

# Coarse Attribute Prediction with Task Agnostic Distillation for Real World Clothes Changing ReID (*Supplementary*)

Priyank Pathak  
 priyank@ucf.edu  
 Yogesh S Rawat  
 yogesh@ucf.edu

Center for Research in Computer Vision  
 University of Central Florida  
 Orlando, FL, USA

## Contents

<b>1</b>	<b>More on CAL Sensitivity problem</b>	<b>1</b>
<b>2</b>	<b>Base Model Details</b>	<b>2</b>
<b>3</b>	<b>Clothes Augmentation (part of Base Model)</b>	<b>3</b>
<b>4</b>	<b>Design Choices (More experiments)</b>	<b>4</b>
<b>5</b>	<b>Effect of Low Quality Images on Soft Biometrics</b>	<b>4</b>
<b>6</b>	<b>Analysis of high-quality and low-quality data (Failure of Traditional methods)</b>	<b>5</b>
<b>7</b>	<b>Visualization of Synthetic Low Quality Images (Used in TBD)</b>	<b>6</b>
<b>8</b>	<b>Additional Training Details</b>	<b>6</b>
<b>9</b>	<b>Dataset Samples (RGB Examples)</b>	<b>7</b>
<b>10</b>	<b>Visualization of Pose Clusters (RGB Examples)</b>	<b>7</b>
<b>11</b>	<b>Ethical Statement</b>	<b>8</b>

## 1 More on CAL Sensitivity problem

Table 1 demonstrates the use of Coarse Attribute Prediction (CAP) for pose and gender in our Base Model and stand-alone CAL [5]. Similar to Triplet loss and foreground augmentation, (shown in main submission), CAL is overly sensitive to foreign loss functions, leading to incompatibility. This strengthens our choice for a two-branch structure (Base Model)

where we can learn attributes like pose and gender via a second branch (the BOT Branch) and communicate the information by sharing the backbone.

Aux.	CAL				Base Model			
	LTCC		PRCC		LTCC		PRCC	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
Vanilla	40.1	18.0	55.2	55.8	41.1	20.3	58.0	59.4
+G	34.9	17.0	45.6	45.5	43.5	21.7	58.8	59.8
+P	34.2	15.9	43.5	41.5	42.5	20.2	59.4	60.6
+CAP	36.0	16.8	43.7	43.5	43.4	20.7	60.2	60.3

Table 1: Contribution of each architectural component on vanilla CAL vs two-branch Base Model (CAP is +Gender+Pose)

From the CAL’s original paper, their ablation study shows extensive incompatibility of CAL with Triplet loss, the most commonly used ReID loss function (table 2).

## 2 Base Model Details

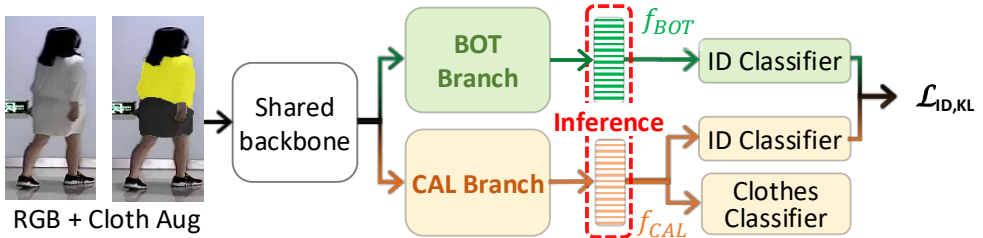


Figure 1: **Base Model:** performance on ‘male’ gender classification for all the datasets. The sharp drop in accuracy for LaST and DeepChange datasets indicates the effect of low-quality images on the off-the-shelf Face recognition model InsightFace [10]

Figure 1 shows our base model for learning clothes invariant features. Base model adopts a two-branch architecture [10, 8], where Bag-of-tricks (BOT [10]) branch is added to CAL (ResNet-50 backbone split identically after the second block).

