

# Supplementary Materials

Anonymous Authors

## 1 EXPERIMENT AND DISCUSSION

### 1.1 Structure of backbone network

In our method, we divide ResNet50 into five layer blocks (e.g., Blocks 1, 2, 3, 4, 5 represent *conv1*, *conv2\_x*, *conv3\_x*, *conv4\_x*, *conv5\_x* in the ResNet50 structure, respectively.). Our backbone network comprises a shared image encoder (denoted as  $\mathcal{F}_{sh}(\cdot)$ ) composed of the first four layer blocks, and two parts without shared parameters: an identity feature extractor (denoted as  $\mathcal{F}_{id}(\cdot)$ ) and a specific feature extractor (denoted as  $\mathcal{F}_{int}(\cdot)$ ), both consisting of the fifth layer block. We compared the performance of the models with different numbers of layer blocks in ( $\mathcal{F}_{sh}(\cdot)$ ) under the cloth-changing setting on the LTCC dataset. The overall experimental results are reported in Table 1. As shown in Table 1, as the structure of the  $\mathcal{F}_{sh}(\cdot)$  expands from Block 1 to Block 1-4, the model develops more robust feature representations. This improvement occurs because our  $\mathcal{F}_{int}(\cdot)$  more accurately extracts interference factors from identity features as the number of shared layer blocks increases. However, when  $\mathcal{F}_{sh}(\cdot)$  is increased to Block 1-5, there is a significant decline in performance. This decline is due to the overlap between interference factor extraction and pedestrian discrimination processes, which fails to disentangle identity features from interference factors.

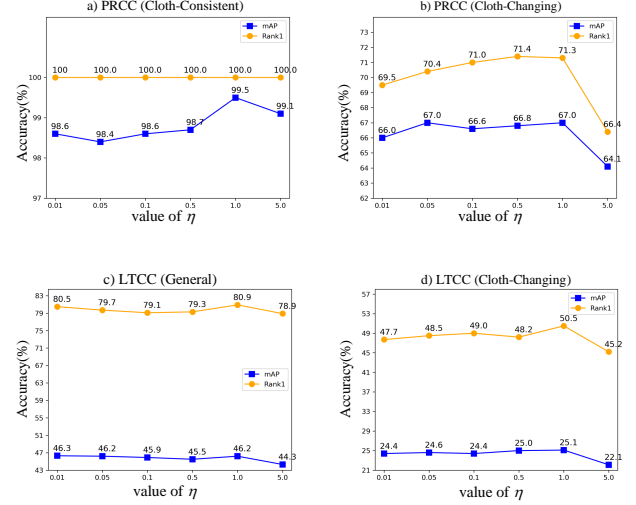
**Table 1: Ablation study of the different structural models under cloth-changing setting on LTCC.**

		LTCC	
$\mathcal{F}_{sh}(\cdot)$	$\mathcal{F}_{id}(\cdot)$ / $\mathcal{F}_{int}(\cdot)$	Rank-1	mAP
Block 1	Block 2-5	48.2	24.1
Block 1-2	Block 3-5	49.0	23.5
Block 1-3	Block 4-5	48.2	23.8
Block 1-4	Block 5	50.5	25.1
Block 1-5	-	42.1	22.0

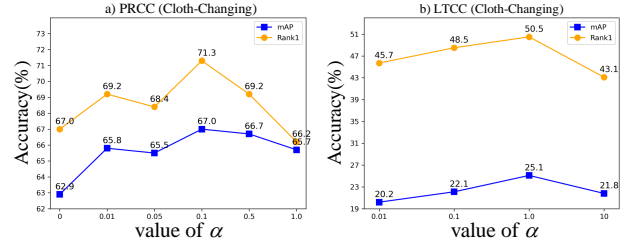
### 1.2 Hyper-Parameter Sensitivity analysis

**Influence of  $\eta$ :** We analyze the sensitivity of the parameter (i.e., the weight  $\eta$  in Eq. (15)) in our method. We tune the values of each parameter, keeping the others fixed, on the LTCC and PRCC. The results are shown in Fig. 1. When  $\eta$  is set to a small value (e.g., from 0.01 to 1.0), the experimental results exhibit an upward trend. However, when  $\eta$  is set to a too large value (e.g., 10), the model results show a significant decline. Based on these experiments, we set the hyper-parameter  $\eta = 1.0$  across all datasets (PRCC, VC-Clothes, LTCC).

