Supplementary Material: Cross-Modality Complementary Learning for Video-based Cloth-Changing Person Re-Identification

Illustration of E-VCCR dataset

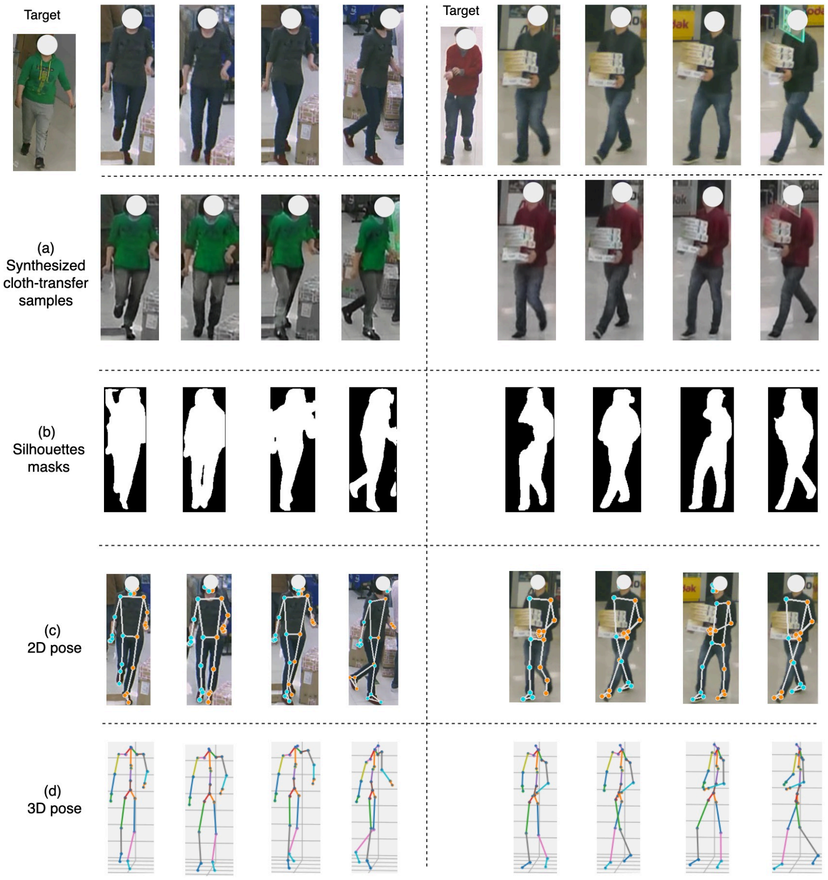
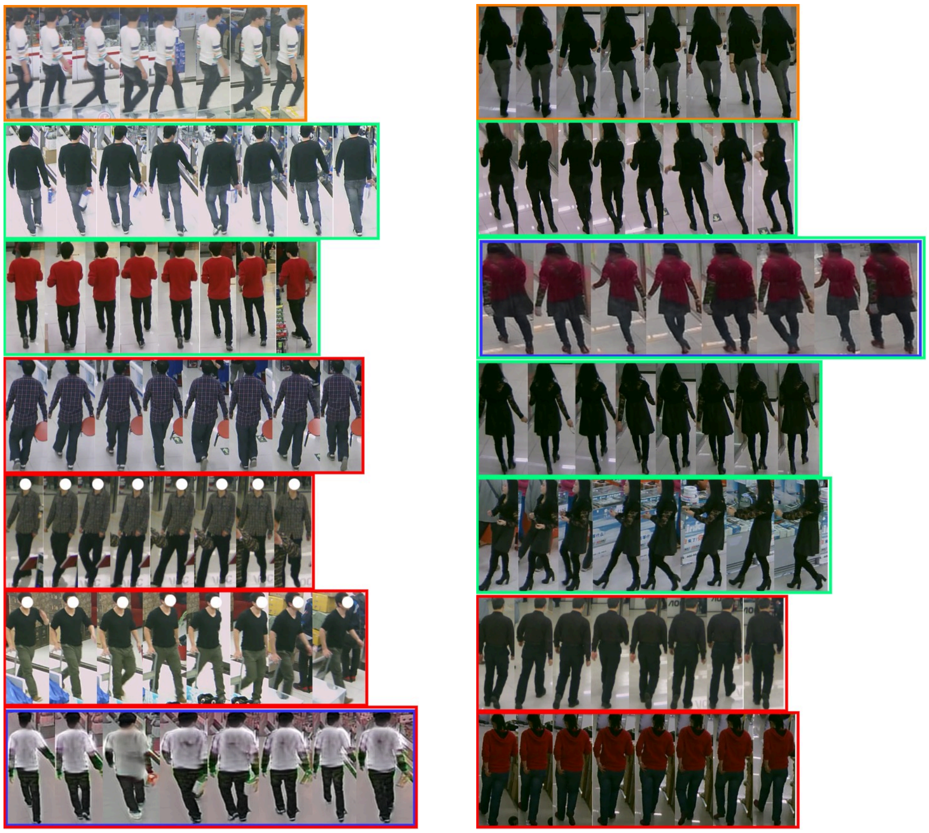


Fig. 5: Illustration of E-VCCR data. Top: original sequence and target clothing for transfer. (a) synthesized sequence. (b) silhouette masks. (c) 2D skeleton-based pose. (d) 3D skeleton-based pose.