CAL branch learns clothes invariant features, which helps the BOT branch via *shared backbone*, and KL Divergence loss on identity classifier of both the branches  $L_{ID,KL}$ . BOT branches simply stabilize CAL adversarial learning (as described as “CAL sensitivity problem”). BOT Branch trains with additional losses and transfers knowledge to the CAL branch via a shared backbone. BOT Branch also learns clothes invariant features Clothes augmented data points, will are likely noisy. Such two-branch architectures have also been studied in existing works [6, 8, 13, 14]. We adopt Huang *et al.* [8] method of “Max-Avg Pooling” (concat of global max and average pool) to produce  $f_{CG}, f_{CA} \in \mathbb{R}^{4096}$  from both the branches. Concatenation of these features is used for inference  $f_{ReID} (= [f_{BOT}, f_{CAL}])$ . Train-only Identity (ID) classifiers predict ID labels on each branch and KL-divergence on these ID logits ( $L_{ID,KL}$ ) helps transfer knowledge across the branches indirectly. Identity logits (via Identity classifiers)  $y_{CG}^{ID}, y_{CA}^{ID}$  predict the identity of the person during training only. We shall

Method	LTCC				PRCC				CCVID [8]			
	CC		genral		CC		SC		CC		general	
	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP
Vanilla	40.1	18.0	74.2	40.8	55.2	55.8	100	99.8	81.7	79.6	82.6	81.3
+ Triplet Loss	34.7	16.6	71.8	37.5	48.6	49.7	100	99.8	81.1	77.0	81.5	78.1
	(5.4↓)	(1.4↓)	(2.4↓)	(3.3↓)	(6.6↓)	(6.1↓)			(0.6↓)	(2.6↓)	(1.1↓)	(3.2↓)

Table 2: Effect of Triplet loss on CAL

collectively refer to these two branches as **Base Model**. Both these branches are briefly summarized below.

**CAL Branch:** This branch focuses on clothing information (orange in the figure) and implements CAL [8], an RGB-only model with Bag-of-Tricks (BOT) [12] framework (excluding Triplet and Center Loss). It uses an additional clothes classifier that predicts cloth labels ( $y_{CA}^{CL}$ ). A clothes Adversarial Loss ( $\mathcal{L}_{CAL}$ ) penalizes the clothes-relevant features for correct clothes predictions ( $\mathcal{L}_{CL,CE}$ ), thereby generating cloth robust features.

**BOT Branch:** Standalone BOT Branch (green in the figure) is a simple ReID model (dilated strides in ResNet-50 last block, with Triplet Loss ( $\mathcal{L}_{Trip}$ )), similar to BOT. This structure is commonly adopted by almost all ReID models due to its compatibility with different loss functions. In our work, we use it to disentangle pose and recognize gender (identity-related features).

We use *clothes augmented images* [8, 19] via body parsing [10] to help the vanilla BOT branch with clothes changing. Table 5 (main submission) indicates that this additional clothes augmentation and two branch structures offer no performance boost, but simply stabilizes CAL for additional training modules. Basemodel (both CAL and BOT branch) is used during inference.

### 3 Clothes Augmentation (part of Base Model)



Figure 2: **Clothing In-paint:** Upper-Lower Body Masks by SCHP [10], shown on LTCC Dataset. Random colors generate new clothing labels.

A lot of ReID models use silhouettes-based body parsing for clothes-changing as an integral part of their framework. Contrary to this, our clothes-augmented samples as used only as additional augmentations and not part of our framework, because of the additional noise

these silhouettes (fine-grained attributes) can bring to the model. Body parsing allows us to randomly change the upper and lower body colors as shown in Figure 2. Noisy silhouette-based clothes augmentation does not affect the identity (*BOT Branch*), gender (*Gender Classifier*), or pose (*POSE Branch*). Specifically for the *CAL Branch*, these clothes have been given their separate clothing labels.

## 4 Design Choices (More experiments)

Table 3 shows the optimal split of the shared backbone between CAL and BOT branches occurs after the second block. Table 4 shows the split of Pose Branch, where we make the architectural choice of split after the first block based on the highest Top-1 accuracy.

