Supplementary Material for “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

This supplementary material accompanies our main manuscript “Fine-grained Shape-Appearance Mutual Learning for Cloth-Changing Person Re-Identification”. We provide further analysis and some visualization results on LTCC [5] and PRCC [6] in this supplementary material.

1. More Analysis and Experiments

Due to the space limitation in main manuscript, we report the results of experiments on the parameters and some components of our proposed **F**ine-grained **S**hape-**A**ppearance **M**utual Learning framework (FSAM) in this supplementary material.

1.1. Analysis on Dense Similarity Loss

In Section 4.2 in main manuscript, we propose the dense similarity loss L_{SI} in Eq. (9) to perform mutual knowledge transfer by encouraging dense knowledge interaction across low-level and high-level features in different layers. To show the effectiveness of this cross-layer knowledge interaction, we conducted experiments and reported the testing rank-k accuracy and mAP in Table S1.

DSIM denotes our final framework with dense similarity loss L_{SI}. The difference between SIM and DSIM is that in SIM, we only minimize the distance between corresponding similarity matrices of two streams without the cross-layer interaction. That is to say, for the similarity matrix of d^{th} convolutional block in one stream, we only minimize the distance between it and the similarity matrix of d^{th} convolutional block in the other stream, where $d \in \{1, 2, 3, 4\}$. SIM* means that we only minimize the distance of the similarity matrices of the last convolutional block without utilizing other intermediate layers.

Table S1. Analysis on our dense similarity loss L_{SI}.

Methods	LTCC		PRCC	
	R-1	mAP	R-1	R-5
SIM*	36.7	14.6	50.5	71.8
SIM	36.2	15.5	51.0	73.1
DSIM (ours)	**38.5**	**16.2**	**54.5**	**77.6**

Specifically, we replace DSIM with SIM and SIM* while maintaining the same parameters and other components in our framework for fair comparison. As shown in Table S1, DSIM outperforms SIM in cloth-changing setting, which validates the effectiveness of cross-layer interaction. We can also observe that SIM cannot achieve much improvement compared with SIM* while DSIM can, which shows that with cross-layer interaction, DSIM mines the knowledge embedded in intermediate layers more effectively.

1.2. The Number of Pose Clusters

In Section 3.2, we propose the pose-specific multi-branch feature learning structure in shape stream to learn pose-specific fine-grained shape features, where we cluster the images into several groups according to their key-point coordinates and divide the 4^{th} convolutional block into several branches accordingly to handle different poses as shown in Figure 2 in main manuscript. To show the effect of different number of pose clusters, we conducted the experiments and showed testing rank-1 accuracy in cloth-changing setting. As shown in Figure S1, we varied the cluster number from 2 to 5. It can be observed that we achieve the best performance when the pose cluster number is 3.

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

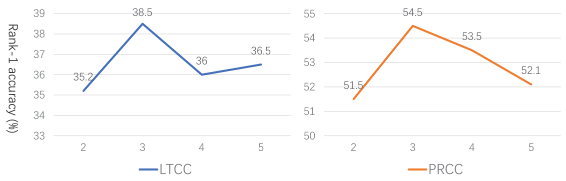
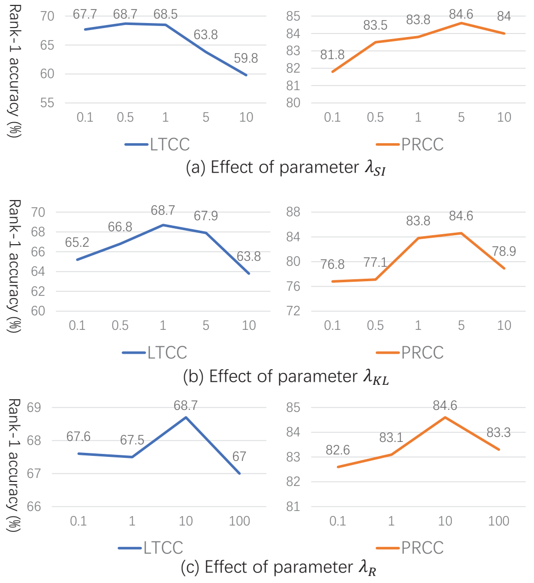


Figure S1. Analysis on the number of pose clusters.



(a) Effect of parameter λ_{SI}

(b) Effect of parameter λ_{KL}

(c) Effect of parameter λ_R

Figure S2. Effect of key parameters. We report the average rank-1 accuracy on validation set.

