Supplementary Materials for
Modality Unifying Network for Visible-Infrared Person Re-Identification

Hao Yu[1], Xu Cheng[1]*, Wei Peng[2], Weihao Liu[3], Guoying Zhao[4]
[1]School of Computer Science, Nanjing University of Information Science and Technology, China
[2]Department of Psychiatry and Behavioral Sciences, Stanford University, USA
[3]School of Computer Science and Technology, Soochow University, China
[4]Center for Machine Vision and Signal Analysis, University of Oulu, Finland
{yuhao,xcheng}@nuist.edu.cn, wepeng@stanford.edu, whliu@stu.suda.edu.cn, guoying.zhao@oulu.fi

1. Applicability with Current SOTAs

In this section, we analyze how much improvement our auxiliary modality can bring to current SOTAs. To perform a fair experiment, all the original performances of the SOTAs are made consistent with their paper claims. We insert the auxiliary generator after the two-stream blocks and feed the generated auxiliary features (Aux.) together with the visible and infrared features to the weight-shared blocks for tri-modality learning. During the training, no additional constraints are used to help cross-modality learning except for their papers. The results are shown in Table A.

Table A. Performance evaluation of applying the proposed auxiliary modality (Aux.) on state-of-the-art VI-ReID methods. Experiments are conducted on SYSU-MM01 (all-search mode) and RegDB (visible to infrared mode) datasets.

Method	SYSU-MM01			RegDB		
	r=1	r=10	mAP	r=1	r=10	mAP
AGW [6]	47.50	84.39	47.65	70.05	86.21	66.37
+Aux.	54.29	88.31	53.16	75.82	89.03	70.44
↑	6.79	3.92	5.51	5.77	2.82	4.07
LBA [4]	55.41	–	54.14	74.17	–	67.64
+Aux.	57.17	–	55.29	74.98	–	67.80
↑	1.76	–	1.15	0.81	–	0.16
MPANet [5]	70.58	96.10	68.24	83.70	–	80.90
+Aux.	72.33	97.14	69.59	83.92	–	81.16
↑	1.75	1.04	1.35	0.22	–	0.26

As illustrated in Table A, the proposed auxiliary modality significantly improves the performance of all SOTAs, even without adding constraints to regulate the relationships among the three modalities. The results validate the superiority of our proposed tri-modality (visible-auxiliary-infrared) learning framework over their original two-modality (visible-infrared) framework in assisting the network in discovering identity-aware and modality-shared patterns from visible and infrared images.

2. Parameter Analysis

In this section, we discuss the impact of different hyperparameters in our MUN, including the spatial pyramid pooling ratios used in the cross-modality learner (CML) and the margin parameter α used in the identity alignment loss (L_{ia}). We do not discuss the value of the balance parameters γ, θ and σ as they are obtained from grid search.

Pooling ratios. We utilize spatial pyramid pooling (SSP) to enable the CML to learn modality-shared patterns from multiple feature scales. The reasonable spatial pooling ratios are discussed in Table B.

Table B. Discussion on pooling ratios in CML. N/A denote not using the SSP. N indicates the number of pooling layers.

N	Ratio	SYSU-MM01		RegDB	
		r=1	mAP	r=1	mAP
0	N/A	71.93	68.66	88.20	85.18
1	{2}	71.06	68.25	85.74	81.98
2	{2,4}	73.57	69.02	88.15	85.36
3	{2,4,6}	73.95	70.04	91.12	86.35
4	{2,4,6,8}	74.98	72.66	93.82	86.39
4	{2,4,6,12}	76.24	**73.81**	95.19	**87.15**
5	{2,4,6,8,16}	**76.25**	73.51	**95.28**	87.10
5	{2,4,6,12,18}	76.21	73.50	94.81	86.83

It is obvious that pleasing results are reached by using the ratio of {2,4,6,12} or {2,4,6,8,16}. To maintain a low computational complexity, we chose the ratio of {2,4,6,12} as the best setting in the paper, which utilizes 4 pooling layers and 4 transposed convolution layers to extract modality-shared patterns from four different feature scales.

