Supplementary Material for WRIM-Net

Yonggan Wu[1,2,3], Ling-Chao Meng[3], Yuan Zichao[3], Sixian Chan[2,4], and Hong-Qiang Wang[2]*

[1] University of Science and Technology of China, Hefei 230026, China
braverywu@mail.ustc.edu.cn
[2] Institute of Intelligent Machines/Zhongqi AI Joint Lab., HIPS, CAS, Hefei, China
hqwang126@126.com
[3] QiXinMingZhi Technology, Hefei 230088, China
[4] Zhejiang University of Technology, Hangzhou 310014, China

A The detail of $\mathcal{L}_{CMKIC\_P_5\_IV}$, $\mathcal{L}_{\mathbf{cls}\_P_4}$ and $\mathcal{L}_{\mathbf{tri}\_P_4}$

$\mathcal{L}_{CMKIC\_P_5\_IV}$ is similar to $\mathcal{L}_{CMKIC\_P_5\_VI}$, and represents its symmetric loss, as shown in the following formula:

$$\mathcal{L}_{CMKIC\_P_5\_IV} = \sum_{i \in \boldsymbol{I_{infra}}} \frac{-1}{\text{top-K}} \sum_{p \in \boldsymbol{P_{vis}(i)}} \log \frac{\exp(\boldsymbol{z_i} \cdot \boldsymbol{z_p}/\tau)}{\sum_{a \in \boldsymbol{A(i)}} \exp(\boldsymbol{z_i} \cdot \boldsymbol{z_a}/\tau)}. \tag{1}$$

$$\mathcal{L}_{cls\_P_4} = \frac{\left(\mathrm{E}(-\log p(\boldsymbol{Q_g})) + \sum_{i=1}^{M} \mathrm{E}(-\log p(\boldsymbol{Q_i}))\right)}{M+1}, \tag{2}$$

where $p()$ is the probability that the feature is correctly predicted by the classifier and E represents the expectation, M represents the number of partitions for the feature Q_g.

$$\mathcal{L}_{tri\_P_4} = E\left[d_p - d_n + \alpha\right]_+, \tag{3}$$

where d_p represents the feature distance belonging to the same ID, d_n represents the feature distance belonging to different IDs, and α is a margin parameter, where $[x]_+ = max(x, 0)$.

B Implementation details

Our proposed method is implemented using the PyTorch Fastreid [1] framework and is trained and inferred on an RTX3090 GPU. For RegDB, the number of split local features is $N = 6$ and $M = 2$, due to two modality images of this dataset is well aligned in the spatial. In each training batch, for the LLCM and RegDB datasets, we randomly selected 8 IDs per modality and sampled 8 random images for each of the selected IDs. For the SYSU-MM01 dataset, we

* Corresponding author.

randomly selected 4 IDs per modality and sampled 16 random images for each of the selected IDs. The input data's image size was adjusted to 384×144 for LLCM and SYSU-MM01, and 256×144 for RegDB. Subsequently, we employed data augmentation methods, including multi-scale cropping, random horizontal flip, random erase, and random grayscale transformation, on the input data. In addition, we used the Adam optimizer and optimized with an initial learning rate of $LR = 3.5 \times 10^{-4}$ and a weight decay value of $WD = 5 \times 10^{-4}$. We decayed the learning rate to the original 0.1 and 0.01 for 80 and 120 training epochs, respectively. The total number of training epochs is 140. The hyper-parameters λ_1 and λ_2 were set to 0.5 and 0.1, respectively. The margin parameter α in the triplet objective function is 0.3, and the temperature parameter τ in the CMKIC loss is 0.07.

C Analysis of parameters λ_1 and λ_2

We evaluate the hyperparameters λ_1 and λ_2 in formula 11 on the LLCM dataset. The Rank-1 and the mAP results of WRIM-Net with different λ_1 and λ_2 are exhibited shown in Fig S1. As the figure shows, setting λ_1 and λ_2 to 0.5 and 0.1 respectively gives the best results.

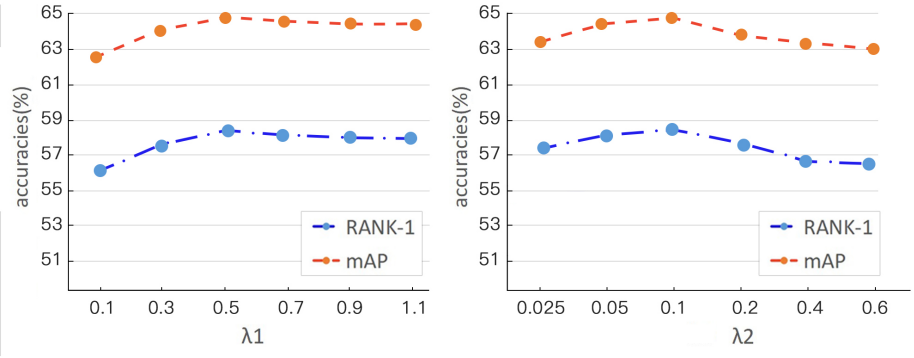

Figure S1. We have evaluated the parameters λ_1 and λ_2 in Formula 9 on the LLCM dataset. First we fix λ_2 at 0.1 and vary λ_1. Next we fix λ_1 at 0.5 and vary λ_2. The mode is VIS to IR.