1.3. Effect of Key Parameters

We evaluated the effect of key parameters λ_{KL}, λ_{SI} and λ_R as shown in Figure S2. Specifically, we conducted cross validation in cloth-changing setting by randomly selecting one-tenth of the identities in the training set as validation set. When evaluating the effect of one of the parameters on validation set, we fixed the other parameters.

Effect of Parameter λ_{SI}. λ_{SI} controls the effect of dense similarity loss L_{SI} in Eq. (12), which aims to mutually transfer knowledge between appearance stream and shape stream in intermediate layer level as illustrated in Section 4.2. To show the effect of λ_{SI}, we varied it from 0.1 to 10 as shown in Figure S2 (a). When λ_{SI} is too small, we cannot effectively transfer the body shape knowledge from shape stream to appearance stream, so that we only gets a relatively low performance. When λ_{SI} gets too large, the mutual knowledge transfer can force the features of two streams to become similar in the very early training process, which also leads to the decrease of performance, as the two streams cannot provide complementary cloth-unrelated features to each other in the remaining training process.

Effect of Parameter λ_{KL}. λ_{KL} controls the effect of Kullback Leibler (KL) Divergence loss L_{KL} in Eq. (12), which aims to transfer knowledge in logits level as illustrated in Section 4.2. To show the effect of λ_{KL}, we also varied it from 0.1 to 10 as shown in Figure S2 (b). As L_{KL} is also proposed for knowledge transfer between the two streams, the impact of λ_{KL} on the performance is similar to that of λ_{SI}.

Effect of Parameter λ_R. λ_R is the weight of parsing knowledge preservation L_R in Eq. (5), which is proposed to preserve the prior parsing knowledge under identity guidance, as illustrated in Section 3.1. To show the effect of λ_R, we varied it from 0.1 to 100 as shown in Figure S2 (c). We observe that the decrease of λ_R causes the decrease of the performance, which is because the regularization of L_R is weakened and the prior parsing knowledge can be lost by identity guidance . When λ_R is too large, we also observe the decrease of performance, as larger λ_R can hinder the learning of id-related and fine-grained shape details on the masks.

Summary. By cross validation, we finally set $\lambda_{SI} = 5$ and $\lambda_{KL} = 5$ for PRCC [6]. For LTCC [5], we set $\lambda_{SI} = 0.5$ and $\lambda_{KL} = 1$. λ_R is set as 10 for both datasets.

1.4. Analysis on number of parameters and complexity

With the dense interactive mutual learning to transfer shape knowledge from shape stream to appearance stream, in testing process we only use the appearance stream, a simple ResNet50. The estimation of mask and keypoint involved in training of shape stream is not required in testing. We evaluate the number of parameters and running time of our model and other models in Table S2. Our method reduces a lot of parameters when we compare training with testing. Also in testing process our final framework achieves much better performance and does not increase the parameters, compared with baseline and other methods like RGA [7] and ISP [8]. In terms of number of parameters and computational cost at test time, our comparison with other methods is fair.

Table S2. Comparisons of network parameters (Params) and training and testing time. Baseline indicates only the appearance stream is used. Note that all experiments are conducted fairly on two Tesla V100 GPUs.

Methods	Training		Testing		PRCC	
	Params	Time	Params	Time	R-1	R-10
RGA [7]	30.13M	0.8h	30.13M	40s	42.3	79.4
ISP [8]	31.68M	16.5h	31.68M	30s	36.6	66.5
Baseline	23.82M	1.2h	23.82M	15s	43.7	73.7
FSAM (ours)	164.27M	12h	23.82M	15s	54.5	86.4

1.5. Analysis on inaccurate pose estimation

In our proposed framework, we use the poses to cluster the pedestrian images into several different views. As shown in Figure 4 in main manuscript, when we cluster poses into three groups, we find that the center poses of the three clusters (average pose within one cluster) can be representative as different views including front, back and side view respectively. We believe that slightly inaccurate pose estimation will not have a large impact on identifying which view the pedestrian belongs to. For example, a person in front view is very unlikely to be estimated to be with the pose of side view. To fully analyze the problem of inaccurate pose estimation, we replace AlphaPose [2] with other pose estimators which have lower precision, including Mask-RCNN [3] and OpenPose [1]. The results can be found in Table S3. We observe that lower precision of pose estimation has not much impact and can still achieve improvement compared with our method without pose clustering.

Table S3. Analysis on inaccuracy of pose estimation. COCO AP are average precision reported on COCO test-dev.

