Learning Modal-Invariant and Temporal-Memory for Video-based Visible-Infrared Person Re-Identification — Supplementary Materials

Xinyu Lin[1], Jinxing Li[1] ✉, Zeyu Ma[1], Huafeng Li[2], Shuang Li[2],Kaixiong Xu[2], Guangming Lu[1], David Zhang[3]
[1]Harbin Institute of Technology, Shenzhen, [2]Kunming University of Science and Technology,
[3]The Chinese University of HongKong, Shenzhen.
{linxinyu0327, lijinxing158}@gmail.com, zeyu.ma@stu.hit.edu.cn, hfchina99@163.com,
{shuangli936, xukaixiong99}@gmail.com, luguangm@hit.edu.cn, davidzhang@cuhk.edu.cn

1. Dataset Collection Description

In this part, we introduce the data collection and permissions.

1.1. Locations

We collect data at two locations, both of which are campuses. We placed 6 cameras on each campus and recruited volunteers to collect the person data. Note that we carried out the collection work with the consent of both schools. To protect anonymity, we do not discuss the detailed collection locations.

1.2. Approval

We collected original data by volunteer recruitment. We obtained consent of each volunteer and every volunteer was aware of the usage of the data: for academic research only. Additionally, we provided financial rewards for volunteers to thank them for their participation in the data collection.

Of course, we have obtained the IRB protocol approval from our institution: The Chinese University of Hong Kong, Shenzhen. The protocol number is: CUHKSZ-D-20220003. If you are interested, you can email us to get the soft copy of the document.

2. Experimental Results and Implementation

2.1. Ablation Study

Experimental results of 'Visible to Infrared' search mode. Due to page limitation, we report the 'Ablation Study' of 'Visible to Infrared' search mode in this supplementary material. Tab. 2 illustrates the effectiveness of our TMR module and the modal-invariant learning module. As we can see, our proposed modules both contribute to performance improvement. Besides, 'Full method^S' in Tab. 1 proves the effectiveness of temporal information for we shuffle the frames in a tracklet to remove the chronological order. Tab. 2 displays the comparisons of our modal-invariant learning with two existing adversarial learning strategies in the 'Visible to Infrared'. It is easy to see that our adversarial learning strategy is superior to that of [1] and [2].

Loss function parameter adjustment. We also evaluate loss functions with different weighted terms. Consider that the term $\mathcal{L}_{adv1}$ to a great extent determines the learning of modal-invariant features, here we mainly test the weight

✉ Jinxing Li is the Corresponding Author.

Table 1. Effectiveness of the TMR module and the adversarial learning module. CMC (%) and mAP (%) are reported. Note that M denotes the modal-invariant learning, while T denotes the temporal memory refinement, and Full method^S denotes full method with shuffled frames in TMR.

Visible to Infrared					
Strategy	R1	R5	R10	R20	mAP
Baseline	59.58	74.43	79.25	83.74	42.61
Baseline + M	60.54	75.84	81.15	85.59	43.59
Baseline + T	62.19	76.11	80.84	84.98	46.02
Full method^S	63.72	77.72	82.70	86.90	45.64
Full method	**64.54**	**78.98**	**82.98**	**87.10**	**47.69**

Table 2. Comparisons of our modal-invariant learning with different adversarial strategies. For a fair comparison, we only replaced our modal-invariant learning with other strategies in our network.

Visible to Infrared					
Strategy	R1	R5	R10	R20	mAP
cmGAN [1]	60.68	75.25	80.29	85.00	44.63
UCDA [2]	64.15	77.47	82.35	86.21	47.08
Our method	**64.54**	**78.96**	**82.98**	**87.10**	**47.69**

Table 3. Evaluations of loss function with different weight.

Infrared to Visible					
Parameter λ	R1	R5	R10	R20	mAP
0.01	60.80	74.76	80.06	85.17	45.54
0.05	63.05	76.81	81.83	86.08	46.50
0.1	62.98	76.40	81.15	85.54	46.19
0.2	63.53	76.81	81.26	85.89	46.49
0.4	**64.62**	**77.12**	**82.37**	**86.81**	**47.32**
0.6	63.46	76.83	81.65	85.71	45.71
0.8	63.48	76.83	82.07	86.34	45.46
1.0	63.74	76.88	81.72	86.28	45.31

Table 4. Evaluations of different pooling strategies in baseline. 'Weighted pooling' means weighted average pooling, the attention scores of which are computed based on Softmax function.

Infrared to Visible					
Strategy	R1	R5	R10	R20	mAP
Average pooling	55.58	70.75	77.01	82.16	40.80
Max pooling	54.47	70.24	76.51	81.83	41.11
Weighted pooling	47.49	64.88	72.43	79.04	36.52

of this term. The whole objective function can be represented as:

$$\mathcal{L} = \lambda \mathcal{L}_{adv1} + \mathcal{L}_{id}^{ce} + \mathcal{L}_{id}^{tri} + \mathcal{L}_{adv2} \quad (1)$$

where λ is the weight of term $\mathcal{L}_{adv1}$. As shown in Tab. 3, as λ changes from 0.2 to 1.0, the experimental results fluctuate slightly, which indicates that our algorithm is very robust. When λ is 0.4, our proposed model achieves best performance.

Baseline pooling strategy. Since the inputs of the network are multiple images, an average pooling layer is utilized in our baseline to fuse the frame-level features obtained from the backbone. Here, we conduct different pooling strategies for baseline, including max pooling and weighted average pooling, as shown in Tab. 4. Note that 'weighted pooling' in Tab. 4 denotes weighted average pooling, and we compute the attention scores based on the Softmax function. As we can see, the max pooling strategy is also adaptive for our baseline.

2.2. Implementation

Synchronization of cross-modal data. To synchronize RGB and IR tracklets, we shuffle all tracklets first and then select the same number of RGB and IR tracklets batch by batch. When all IR tracklets are selected, we shuffle all tracklets again.

References

[1] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, volume 1, page 2, 2018. 1

[2] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *ICCV*, pages 8080–8089, 2019. 1