Supplementary Materials of TOPLight: Lightweight Neural Networks with Task-Oriented Pretraining for Visible-Infrared Recognition

Hao Yu[1], Xu Cheng[1]*, Wei Peng[2]
[1]School of Computer Science, Nanjing University of Information Science and Technology, China
[2]Department of Psychiatry and Behavioral Sciences, Stanford University
{yuhao,xcheng}@nuist.edu.cn, wepeng@stanford.edu

1. Applicability with Deep Networks

In this section, we evaluate the effectiveness of the proposed TOP and FDR modules on conventional deep networks, including ResNet-50 [5], ConvNeXt-T [8], and Swin transformer [7], to demonstrate the applicability and generality of the proposed methods. We first pretrain the original one-path network on the ImageNet-1k dataset and then transform the network into a dual-path network by initializing two stem blocks. We then perform adaptive training on the ImageNet-mini dataset using the proposed task-oriented augmentation scheme, identity-consistency loss, and fake domain loss. The settings for applying TOP to deep networks are presented in Table 1, while Table 2 shows the corresponding experimental results.

It is evident that the proposed TOP strategy and FDR module significantly improve the VI-ReID performance of conventional deep neural networks. However, we also discovered that transformer-based networks are less affected by the sample variations created by task-oriented augmentation, resulting in less significant improvements in TOP compared to convolution-based networks. Furthermore, the results presented in Table 2 validate that our proposed methods are model-agnostic and can effectively enhance the VI-ReID performance of mainstream architectures.

Table 1. Settings of applying the proposed TOP for deep networks. The red and blue fonts denote the specific settings for the transformer and convolution-based networks, respectively.

(a): pretraining settings for deep networks		
Settings	On ImageNet-1K	On ImageNet-mini
Total epochs	200	70
Batch size	1024	36×2
Image size	224×224	224×224
Augmentations	Stanard [8]	**Task-oritend**
Optimizer	SGD / AdamW	SGD
Learning rate (Lr)	0.5 / 0.001	0.1 / 0.001
Lr decay	circle / cosine	cosine
Lr warm up	First 10 epoch	None
Weight decay	4E-5 / 5E-2	5E-5 / 1E-2
Emveriment	Pytorch-1.8, FP16	Pytorch-1.8
GPU	4 × Tesla V100-32G	RTX 3060-12G
Training cost	≈ 70h	≈ 10h
(b): Task-oriented augmentation settings		
Augmentations	Probability	Target
Crop & Filp	0.5	All path
Deform	0.2	
Colour jitter Channel shuffle RGB shift	Random pick	Block1-1 path
Compress	0.2	Block1-2 path
Defocus	0.2	
Sharpen gray	1	

2. Applicability with Latest SOTAs

This section investigates the impact of applying TOP to the current state-of-the-art VI-ReID methods. To ensure a fair experiment, we guarantee that all original performance metrics of the adopted SOTAs are consistent with their paper claims, and our ablation experiments are based on their official implementation codes. Specifically, we use the TOP strategy instead of the standard ImageNet-1k pretraining to prepare their dual-path feature extraction backbone for VI-ReID. The detailed settings can be found in Table 1. The experimental results are presented in Table 3.

It is clear that the TOP strategy consistently improves the performance of all the SOTAs without any increase in computational burden. This result validates the practicality and universality of our proposed TOP strategy for enhancing the performance of existing VI-ReID methods.

3. VI Adaptation on the Large Datasets

In the paper, a novel task-oriented pretraining strategy is proposed to improve the performance of lightweight net-

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

Table 2. Evaluation of applying the TOP strategy and FDR module to deep networks. Experiments are conducted on SYSU-MM01 (all-search mode) and RegDB (visible to infrared mode) datasets. Rank-1 (%) and mAP (%) are reported.

Method	SYSU-MM01		RegDB	
	r=1	mAP	r=1	mAP
ResNet-50 [5]	56.98	54.72	76.86	71.30
+ TOP	63.77	58.40	79.72	74.28
+ FDR	68.45	64.93	85.59	75.85
	↑ 11.47	↑ 10.21	↑ 8.73	↑ 4.55
ConvNeXt-Tiny [8]	58.72	55.31	78.25	72.64
+ TOP	64.15	61.82	80.29	74.06
+ FDR	69.53	65.70	86.87	81.49
	↑ 10.81	↑ 10.39	↑ 8.62	↑ 8.85
ViT-B/16 [4]	52.17	51.81	75.31	70.37
+ TOP	53.59	52.01	77.84	73.27
+ FDR	58.45	55.12	79.01	74.99
	↑ 6.28	↑ 3.31	↑ 3.70	↑ 4.62
Swin-Tiny [7]	58.24	55.16	78.39	72.68
+ TOP	61.08	58.45	79.25	74.40
+ FDR	67.59	61.73	83.80	77.91
	↑ 9.35	↑ 6.57	↑ 5.41	↑ 5.23

Table 3. Evaluation of applying TOP on the latest SOTAs.