Pose estimators	COCO AP	LTCC		PRCC	
		R-1	mAP	R-1	R-5
w/o pose clustering	-	35.5	15.8	53.1	73.9
AlphaPose [4] (ours)	73.3	38.5	16.2	54.5	77.6
Mask-RCNN [R1]	67.0	37.2	16.2	54.7	76.5
OpenPose [R2]	61.8	37.5	15.8	54.8	77.3

2. Visualization

Matching Examples. To have better visual understanding, we show some matching examples of baseline (a) and our FSAM (b) in Figure S3 and Figure S4. Compared to baseline that tends to match pedestrians with clothes of similar colors, our proposed FSAM correctly matches the same pedestrian even when the clothes are changed, which validates that we successfully learn the cloth-unrelated features for retrieval.

Visualization of Fine-grained Mask. We provide more examples of our fine-grained masks in Figure S5 and S6. Comparing fine-grained masks (c) with initial estimated coarse masks (b), we observe that our fine-grained masks are able to capture id-related and discriminative shape details and alleviate the problem of part missing caused by domain gap.

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *TPAMI*, 2019. 3

[2] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 3

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 3

[4] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *arXiv preprint arXiv:1910.09777*, 2019. 5

[5] 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, 4, 5

[6] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE TPAMI*, 2019. 1, 2, 4, 5

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

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

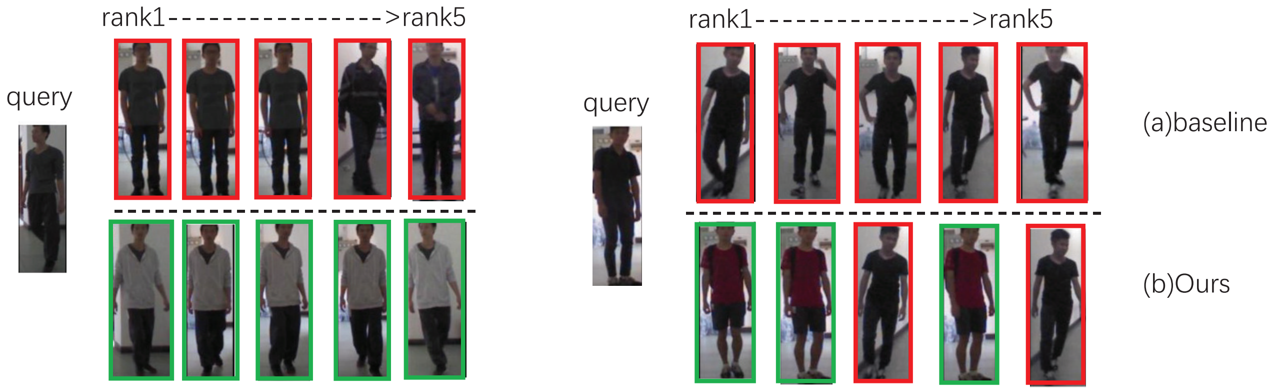


Figure S3. Some matching examples of baseline and our FSAM on PRCC [6] are shown. The correct matches are indicated by green bounding boxes while the incorrect matches are indicated by red bounding boxes.

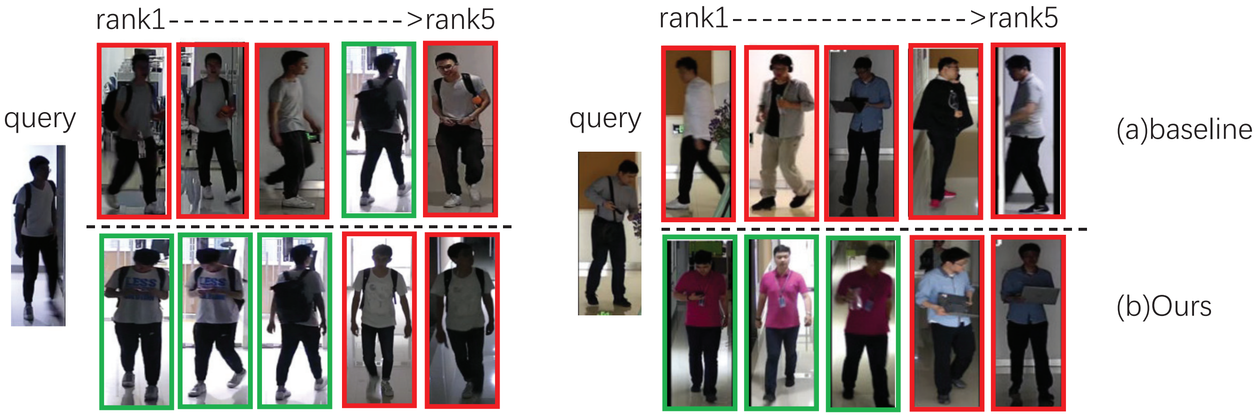


Figure S4. Some matching examples of baseline and our FSAM on LTCC [5] are shown.