D Impact of Different Configurations of MIIM

The results of different configurations of MIIM are presented in Table 4 of the main text. Table S1 supplements the details provided in Table 4. Notably, employing separate MIIM after Blocks 1 and 2 individually leads to significantly greater performance improvements compared to placing them after Blocks 3 and 4. We analyze the impact of configuring all MIIMs after the four blocks as separate or shared on network performance, as illustrated in table S2. From table

Table S1: Impact of different placement of MIIMs on LLCM. The MIIM placed after Block 1 and 2 is the separate MIIM, while the MIIM placed after Block 3 and 4 is the shared MIIM. Shared MIIM are weight shared, while separate MIIM are not.

Block1	Block2	Block3	Block4	Rank-1	mAP
×	×	×	×	54.83	61.27
✓	×	×	×	56.68	62.83
×	✓	×	×	57.34	63.53
×	×	✓	×	55.26	61.98
×	×	×	✓	55.03	61.76
✓	✓	×	×	57.82	64.03
×	×	✓	✓	56.16	62.72
✓	✓	✓	✓	**58.38**	**64.75**

Table S2: Impact of configuring all MIIMs as separate or shared after the four blocks on network performance on LLCM. "WRIM-Net (Share)" denotes the configuration where all MIIMs are shared after the four blocks, while "WRIM-Net (Separate)" denotes the configuration where all MIIMs are separate after the four blocks.

	Rank-1	mAP
WRIM-Net (share)	57.63	64.15
WRIM-Net (separate)	57.12	63.53
WRIM-Net	**58.38**	**64.75**

S2, we observe a decrease in metrics when replacing MIIM with all shared or all separate MIIM. This suggests that implementing specific-modality (separate MIIM) multi-dimensional information mining in the shallower layers of the network and shared-modality multi-dimensional information mining in the deeper layers is a more sensible design. According to Table S2, WRIM-Net (share) displays a Rank-1 decrease of 0.75% and a mAP decrease of 0.6% compared to the complete configuration. This implies that separate MIIM plays a role in the shallow layers to effectively mine specific-modality information.

E Ablation of fusion modes within the MIIM's internal SCR module

In neural networks, the common fusion mode for modules involves adding the original features to those processed by the module (such as our MIIM), termed the ADD mode. Therefore, we compare our designed sigmoid multiplication fusion method with the ADD mode. Additionally, we also conducted an ablation study on the multiplication without sigmoid mode, as depicted in Table S3. According to Table S3, replacing the sigmoid multiplication fusion mode with the ADD mode or adopting multiplication without sigmoid mode significantly reduces metrics. This demonstrates the effectiveness of our module fusion mode design.

F Ablation of Auxiliary-Information in AICL

The proposed AICL approach involves leveraging auxiliary information from the block 3 layer to enhance the mining of modality-invariant information. Here, we perform an ablation study on the auxiliary information by excluding the corresponding loss guidance for the $\boldsymbol{P_3}$ layer and disregarding the information from the $\boldsymbol{P_3}$ layer. The results, as shown in Table S4, indicate that without

Table S3: Ablation of fusion modes within the MIIM's internal SCR module on LLCM. $'W'$ indicates the use of sigmoid multiplication fusion, while $'W/O'$ denotes multiplication without sigmoid. $'W/O(ADD)'$ signifies the replacement of sigmoid multiplication fusion mode with ADD mode.

Sigmoid	Rank-1	mAP
W/O	56.08	62.98
$W/O(ADD)$	56.74	63.55
W	58.38	64.75

Table S4: Ablation of Auxiliary Information in AICL on LLCM. WRIM-Net (no Auxiliary Information) represents the model without the utilization of auxiliary information.

	Rank-1	mAP
WRIM-Net (no Auxiliary-Information)	57.32	63.43
WRIM-Net	**58.38**	**64.75**

using auxiliary information, there is a decrease of 1.06% and 1.32% in Rank-1 and mAP, respectively. This demonstrates the effectiveness of the auxiliary information design in the AICL approach.

References

1. He, L., Liao, X., Liu, W., Liu, X., Cheng, P., Mei, T.: Fastreid: A pytorch toolbox for general instance re-identification. arXiv preprint arXiv:2006.02631 (2020)