**Influence of  $\alpha$ :** We analyze the sensitivity of the parameter in our method, (i.e., the weight  $\alpha$  in Eq. (15)). The results are shown in Fig. 2. When  $\alpha$  is set to a small value, the model cannot extract



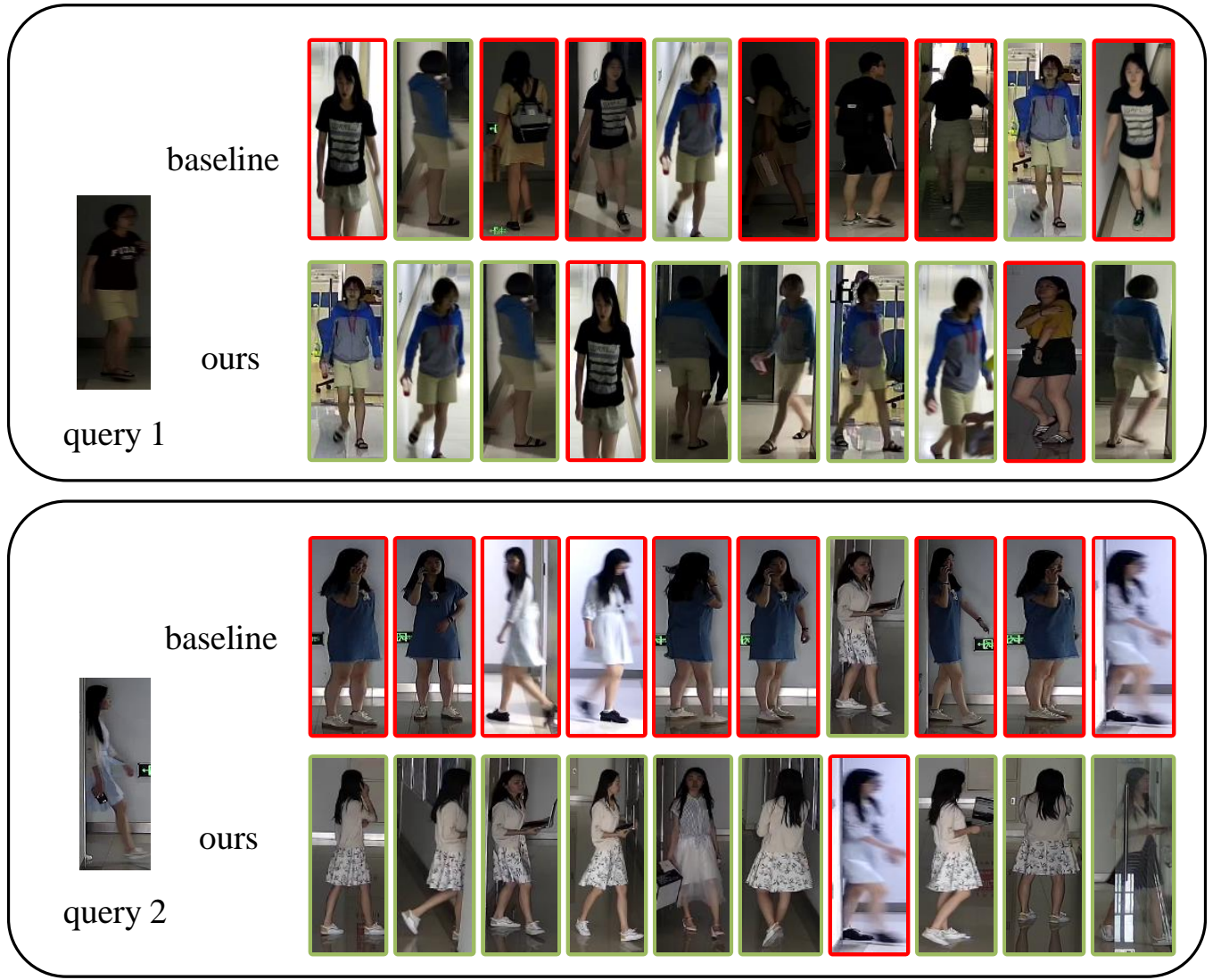
**Figure 1: Parameter sensitivity analysis on  $\eta$  in Eq. (15) on PRCC and LTCC.**