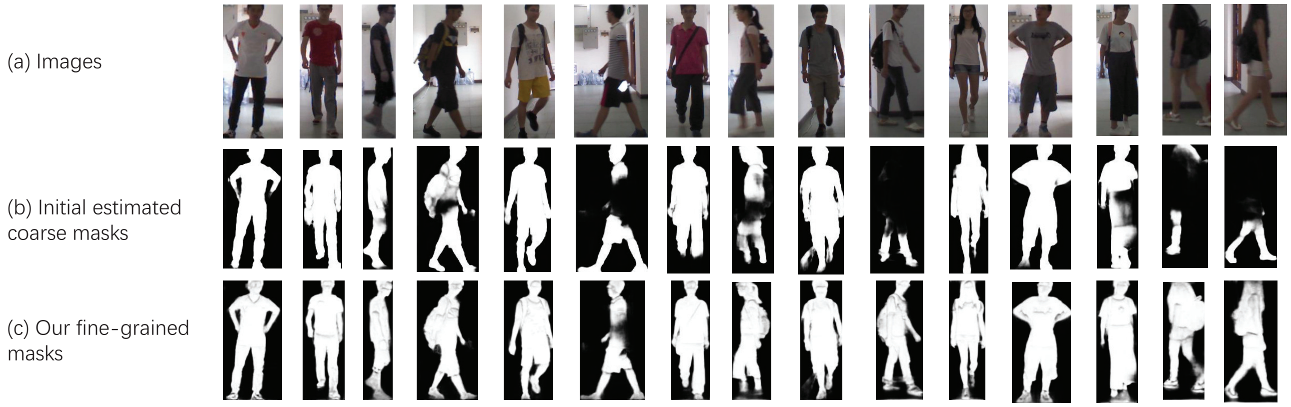

(a) Images
(b) Initial estimated coarse masks
(c) Our fine-grained masks

Figure S5. Visualization on PRCC [6]. We provide examples of our fine-grained masks and the initial coarse masks estimated by off-the-shelf human parsing model SCHP [4].

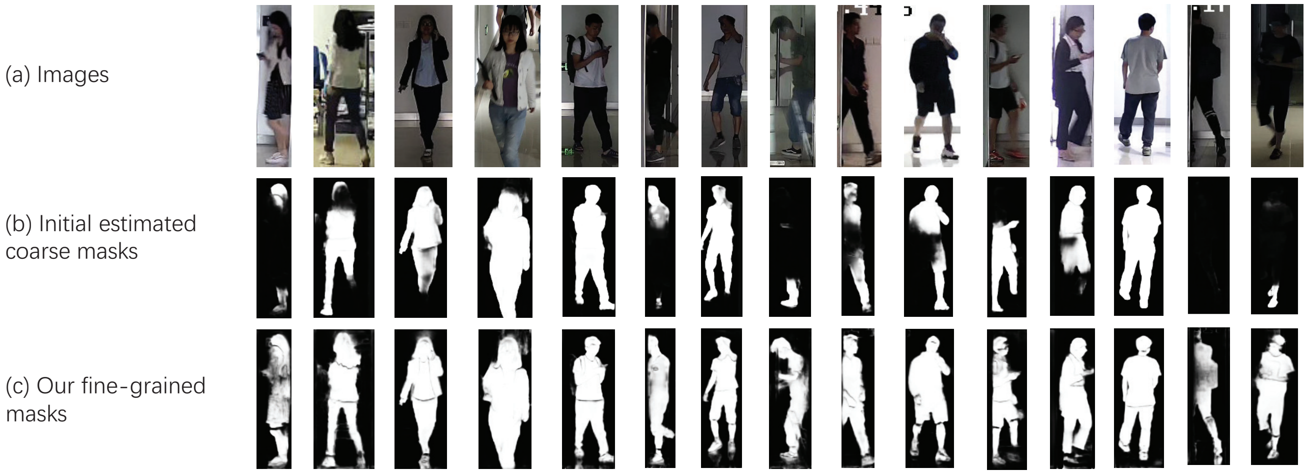

(a) Images
(b) Initial estimated coarse masks
(c) Our fine-grained masks

Figure S6. Visualization on LTCC [5]. We provide examples of our fine-grained masks and the initial coarse masks estimated by off-the-shelf human parsing model SCHP [4].