Shared Backbone Split	LTCC		PRCC	
	R-1	mAP	R-1	mAP
Disjoint	40.1	19.4	57.9	59.3
After 1 <sup>st</sup> Block	39.8	19.5	57.3	59.1
After 2 <sup>nd</sup> Block	<b>41.1</b>	<b>20.3</b>	58.0	<b>59.4</b>
After 3 <sup>rd</sup> Block	<b>41.1</b>	19.0	<b>58.8</b>	58.6

Table 3: **Backbone Split**: Disjoint means separate backbones (Base Model)

Pose Branch Split	LTCC		PRCC	
	R-1	mAP	R-1	mAP
1 <sup>st</sup> Block	<b>46.4</b>	21.5	<b>64.0</b>	63.2
2 <sup>nd</sup> Block	44.4	21.5	62.6	<b>64.2</b>
3 <sup>rd</sup> Block	45.9	<b>21.6</b>	63.8	63.9

Table 4: **Pose Branch Split** on shared base in RLQ.

## 5 Effect of Low Quality Images on Soft Biometrics

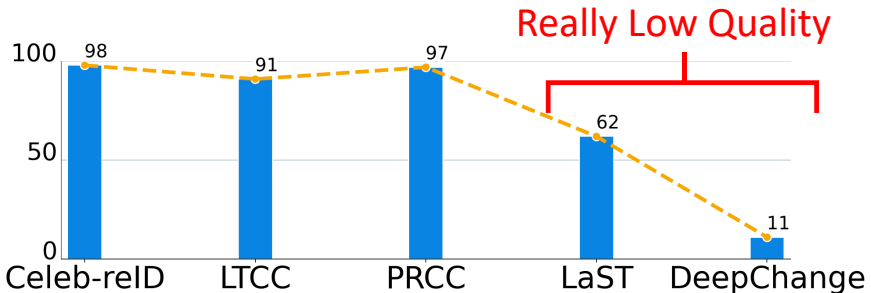


Figure 3: **Gender (soft biometrics)**: performance on ‘male’ gender classification for all the datasets. The sharp drop in accuracy for LaST and DeepChange datasets indicates the effect of low-quality images on the off-the-shelf Face recognition model InsightFace [2]

Traditionally gender classification is used during inference time to filter out wrong genders for increasing ReID accuracy, commonly referred to as “Soft Biometrics” [4, 10]. Figure 3 shows the effect of low-quality images on the gender recognition capability of Face Recognition Model InsightFace [10]. This is one of the key reasons we opt for manual annotation of gender labels in the training set.

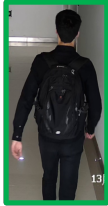


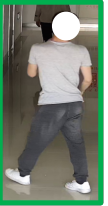

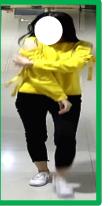
	HQ	Pixelation & OOF	Motion Blur	HQ	Pixelation	HQ
LTCC Data Points						
BLUR: FFT	9.24	9.92	5.33	8.78	8.66	13.31
Blur: (Laplacian)	255	110	210	234	139	206
Blur (SVD)	0.60	0.73	0.77	0.71	0.66	0.68
Motion Blur: Laplacian	112.61	99.66	65.90	115.68	111.35	133.46

Figure 4: Edge detection scores for Fast Fourier Transform (FFT), Laplacian, Singular value decomposition (SVD), for various artifacts. High-quality (HQ) images (shown in green) should have a distinctively separate value compared to low-quality (LQ) images (shown in red), but there are similar conflicting values as shown in red.

## 6 Analysis of high-quality and low-quality data (Failure of Traditional methods)

Distinguishing between high-quality (HQ) and low-quality (LQ) data points, especially in the presence of artifacts like pixelation, motion blur, and out-of-focus blur, is a challenging task. While pixelation can be easily detected in images simply by looking at the spatial dimension ( $\leq 64 \times 64$ ), it’s not possible to identify pixelated images once they’ve been resized. A common alternative for detecting these artifacts, thereby separating LQ images, is to use edge detectors such as Laplacian filters, Singular Value Decomposition (SVD), and Fast Fourier Transform (FFT). These tools score images, categorizing those above (or below) certain predefined thresholds as either blurry or high-quality. However, as shown in the fig. 4, these metrics often produce conflicting values (in red) for HQ (green) and LQ (red) data points, making it difficult to establish a clear boundary between the two. Due to the lack of a consistent trend in these conflicting values, an external high-quality dataset, Celeb-ReID [10], is preferred for generating synthetic HQ-LQ image pairs and studying their relationships.

