Visible-Infrared Person Re-Identification via Semantic Alignment and Affinity Inference (Supplementary Materials)

This supplementary material presents additional details and results not included in the main paper due to space limitations. We organize the contents as follows:

- Section A provides analysis of the hyper-parameters.
- Section B compares AIM with SIM.
- Section C shows details of the classification loss.
- Section D presents visualizations of retrieval results.

A. Analysis of the Hyper-parameters

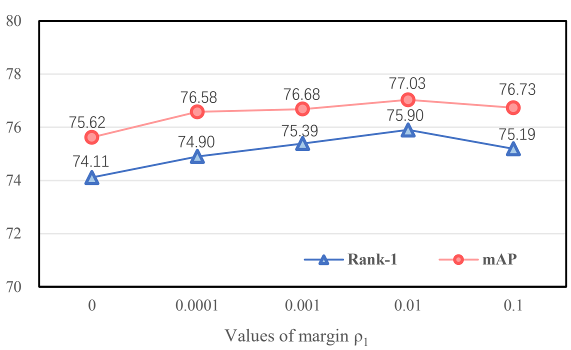

Figure 1. The sensitive graph of the margin ρ_1 on the SYSU-MM01 dataset in all-search and single-shot mode.

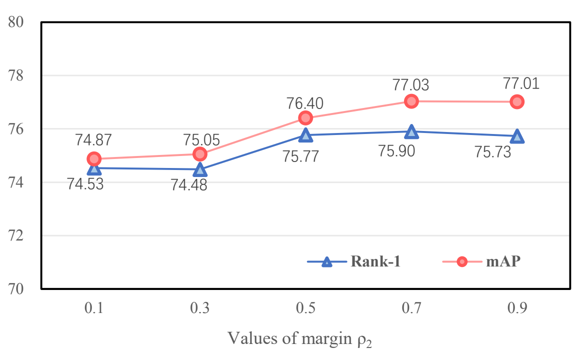

Figure 2. The sensitive graph of the margin ρ_2 on the SYSU-MM01 dataset in all-search and single-shot mode.

Evaluation of parameters ρ_1 and ρ_2 in Eq. (5). We evaluate the effects of margin ρ_1 and margin ρ_2 in Eq. (5). We first set ρ_2 to 0.7 and experiment with different values for ρ_1. As shown in Figure 1, the performance improves with the increase of ρ_1 until it reaches 0.01. Increasing ρ_1 enhances the diversity of features, thereby reducing the risk of overfitting in the network. However, if ρ_1 is set too high, it may cause features of the same pedestrian to be dispersed, leading to reduced discriminability and making it difficult to identify pedestrians accurately. Hence, we set ρ_1 to 0.01 to balance diversity and discriminability.

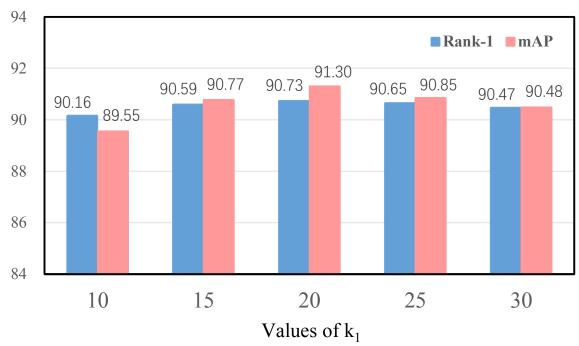

Figure 3. The sensitive graph of k_1 on the SYSU-MM01 dataset in indoor-search and multi-shot mode.

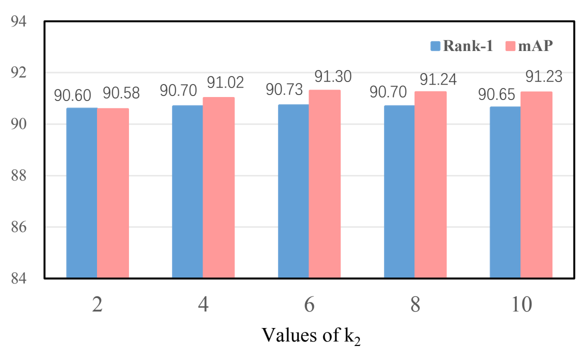

Figure 4. The sensitive graph of k_2 on the SYSU-MM01 dataset in indoor-search and multi-shot mode.

We set ρ_1 to 0.01, and vary ρ_2 from 0.1 to 0.9 with an interval of 0.2. As shown in Figure 2, we observe an initial improvement in performance with the increase of ρ_2, which reaches its maximum value at 0.7. Hence, we set ρ_2 to 0.7.

