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

2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-0129-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPR52729.2023.00345

Abstract

Visible-infrared recognition (VI recognition) is a challenging task due to the enormous visual difference across heterogeneous images. Most existing works achieve promising results by transfer learning, such as pretraining on the ImageNet, based on advanced neural architectures like ResNet and ViT. However, such methods ignore the negative influence of the pretrained colour prior knowledge, as well as their heavy computational burden makes them hard to deploy in actual scenarios with limited resources. In this paper, we propose a novel task-oriented pretrained lightweight neural network (TOPLight) for VI recognition. Specifically, the TOPLight method simulates the domain conflict and sample variations with the proposed fake domain loss in the pretraining stage, which guides the network to learn how to handle those difficulties, such that a more general modality-shared feature representation is learned for the heterogeneous images. Moreover, an effective fine-grained dependency reconstruction module (FDR) is developed to discover substantial pattern dependencies shared in two modalities. Extensive experiments on VI person re-identification and VI face recognition datasets demonstrate the superiority of the proposed TOPLight, which significantly outperforms the current state of the arts while demanding fewer computational resources.

1. Introduction

Identity recognition technologies have provided numerous reliable solutions for monitoring systems, which strive to match the face (face recognition [6, 7]) or pedestrian (person re-identification [42]) images of the same identity. However, the majority of previous efforts only consider visible images. In real-life practice, many surveillance cameras can switch to infrared imaging mode at night. Thus, the essential cross-modality visible-infrared recognition (VI recognition) technology has been developed to match the visible and infrared photographs of the same people.

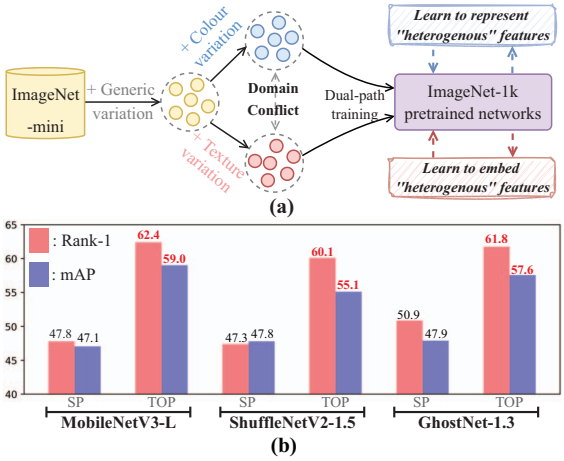

Figure 1. (a) The task-oriented pretraining strategy; (b) Performance comparison of the standard ImageNet-1k pretraining scheme (SP) and the proposed task-oriented pretraining scheme (TOP) on the SYSU-MM01 dataset [35] (All-Search mode).

Recently, visible-infrared person re-identification (VI-ReID) [26, 38] and visible-infrared face recognition [8, 13, 44] have been widely studied. The key issue is identifying the modality-shared patterns. To this end, several works [29, 30] use generative adversarial networks (GANs) to implement cross-modality alignment at the pixel and feature levels. Others [4, 26, 38] design the dual-path feature extraction network, coupled with inter-feature constraints, to close the embedding space of two modalities. However, these methods utilize at least one pretrained ResNet-50 [12] backbone to extract solid features, which makes them unsuitable for edge monitoring devices. Recent works [4, 38] employ auxiliary models (e.g., pose estimation, graph reasoning) to relieve the modality discrepancy, which enhances the performance on academic benchmarks but reduces the real-time inference speed. Compared with conventional deep networks (e.g., ResNet, ViT), lightweight networks [11, 15, 24] can extract basal features rapidly. In VI recogni-

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

979-8-3503-0129-8/23/$31.00 ©2023 IEEE
DOI 10.1109/CVPR52729.2023.00345

tion tasks, however, the vast modality discrepancy renders the performance of lightweight networks significantly inferior to that of conventional deep networks. The main reason is that lightweight networks lack the ability to identify modality-shared patterns from heterogeneous images.