Margin parameter. The margin parameter α is used in the identity alignment loss. It helps to regulate a larger feature distance between negative sample pairs and a smaller feature distance between positive sample pairs. As shown in Figure A, the best performance on both the SYSU-MM01 and RegDB datasets is obtained by setting $\alpha = 0.55$.

*Corresponding Author (Email: xcheng@nuist.edu.cn)

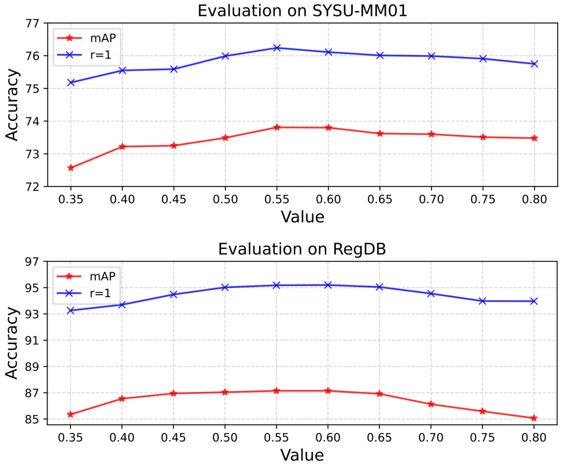

Figure A. Evaluation on different margin values.

3. Applying Prototype Scheme in Other Works

To avoid the inconsistency issue in learned feature relationships, we propose the prototype scheme, which dynamically calculates the modality prototypes based on the learned representations in each training iteration and then uses these prototypes to perform the cross-modality alignment. In this section, we discuss the impact of using our prototype scheme to upgrade current alignment-based VI ReID works (LBA [4] and MPANet [5]).

Table C. Evaluation on applying the proposed modality prototype scheme to other alignment-based methods.

Method	SYSU-MM01			RegDB		
	r=1	r=10	mAP	r=1	r=10	mAP
LBA [4]	55.41	–	54.14	75.17	–	67.64
+Prototype	56.19	–	55.51	76.75	–	67.98
↑	0.78	–	1.37	1.58	–	0.34
MPANet [5]	70.58	96.10	68.24	83.70	–	80.90
+Prototype	71.15	96.18	68.29	83.95	–	81.49
↑	0.57	0.08	0.05	0.25	–	0.59

As shown in Table C, the proposed prototype scheme can consistently improve the cross-modality matching performance under various testing scenarios, which shows its superiority in representing global modality-related information for performing robust modality alignment.

4. Designs in Auxiliary Generator

In this section, we discuss the impact on different designs and structures of the proposed auxiliary generator (IML+CML). We analyze how different configurations and architectures impact the overall performance.

Layer scale. The layer scale scheme is designed to control the modality-specific and modality-shared patterns learned in the auxiliary modality. In Table D we can notice that this scheme can bring consistent advantages as it enables our auxiliary modality to dynamically relieve the changing modality discrepancy in each training iteration.

Table D. Performance evaluation with or without using the layer scale scheme

Layer scale	SYSU-MM01			RegDB		
	r=1	r=10	mAP	r=1	r=10	mAP
×	75.15	97.77	72.52	93.82	97.6	85.92
✓	**76.24**	**97.84**	**73.81**	**95.19**	**98.93**	**87.15**

Series vs parallel. We discuss the reasonable scheme to form the auxiliary generator by combining the CML and IML. There are two available schemes, namely the series and the parallel schemes, as shown in Figure B.

For the series scheme, we use the IML to identify the modality-specific patterns in visible and infrared features, respectively. Then, the outcomes of two IMLs are fed into the CML for cross-modality pattern learning. For the parallel scheme, we simultaneously put the features of two modalities into the CML and IML and then fuse the outputs of two IMLs and one CML by using an additional 1 × 1 point-wise convolution layer.

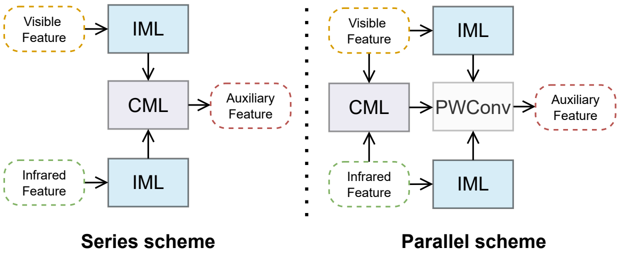

Figure B. Illustration of the series and parallel scheme to combine the proposed cross-modality learner (CML) and Intra-modality learner (IML) for constructing the auxiliary generator.

