li, Dapeng Chen, Rui Zhao, Junjie Yan, and Yu Qiao. Cocas: A large-scale clothes changing person dataset for re-identification. In *CVPR*, 2020. 1, 2

[49] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 2

[50] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. Multiple expert brainstorming for domain adaptive person re-identification. *arXiv preprint arXiv:2007.01546*, 2020. 3

[51] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, 2018. 2, 5, 8

[52] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, 2019. 2, 8

[53] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, 2020. 1, 2, 3, 6, 7, 8

[54] Fang Zhao, Shengcai Liao, Guo-Sen Xie, Jian Zhao, Kaihao Zhang, and Ling Shao. Unsupervised domain adaptation with noise resistible mutual-training for person re-identification. In *ECCV*, 2020. 3

[55] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose-invariant embedding for deep person re-identification. *IEEE TIP*, 2019. 2

[56] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 6, 8

[57] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. abs/1708.04896, 2017. 6

[58] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. *ECCV*, 2020. 2, 3, 6, 7, 8

[59] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. In *NeurIPS*, 2018. 2