To address this issue, we present an effective task-oriented pretraining (TOP) strategy. As shown in Fig. 1(a), we first train a lightweight network on the ImageNet-1k dataset to learn vision prior knowledge. After that, the trained network is transformed into the dual-path network and further trained by using task-oriented data augmentations, identity consistency loss and fake domain loss on the ImageNet-mini dataset [18]. The task-oriented pretraining (TOP) strategy simulates the sample differences in VI scenes and teaches the network how to represent and embed discrepant features. Fig. 1(b) reports the performance of three lightweight networks in the VI-ReID task. Compared with the ImageNet-1k pretraining, our TOP strategy can remarkably improve the baseline performance.

Another weakness of lightweight networks is that few feature maps are learned from raw images for rapid inference. In the VI recognition scene, it is challenging to discover modality-shared patterns with so few learned feature maps. In practice, the network can focus on a group of aggregated modality-specific patterns that offer the most gradient for identity classification. In contrast, the fine-grained and modality-shared patterns, which are crucial for achieving robust cross-modality matching, are neglected.

Based on the above observations, we present a novel fine-grained dependency reconstruction (FDR) module to help lightweight networks learn modality-shared and fine-grained patterns. Specifically, inspired by the horizontal slice scheme [1], we first slice feature maps horizontally and vertically to extract fine-grained patterns from diversified local regions. Then, the original spatial dependencies of these patterns are eliminated by using pooling operations. Further, the cross-modality dependencies are built by using up-sampling layers to amplify the modality-shared parts from these patterns. At last, to avoid overfitting, the shuffle attention is designed to re-weight the channel dependencies of all the feature maps, which spreads attention to local patterns as much as possible. In general, the major contributions of this paper can be summarized as follows.

- We propose an effective task-oriented pretrained lightweight neural network (TOPLight) for VI recognition. To the best of our knowledge, it is the first work to develop a paradigm for VI recognition on edge devices with an extremely low computation budget.
- An effective task-oriented pretraining strategy is proposed to enhance the heterogeneous feature learning capacity of lightweight networks with task-oriented augmentations and the proposed fake domain loss.
- A fine-grained dependency reconstruction module is designed to mine cross-modality dependencies.
- Extensive experiments demonstrate that the proposed method outperforms the state-of-the-art methods on mainstream VI-ReID and VI face recognition datasets by a remarkable margin and extremely low complexity.

2. Related Work

Visible-Infrared Face Recognition (VI-FR). It aims to match face images across visible and infrared modalities. Initially, handcraft features are designed for VI-FR. In [3,9], authors utilize the Local Radon Binary Pattern (LRBP) as the general representation to perform identity matching. Difference-of-Gaussian (DoG) filtering and Multi-scale Block Local Binary Patterns (MB-LBP) are widely adopted [20, 50] to acquire robust representations for VI-FR. Recently, deep learning methods have achieved considerable success in VI-FR, focusing on learning modality-shared features with CNN networks [13,14,22] or disentangling the modality-invariant representations with GAN networks [46]. In [37], a dual-generation method is proposed to disentangle modality-invariant patterns from visible and infrared images, which helps relieve the modality discrepancy. After that, Duan *et al.* [8] proposed a pose-aligned cross-spectral hallucination (PACH) network to eliminate identity-independent patterns at multiple stages.

Visible-Infrared Person Re-identification (VI-ReID). It aims to match visible and infrared pedestrian images according to the feature similarity ranking [47]. The zero-padding scheme [35] was first proposed to handle VI-ReID. In the meantime, they established the first large-scale VI-ReID dataset, named SYSU-MM01. Afterwards, a bi-directional center-constrained network [39] was presented to simultaneously optimize intra- and inter-modality discrepancies. To help the network learn modality-invariant patterns, X-modality [19] is introduced to bridge the gap between visible and infrared modality. Recently, a dense alignment learning method [26] was presented to establish the cross-modality feature correspondence at the pixel level. Ye *et al.* [38] introduced an attentive graph to discover tri-level relations between visible and infrared images. In addition, several GAN-based methods implement cross-modality alignment [29, 30] or modality compensation [31, 45] to help the network learn robust and compact knowledge. However, despite the promising results, these works are too complex to be utilized in the real world.