Table E. Evaluation on the series and parallel schemes to construct the auxiliary generator

Combination	SYSU-MM01			RegDB		
	r=1	r=10	mAP	r=1	r=10	mAP
Parallel	73.45	96.99	70.16	92.53	96.87	84.60
Series	**76.24**	**97.84**	**73.81**	**95.19**	**98.93**	**87.15**

Experimental results are shown in Table E, where the series scheme is selected in this paper due to its better performance. It enables our auxiliary generator to reveal the discriminative patterns in each modality and learn the modality-shared subset.

Pile up learners. Based on the series scheme, two IMLs and one CML are regarded as a basic unit to learn modality-related knowledge. Here, we discuss the impact of piling up multiple CMLs and IMLs to generate the auxiliary modality. Figure C gives the piling-up manner. To ensure the coherence of the input/output, we remove the 1 × 1 point-wise convolution layers in piled CMLs except for the last one.

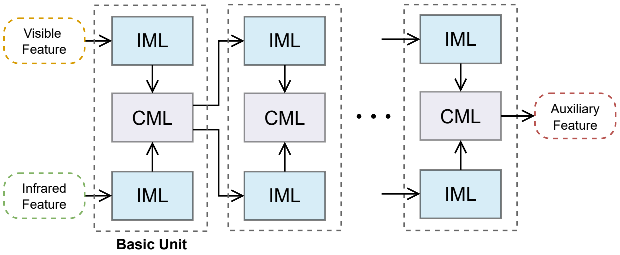

Figure C. Illustration on the manner of piling up multiple CMLs and IMLs.

The results are shown in Table F. Despite the fact that piling up learners can bring advantages in terms of Rank-1 accuracy, it may lead to overfitting issues that negatively impact the mean average precision (mAP) metric. Additionally, employing more learners to generate the auxiliary feature explicitly introduces additional floating-point operations (FLOPs) and parameters, but without yielding significant improvements in accuracy. Thus, we opt to utilize two IMLs and one CML to construct our auxiliary generator.

Table F. Evaluation on piling up multiple IMLs and CMLs to generate the auxiliary modality. Here N denotes the number.

N of IML	N of CML	SYSU-MM01		RegDB	
		r=1	mAP	r=1	mAP
2	1	76.24	**73.81**	95.19	87.15
4	2	76.29	73.77	95.22	**87.16**
6	3	76.85	73.71	**95.34**	87.10
8	4	**76.92**	73.18	95.33	87.04
10	5	76.87	73.19	95.14	86.53

micro designs. We discuss the impact of different micro designs in CML and IML, including the activation and normalization layers, integration manner of features, and inverted bottleneck ratios. Experimental results are shown in Table G, where the "MUN" indicates the performance of using the default settings reported in our paper.

Table G. Evaluation on different micro designs in IML and CML

micro design	SYSU-MM01		RegDB	
	r=1	mAP	r=1	mAP
MUN	76.24	**73.81**	**95.19**	**87.15**
Activation: ReLU → GeLU [3]	76.21	73.80	95.11	87.10
Activation: ReLU → SILU [2]	75.22	72.58	94.17	86.50
Normalization: BN → LN [1]	76.23	73.75	94.98	87.12
Integration: Concat → Add	72.87	68.95	91.77	84.68
Neck ratio: 4 → 8	**76.38**	73.80	95.18	87.06
Neck ratio: 4 → 2	75.54	72.09	93.82	85.51

In Table G, it is evident that certain popular micro designs, such as GeLU activation and layer normalization, do not yield substantial improvements. In order to strike a balance between computational cost and accuracy, we opt to retain the conventional CNN settings within the CML and IML architectures. This entails employing batch normalization to mitigate overfitting and utilizing the ReLU activation layer to enhance nonlinearity. Furthermore, we set the inverted bottleneck ratio to 4, ensuring a better trade-off between computational efficiency and performance.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[2] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018. 3

[3] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 3

[4] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12046–12055, 2021. 1, 2

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4330–4339, 2021. 1, 2

[6] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 1