An illustration of E-VCCR data is shown in Figure 5, in which the following sequences are visualized: original sequence and target clothing for transfer, synthesized sequence, silhouette masks, 2D skeleton-based pose, and 3D skeleton-based pose.



(a) Our model is able to correctly match gallery sequences at rank-1 and rank-2 despite moderate clothing and viewpoint changes. Meanwhile, under similar-clothing scenario caused by the generated distractor at rank-6, gait embedding learnt by our model shows effectiveness in distinguishing individuals.

(b) The generated distractor of same identity as query at rank-2 is correctly matched, while the sequence of different identity but similar clothing is ranked 6 based on similarity, showing the robustness of our proposed framework. Correct matches from rank-1 to rank-4 show high Re-ID accuracy produced by our model under real-world Re-ID environment.

Fig. 6: Visualization of ranking list on E-VCCR. The top row with orange bounding box is the query sequence. The following 6 sequences are gallery sequences ranked by pair-wise cosine similarities to the query sequence. Correctly matched sequences are covered by green bounding boxes, while sequences of different identities are covered by red bounding boxes. The generated cloth-transfer distractors are covered by blue bounding boxes.

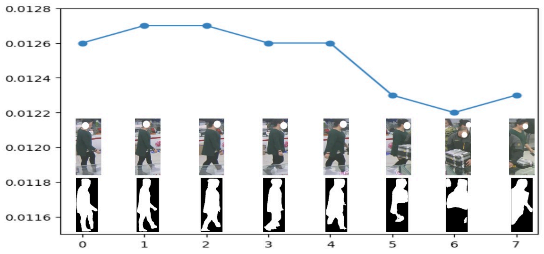


Fig. 7: Per-frame attention scores learnt by ATA module.

Method	VCCR		E-VCCR	
	R-1	mAP	R-1	mAP
Avg. Pooling	50.5	41.1	43.0	35.2
Set Pooling [6]	51.6	41.8	44.1	35.9
ATA	**54.2**	**44.2**	**46.3**	**38.6**

Table 7: Effectiveness of ATA module

Visualized Ranking List on E-VCCR

We provide a qualitative analysis of experimental results on E-VCCR dataset in Figure 6. The top row with orange bounding box is the query sequence. It is followed by the top-6 gallery sequences ranked by pair-wise cosine similarities. Green bounding boxes denote the correctly matched sequences, while red bounding boxes represent wrong results. Sequences covered by blue bounding boxes are generated cloth-transfer distractors. From Figure 6a, it can be seen that our model is robust to moderate clothing changes and viewpoint variations, shown by matched sequences at rank-1 and rank-2. Our model shows effectiveness of using gait and body shape to assist Re-ID learning in real-world scenarios, shown by rank-6 of the generated distractor sequence of similar clothing but different identity. Figure 6b further demonstrates the robustness of our framework, shown by the correct match at rank-2 of a generated distractor of same identity but similar clothing with sequence at rank-6.

Attention-based Temporal Aggregation (ATA).

From Table 7, on VCCR, ATA [49] makes an improvement of 3.7% and 2.6% in mAP compared to Avg. Pooling and Set Pooling [6], respectively. This is because ATA leverages GRU layers to implicitly capture temporal dependencies from sequences. Moreover, the employed attention mechanism enables ATA to assign higher attention scores to the most informative frames, as illustrated in an example sequence of frames in Figure 7. Features from these frames thus contribute more to the final person representation, while contribution of noisy frames caused by occlusion is minimized, leading to higher discriminative power for Re-ID.

Setting	CC		Standard	
	R-1	mAP	R-1	mAP
2 layers, size 512	52.8	42.9	88.0	65.2
2 layers, size 1024	**54.2**	**44.2**	**89.3**	**66.9**
3 layers, size 512	53.4	43.2	88.3	65.9
3 layers, size 1024	53.6	43.7	88.7	66.4

Table 8: Analysis on several self-attention configurations of ATA module. Experiments are on VCCR.

Method	VCCR		CCPG	
	R-1	mAP	R-1	mAP
WFP [21]	52.3	43.1	81.7	54.8
MSF [82]	51.1	42.0	79.9	53.1
Concat	**54.2**	**44.2**	**84.1**	**56.2**

Table 9: Analysis on appearance and gait fusion methods.

Self-attention configurations for ATA

The goal of using self-attention layers in our ATA module is to learn a coefficient for each frame which represents its contribution to in the final vector, producing a more fine-grained output. Table 8 reports experiments conducted on different configurations of self-attention in terms of number of layers and layer size. We found that using two self-attention layers of size 1024 yields the best results.

Feature Fusion

We apply concatenation to fuse texture and gait embeddings for final person representation, and compare its effectiveness with the Weight Prediction Fusion (WPF) [21] and Matching Score Fusion (MSF) [82] as shown in Table 9. WFS projects the embeddings onto a common feature space, then performs summation. MSF computes similarity score for each embedding, them sum up the scores for sequence-to-sequence matching. It can be seen that simple concatenation achieves the highest results since transforming the embeddings is not able to preserve their discriminative power for Re-ID.