3. Methodology

3.1. Overview of The Proposed Method

The major pipeline of the proposed TOPLight is illustrated in Fig. 2, which can be summarized as three steps.

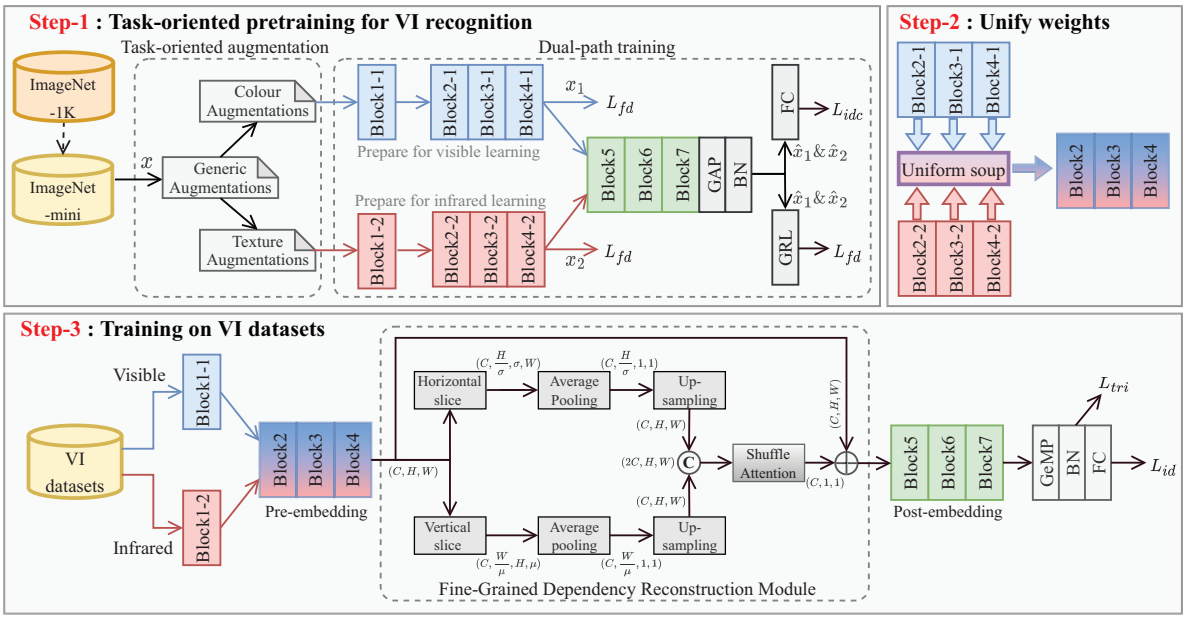

Figure 2. The overall pipeline of our method. We take three steps to adapt a lightweight network to the VI recognition task.

Step-1. To adapt the lightweight networks to VI recognition scenarios, we construct a dual-path network based on the lightweight model trained on ImageNet-1k (similar in [26]) and further train the dual-path network on ImageNet-mini [18] with task-oriented augmentations. The identity consistency loss (L_{idc}) and fake domain loss (L_{fd}) are designed to supervise the domain conflict learning process.

Step-2. We utilize the uniform soup scheme [34] to unify two pre-embedding block groups. This step aims to tune the structure of the pretrained network suited for VI training, which preserves two separate blocks (Block1-1 and Block1-2) to extract low-level features from two modalities.

Step-3. In this stage, the pretrained dual-path network is further trained on VI recognition datasets. A novel fine-grained dependency reconstruction (FDR) module is integrated into the network, which can construct fine-grained pattern dependencies between visible and infrared features.