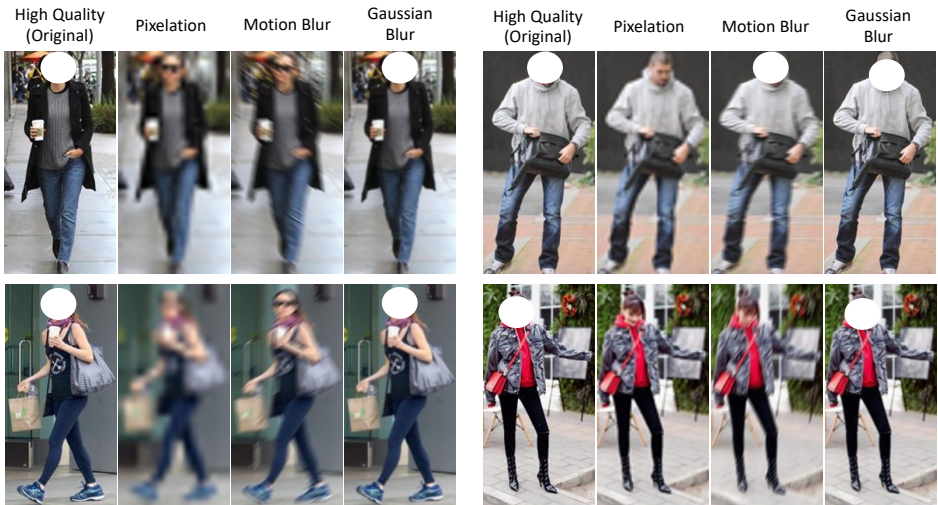


Figure 5: Synthetic Low-Quality images, with images generated from Celeb-ReID. Gaussian blur simulates out-of-focus blur.

## 7 Visualization of Synthetic Low Quality Images (Used in TBD)

We use Celeb-ReID [1], as a source of High-Quality images. We simulate real-world artifacts on it, namely pixelation, motion blur, and Gaussian blur (out-of-focus blur) as shown in Figure 5. The method of generating these artifacts is explained in the *under Implementation Details in main submission*.

## 8 Additional Training Details

All branches are ImageNet pre-trained ResNet-50 blocks. We use data augmentation like random horizontal flipping, erasing, and cropping along with clothes augmentation. Raw RGB images along with their synthetic clothing samples are normalized ( $\mu = [0.485, 0.456, 0.406]$ ,  $\sigma = [0.229, 0.224, 0.225]$ ). We use Adam optimizer with warmup learning rate ( $\text{lr}=0.00035$ ). No hyperparameters were done, thus all loss terms were assigned a weight of 1.0. Models are trained on with 4 positive samples per batch (for Triplet loss). Gender labels can be manually generated ( $0^{\text{th}}$  class in doubt), given the limited number of unique IDs in the training data, and errors in the genre recognition model on low-quality datasets. Pose vectors are computed on resized images, *i.e.* resized poses. For TAD, the Teacher Base Model is pre-trained on the Celeb ReID dataset. All results are an average of two runs, with evaluation at every 10th epoch. We report the best Top-1 and mAP scores obtained throughout all the evaluations.

Branches are ImageNet initialized (no pre-training), with output ‘Max-Avg’ pooled (concat of max and avg pool [8]). Model is trained end-to-end for 200 epochs (300 epochs for LaST) with 40 batch sizes. Gender is manually annotated as 1(male), and 0(female). Al-