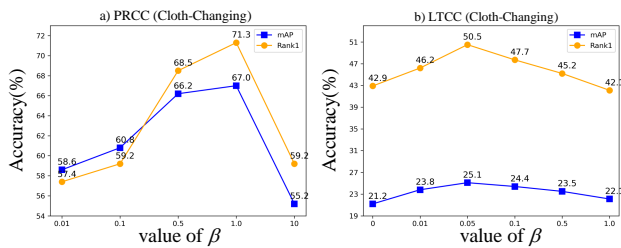
**Figure 2: Parameter sensitivity analysis on  $\alpha$  in Eq. (15) on PRCC and LTCC under cloth-changing setting.**

accurate interference factors, resulting in a reduced decoupling effect. When  $\alpha$  is set to a large value, the model focuses more on extracting interference factors and ignore the extraction of identity features, resulting in a declined effect. Based on these experiments, we set  $\alpha$  to 0.1 and 1.0 for the PRCC and LTCC datasets, respectively.

**Influence of  $\beta$ :** We analyze the sensitivity of the parameter in our method, (i.e., the weight  $\beta$  in Eq. (8)). The results are shown in Fig. 4. The model is very sensitive to the changes in the value of  $\beta$ . As shown in the results of Fig. 4 a), when  $\beta$  is set to a small value (e.g., 0.01 ~ 0.1), the model has insufficient ability to match pedestrian images and cloth-erased images. When  $\beta$  is set to a large value (e.g., 10), the model overly focuses on identical parts of the pedestrian and clothes-erased images, introducing background interference into the identity feature. When the value of  $\beta$  is set within the appropriate range (e.g., 0.5 ~ 1.0), the model benefits from the learned identity features to be discriminative among different pedestrians while irrelevant to diversified cloth texture. Based on these experiments, we set  $\beta$  to 1.0 and 0.05 for the PRCC and LTCC datasets, respectively.



**Figure 3: According to the top 10 retrieval results of baseline and our method on the LTCC dataset under the cloth-changing setting. Images in green and red boxes are positive and negative results, respectively.**



**Figure 4: Parameter sensitivity analysis on  $\beta$  in Eq. (8), on PRCC and LTCC under cloth-changing setting.**

### 1.3 Visualization

To intuitively demonstrate the effect of our model on CC-ReID, we visualize the top-10 ranked retrieval results for both the baseline method and our approach on the LTCC dataset, under the cloth-changing setting. As defined in a previous article, the baseline method (Fig. 3) involves training the dual-stream backbone network using  $L_{ce}$  from Eq. (1) and  $L_{tri}$  from Eq. (2). As shown in Fig. 3, it can be observed that the baseline will return incorrectly matched pedestrian images due to the interference of lighting, viewpoint and clothes. For example, as shown in query 1, the retrieval results of the baseline method incorrectly match different pedestrians with similar lighting and clothes. The retrieval results of our method shows robustness to richer backgrounds and lighting, and largely

overcome the interference from cloth factors. These results confirm that our method can effectively resist the influence of these interference factors and shows better robustness than the baseline.

291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348