3.2. Task-oriented Pretraining Strategy

The core motivation of the task-oriented pretraining strategy is to enable lightweight networks to learn prior knowledge related to cross-modality matching during the pretraining stage. Specifically, we first transform the lightweight network trained on the ImageNet-1k dataset into the dual-path network, which consists of two stem blocks, two pre-embedding block groups and one post-embedding block group. Then, the dual-path network is further trained on the ImageNet-mini dataset. During this training, task-oriented data augmentation (DA) is utilized to create visual differences in training samples, which intends to simulate the discrepancy between visible and infrared

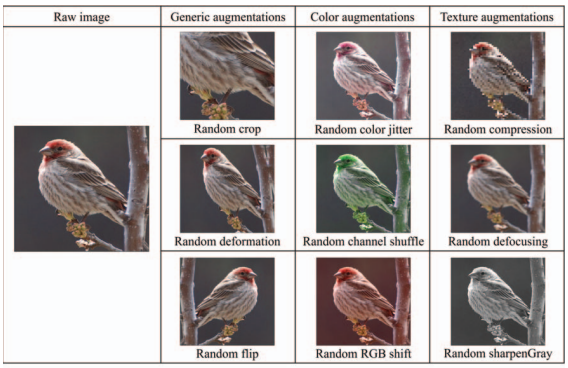

Figure 3. The effects of each data augmentation.

images and enforce the network to pay more attention to the identity-aware but colour-irrelevant patterns. Thus, we design the **generic**, **colour** and **texture** data augmentation (DAs) to achieve the training process, as shown in Fig. 3.

For the generic DAs, we increase the overall sample diversity by using random crop, flip and deformation. In the colour DAs, we randomly perform colour jitter, channel shuffle and RGB shift for each sample in the Block1-1 path. The colour DAs are utilized to disturb the strong colour prior knowledge learned from the ImageNet-1k dataset, forcing the network to pay more attention to the structural patterns of visible images (which also exist in infrared images), not just colour patterns. In the texture DAs, we randomly perform compression, defocusing and sharpenGray for each sample in the Block1-2 path, which intends to remove the colour information and simulate the terrible imaging quality under night surveillance conditions. During the TOP, we combine these three DAs to create visual differences that are close to the VI recognition scenes.

However, in the above manner, the network can lazily learn from one path to avoid embedding visually discrepant samples. Meanwhile, the sample differences generated by augmentations are insufficient to simulate the serious domain conflict in VI recognition scenes. Thus, we develop a fake domain loss to aggravate the variation of mid-level features during the training and motivate the network to discover shared patterns among them.

Assume $\mathbf{x}$ denotes a raw sample from the ImageNet-mini dataset; after augmentations and feature learning, the $\mathbf{x}_1$ and $\mathbf{x}_2$ denote the outcome features of Block4-1 and Block4-2, respectively. $\mathbf{y}$ is the shared ground-truth label. We concatenate $\mathbf{x}_1$ and $\mathbf{x}_2$ and feed them into Block5. The output of Block7 is denoted as $\hat{\mathbf{x}} = \{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2\}$. Based on the above definitions, the fake domain loss can be defined by:

$$L_{fd}^1 = L_d(\mathbf{x}_1, \mathbf{d}_1) + L_d(GRL(\hat{\mathbf{x}}_1), \mathbf{d}_1), \tag{1}$$

$$L_{fd}^2 = L_d(\mathbf{x}_2, \mathbf{d}_2) + L_d(GRL(\hat{\mathbf{x}}_2), \mathbf{d}_2), \tag{2}$$

where $\mathbf{d}_1$ and $\mathbf{d}_2$ are the corresponding pseudo domain labels for $\mathbf{x}_1$ and $\mathbf{x}_2$. We use the domain labels to pretend that $\mathbf{x}_1$ is from one domain and $\mathbf{x}_2$ is from the other. L_d denotes the domain classification loss implemented with the logistic classifier. GRL is the gradient reversal layer proposed in [10], which aims to reverse the optimization target of the final domain classifier. In this manner, L_{fd}^1 and L_{fd}^2 generate two contradictory learning procedures: self-domain constraints are individually set on $\mathbf{x}_1$ and $\mathbf{x}_2$, which enforce them to be representative and specific for the two fake domains we pretended, respectively. The inverted-domain constraints are set on the final features after Block7 ($\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$) by reversing gradients, which urges them to be similar so as to confound the domain classification loss L_d.