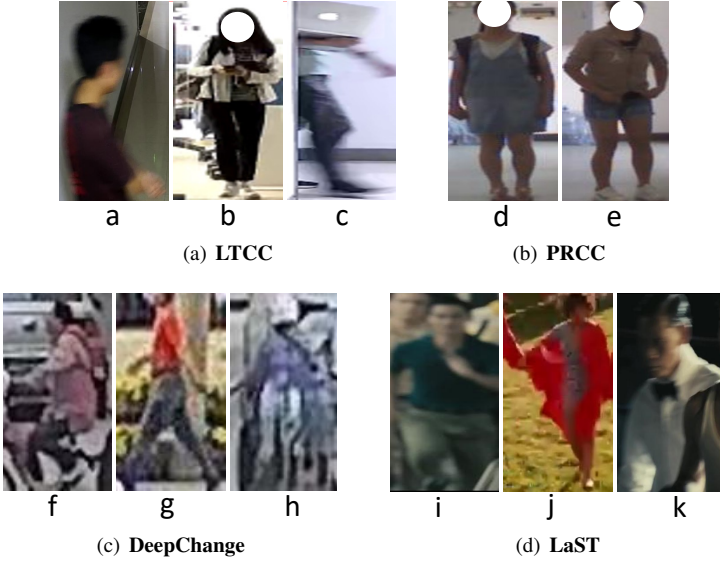


Figure 6: **Dataset challenges:** PRCC & DeepChange have pixelation (b, f, g, h, j) and out-of-focus blur (a, d, e, k). LaST & LTCC additionally have motion blur (c, i, k)

phaPose [8] is used to generate fine-grained 2D skeletons. 13 body lines Pose vectors are clustered with an empirically found cluster size of 15 ( $0^{th}$  cluster for noisy poses). We use Celeb-ReID [10] without labels as an external HQ dataset. In PRCC and DeepChange, **pixelation** and **out-of-focus** (OOF) is simulated with  $\frac{1}{2}$  probability. **Motion blur** is added for LTCC and LaST. Pixelation is downsampling to resolution  $\in [16 \times 16, 64 \times 64]$ , and resizing back. *OOF* blurring is Gaussian blur with a random kernel  $\in [5, 21]$ . *Motion blur* is obtained by applying a random size kernel  $\in [8, 20]$  of 1s rotated randomly  $\in [0^\circ, 180^\circ]$ . AlphaPose [9] as 2D skeletons consisting of 17 joints, 13 body lines (stretch and angles), clustered with an empirically found cluster size of 15 ( $0^{th}$  cluster for noisy poses).

## 9 Dataset Samples (RGB Examples)

**LTCC**[13] has diverse poses from 12 indoor camera views, with noticeable motion, out-of-focus (OOF) blurring, and pixelation (fig. 6(a)). **PRCC**[17] is relatively high-resolution and has some OOF and pixelation, featuring 3 indoor camera views (fig. 6(b)). **DeepChange**[16] is captured across different weather conditions (12 months) in 17 camera views with massive pixelation and OOF (fig. 6(c)). **LaST**[14] consists mostly of web/movie images with similar defects as LTCC (fig. 6(d)).

## 10 Visualization of Pose Clusters (RGB Examples)

We obtain pose representation from off-the-shelf pose detector AlphaPose [9], which consists of key points (joints) and body lines (distance between each body joint). We cluster pose vectors using K-means with optimal performance at around 15 clusters as indicated Figure 7. We have two major spikes in accuracy, around the 15th and 25th clusters.



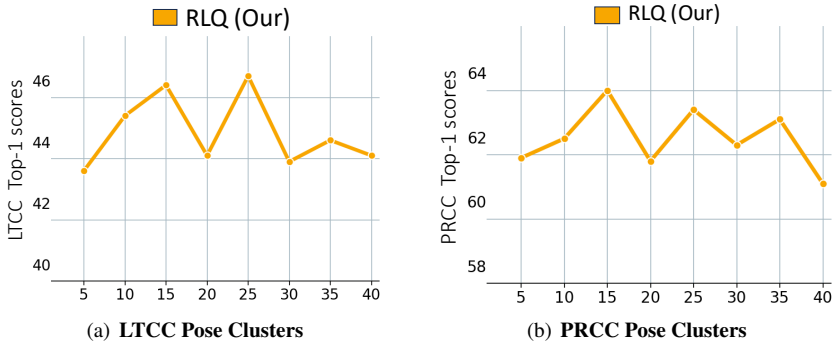


Figure 7: **Optimal K-means cluster size** of pose vectors on the LTCC and PRCC dataset.