Evaluation of parameters k_1 and k_2 in AIM. We set k_2

Figure 5. The visualization of retrieval results on SYSU-MM01 in all-search and multi-shot mode. We use different distance measurement methods (*e.g.,* cosine distance, SIM [1], and AIM) to match pedestrian images. The green border represents the correct example, and the red border denotes the wrong example.

to 6, and change k_1 from 5 to 30 with an interval of 5. As shown in Figure 3, the performance is improved with the increase of k_1 at first, and achieve the best performance when k_1 = 20. Therefore, we set the value of k_1 to 20.

We set k_1 to 20, and change k_2 from 2 to 10 with the interval of 2. As shown in Figure 4, the performance is improved with the increase of ρ_2 at first, and achieve the best results when k_2 = 6. Hence, we set k_2 to 6.

B. Compare AIM with SIM

We add AIM and SIM [1] to several methods (*e.g.*, AlignGAN [2], AGW [4], and MPANet [3]) for comparison. As shown in Table 1, AIM achieves better results than SIM. This result further proves the effectiveness of AIM.

Table 1. Evaluation of AIM and SIM on other methods on SYSU-MM01 under multi-shot setting. We retrain the models.

Method	all-search		indoor-search	
	Rank-1	mAP	Rank-1	mAP
AlignGAN*+SIM	50.85	44.96	57.90	56.84
AlignGAN*+AIM	**51.63**	**50.65**	**58.32**	**61.94**
AGW*+SIM	54.17	54.21	62.95	65.31
AGW*+AIM	**55.01**	**55.18**	**63.84**	**67.51**
MPANet*+SIM	78.40	77.32	84.92	85.15
MPANet*+AIM	**78.52**	**78.27**	**85.96**	**87.31**

C. Details of the classification loss

For both baseline and our model, we utilize the classification loss $\mathcal{L}_{id}$ to guide the model to focus on identical information. The classification loss can be defined as:

$$\mathcal{L}_{id} = \mathcal{L}_{sh\_id} + \mathcal{L}_{sp\_id} + \mathcal{L}_{cm\_id}. \quad (1)$$

$\mathcal{L}_{sh\_id}$ uses a shared classifier for both modalities as:

$$\mathcal{L}_{sh\_id} = -logP(y_{\mathrm{v}}|C(\tilde{\mathbf{F}}_{\mathrm{v}})) - logP(y_{\mathrm{r}}|C(\tilde{\mathbf{F}}_{\mathrm{r}})), \quad (2)$$

where $P(\cdot)$ is the probability of correct prediction, y_{v} and y_{r} are labels, and $C(\cdot)$ is a shared classifier. $\mathcal{L}_{sp\_id}$ replace $C(\cdot)$ with modality-special classifiers $C_{\mathrm{v}}(\cdot)$ and $C_{\mathrm{r}}(\cdot)$. Following [3], we use mean classifiers $\overline{C}_{\mathrm{r}}(\cdot)$ and $\overline{C}_{\mathrm{v}}(\cdot)$ as:

$$\mathcal{L}_{cm\_id} = \overline{C}_{\mathrm{r}}(\tilde{\mathbf{F}}_{\mathrm{v}})\log\frac{\overline{C}_{\mathrm{r}}(\tilde{\mathbf{F}}_{\mathrm{v}})}{C_{\mathrm{v}}(\tilde{\mathbf{F}}_{\mathrm{v}})} + \overline{C}_{\mathrm{v}}(\tilde{\mathbf{F}}_{\mathrm{r}})\log\frac{\overline{C}_{\mathrm{v}}(\tilde{\mathbf{F}}_{\mathrm{r}})}{C_{\mathrm{r}}(\tilde{\mathbf{F}}_{\mathrm{r}})}. \quad (3)$$

D. Visualization of retrieval Results

For each query image, we retrieve the top 10 gallery images with the highest similarity and rank them in descending order of similarity. As shown in Figure 5, AIM can achieve more stable matching results than other methods.

References

[1] Mengxi Jia, Yunpeng Zhai, Shijian Lu, Siwei Ma, and Jian Zhang. A similarity inference metric for rgb-infrared cross-modality person re-identification. In *IJCAI*, page 1026–1032, 2020. 2, 3

[2] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*, pages 3623–3632, 2019. 3

[3] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *CVPR*, pages 4330–4339, 2021. 3

[4] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *TPAMI*, 44(6):2872–2893, 2021. 3