Method	SYSU-MM01		RegDB	
	r=1	mAP	r=1	mAP
AGW [12]	47.50	47.65	70.05	66.37
+TOP	**62.45** ↑14.95	**60.09** ↑12.44	**81.96** ↑11.91	**75.08** ↑8.71
CMT [1]	71.88	68.57	91.97	84.46
+TOP	**73.45** ↑1.57	**69.93** ↑1.36	**93.58** ↑1.61	**85.59** ↑1.13
SMCL [2]	67.39	61.78	83.05	78.57
+TOP	**71.45** ↑4.06	**64.32** ↑2.54	**88.57** ↑5.52	**82.05** ↑3.48
MPANet [3]	70.58	68.24	82.80	80.70
+TOP	**74.25** ↑3.67	**71.13** ↑2.89	**89.62** ↑6.82	**83.54** ↑2.84

works for VI recognition, where the critical step is the VI scene adaptation training on the few-shot ImageNet-mini dataset [10]. We also wish to investigate the influence of performing VI adaptation training on larger ImageNet-like datasets. Specifically, for the basic settings to construct an ImageNet-like dataset, we denote the number of identities as N and fix the number of samples within each identity to 500 + N. Then, we gradually increase the number of identities (N) to construct different subsets of the ImageNet-1K dataset with various sizes, which we refer to as ImageNet-X. All the identities and image samples are randomly selected from the ImageNet-1k dataset using a random seed of 42 in the $Numpy$. The MobileNetV3-L [6] pretrained on ImageNet-1k is used as the baseline network. Table 4 displays the experimental results.

It can be noticed that VI adaptation training carried out on larger datasets may help to increase the pretraining accuracy on the ImageNet-X datasets we constructed, but it doesn't bring significant improvement for the VI-ReID task. The VI-ReID performance even begins to degrade when N exceeds 500. Two factors may be responsible:

(1): We employ a task-oriented augmentation strategy to simulate the visual differences in VI scenes, which can also be viewed as introducing additional noise during the training. Training with numerous noised image samples may mislead the network to concentrate on learning the noise representation, thereby preventing it from learning identity-aware patterns among the heterogeneous features.

(2): When performing VI adaptation training with fewer training samples, the specific generalization capacity for few-shot learning can be enhanced, and the two currently available VI-ReID datasets (SYSU-MM01 [11] and RegDB [9]) are both few-shot datasets, containing fewer than 50,000 image samples each.

Thus, we perform the VI adaptation training on the ImageNet-mini dataset, which has 100 identities (N) and 600 images per identity. This solution can bring advantages in terms of performance and training speed.

Table 4. Performance evaluation of performing TOP on different subsets of ImageNet-1k datasets. Metrics including TOP-1 (%), ToP-5 (%), Rank-1 (%), mAP (%) accuracy are reported.

Setting	On ImageNet-X		On SUSY-MM01	
N	Top-1 (%)	Top-5 (%)	r=1 (%)	mAP (%)
100	80.92±0.28	85.57±0.15	59.75±0.13	54.73±0.09
200	80.99±0.25	86.06±0.13	59.76±0.12	54.71±0.09
300	81.34±0.24	86.17±0.11	59.79±0.19	54.83±0.11
400	81.52±0.26	86.20±0.15	59.77±0.21	54.74±0.10
500	81.67±0.21	86.25±0.14	58.92±0.33	54.55±0.15
600	81.69±0.17	86.23±0.11	57.29±0.35	54.38±0.17

4. Motivations of Augmentations

There are three major motivations that inspired us to design the colour and texture augmentations.

Figure 1. Visualizations. The colour and texture augmentations are utilized to simulate the modality discrepancy in VI scenes.

(1) To enable the network to learn prior knowledge related to cross-modal matching in the pretraining stage, we jointly use colour and texture augmentations to simulate the difference between infrared and visible images in ImageNet samples, as shown in Fig. 1.

(2) The colour augmentations (e.g., random RGB shift) can disturb the strong colour prior knowledge learned from the ImageNet dataset. We hope the network pays more attention to the structural patterns of visible images (which also exist in infrared images), not just colour patterns.

(3) The texture augmentations (e.g., random defocusing) are also used to simulate the terrible imaging quality (e.g., defocus) under night surveillance conditions.

5. Details about the Gradient Reversal Layer

In this paper, we use the gradient reversal layer (GRL) to create an adversarial training process that forces the network to discover shared patterns from heterogeneous images. The GRL can be defined as:

$$GRL_\lambda(\mathbf{x}) = \mathbf{x}, \quad \frac{d(GRL_\lambda)}{d\mathbf{x}} = -\lambda\mathbf{I}. \tag{1}$$

It establishes opposite optimization targets before and after the current layer by reversing the gradient when performing backpropagation. λ is used to control the reversal ratio.

References

[1] Jiang , $et.al$. Cross-modality transformer for visible-infrared person re-identification. ECCV, 2022. 2

[2] Ling , $et.al$. A multi-constraint similarity learning with adaptive weighting for visible-thermal person re-identification. IJCAI, 2021. 2

[3] Wu , $et.al$. Discover cross-modality nuances for visible-infrared person re-identification. CVPR, 2021. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[6] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 2

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1, 2

[8] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1, 2

[9] Dat Tien Nguyen, Hyung Gil Hong, Ki Wan Kim, and Kang Ryoung Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3):605, 2017. 2

[10] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 2

[11] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017. 2

[12] 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. 2