Figure 8: **Actual Pose Clusters** Pose clustering on the LTCC dataset, with 15 clusters. Clusters are re-numbered to better fit. Clusters from 0 - 4 are shown.

For better understanding, we have shown these pose clusters with their RGB images in Figure 8 and Figure 9. We have also highlighted the commonalities across RGB images in their pose clusters. Pose Cluster 0 consists mainly of extreme outliers with most artifacts where the pose model couldn't detect any pose, thereby assigning the default 0th cluster.

## 11 Ethical Statement

Our research targets real-world problems in ReID. However, we recognize the potential risk for misuse in tracking and targeting individuals. To safeguard against this, we will release some aspects of codes only via emails, for academic institutions, ensuring work's full potential is limited to academic research.





Figure 9: **Actual Pose Clusters:** Pose clustering on the LTCC dataset, with 15 clusters. Clusters are re-numbered to better fit. Clusters from 5 - 15 are shown.

## References

- [1] Zhenyu Cui, Jiahuan Zhou, Yuxin Peng, Shiliang Zhang, and Yaowei Wang. Dcr-reid: Deep component reconstruction for cloth-changing person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):4415–4428, 2023. doi: 10.1109/TCSVT.2023.3241988.
- [2] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.
- [3] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017.

- [4] Hiren Galiyawala, Kenil Shah, Vandit Gajjar, and Mehul S Raval. Person retrieval in surveillance video using height, color and gender. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.
- [5] Xinqian Gu, Hong Chang, Bingpeng Ma, Shutao Bai, Shiguang Shan, and Xilin Chen. Clothes-changing person re-identification with rgb modality only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1060–1069, 2022.
- [6] Peixian Hong, Tao Wu, Ancong Wu, Xintong Han, and Wei-Shi Zheng. Fine-grained shape-appearance mutual learning for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10513–10522, June 2021.
- [7] Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. Celebrities-reid: A benchmark for clothes variation in long-term person re-identification. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.
- [8] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing status awareness for long-term person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11895–11904, 2021.
- [9] Xuemei Jia, Xian Zhong, Mang Ye, Wenxuan Liu, and Wenxin Huang. Complementary data augmentation for cloth-changing person re-identification. *IEEE Transactions on Image Processing*, 31:4227–4239, 2022. doi: 10.1109/TIP.2022.3183469.
- [10] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. doi: 10.1109/TPAMI.2020.3048039.
- [11] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern recognition*, 95:151–161, 2019.
- [12] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.
- [13] 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.
- [14] Xiujun Shu, Xiao Wang, Xianghao Zang, Shiliang Zhang, Yuanqi Chen, Ge Li, and Qi Tian. Large-scale spatio-temporal person re-identification: Algorithm and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 4390–4403, 2022.
- [15] Dingyi Wang and Haishun Du. Double-stream network for clothes-changing person re-identification based on clothes related feature suppression and attention mechanism. In Wang Yongtian and Wu Lifang, editors, *Image and Graphics Technologies and Applications*, pages 208–219, Singapore, 2023. Springer Nature Singapore. ISBN 978-981-99-7549-5.

- [16] Peng Xu and Xiatian Zhu. Deepchange: A large long-term person re-identification benchmark with clothes change. *arXiv preprint arXiv:2105.14685*, 2021.
- [17] Qize Yang, Ancong Wu, and Wei-Shi Zheng. Person re-identification by contour sketch under moderate clothing change. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2029–2046, 2019.
- [18] Junhao Zheng, Xiaoman Hu, Tianyi Xiang, and Patrick P. K. chan. Dual-path model for person re-identification under cloth changing. In *2020 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 291–297, 2020. doi: 10.1109/ICMLC51923.2020.9469545.
- [19] Zihui Zhou, Hong Liu, Wei Shi, Hao Tang, and Xingyue Shi. A cloth-irrelevant harmonious attention network for cloth-changing person re-identification. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 989–995, 2022. doi: 10.1109/ICPR56361.2022.9956160.