During this procedure, Block(2,3,4)-1 and Block(2,3,4)-2 are trained to extract two types of strongly distinguished "heterogeneous features". In comparison, Block(5,6,7) are trained to embed these discrepant features and learn their shared patterns. We manually create this conflict to force the lightweight network to learn how to represent heterogeneous features and discover shared patterns from them.

To ensure that the embedded features are identity-related, the identity consistency loss (L_{idc}) is introduced as:

$$L_{idc} = log[P(y|p(\hat{\mathbf{x}}_1))] + log[P(y|p(\hat{\mathbf{x}}_2))], \tag{3}$$

where $p(\hat{\mathbf{x}}_1)$ and $p(\hat{\mathbf{x}}_2)$ denote the identity prediction results from the same linear classifier. $\mathbf{y}$ is the ground-truth label for both $\mathbf{x}_1$ and $\mathbf{x}_2$. Here, we utilize one classifier to predict two conflicted features with shared logits, aiming to further improve the embedding capacity of Block(5,6,7).

The overall loss in the TOP is written as follows.

$$L = L_{idc} + L_{fd}^1 + L_{fd}^2. \tag{4}$$

After the pretraining, we utilize the uniform soup scheme to unify two groups of pre-embedded blocks in turn (Step-2 in Fig. 2). The final lightweight network for VI-recognition training consists of two independent stem blocks (Block1-1 and 1-2), one pre-embedding block group (Block2,3,4) and one post-embedding block group (Block5,6,7), as shown in the Step-3 of Fig. 2.

3.3. Fine-grained Dependency Reconstruction

After the task-oriented pretraining, the lightweight network learns how to extract modality-invariant features from heterogeneous images. However, it lacks the ability to discover fine-grained patterns across the modalities. Therefore, we introduce a practical fine-grained dependency reconstruction module to help the lightweight network discover plentiful fine-grained patterns by constructing cross-modality feature dependencies in the VI training phase.

Specifically, let us denote $S^v = \{\mathbf{x}_i^v | i = 1, 2, ..., N^v\}$ and $S^r = \{\mathbf{x}_i^r | i = 1, 2, ..., N^r\}$ as the visible and infrared training sets, respectively; N^v and N^r indicate the total numbers of visible and infrared samples, respectively. During the training, we feed $\mathbf{x}^v$ and $\mathbf{x}^r$ into Block1-1 and Block1-2 to extract low-level features of two modalities, respectively. Then, we concatenate these low-level features and feed them into the Block(2,3,4) for the initial feature embedding. Based on the outcome of Block4 ($\mathbf{Z} = \{\mathbf{Z}^v, \mathbf{Z}^r\}$), the fine-grained and cross-modality feature dependencies are constructed to help the network learn substantial modality-shared patterns by the FDR module.

In order to extract fine-grained patterns from feature maps, we slice $\mathbf{Z}$ into parts along the horizontal and vertical directions. For each spatial map $\{\mathbf{Z}\}_{m=0}^{C-1} \in \mathbb{R}^{H \times W}$, the horizontal patterns are obtained by:

$$\mathbf{h}^f = Concat[\underset{\sigma \times W}{Window}(\{\mathbf{Z}^f\}_{m=0}^{C-1})], f \in \{v, r\}, \tag{5}$$

where $\mathbf{Z}^f \in \mathbb{R}^{C \times H \times W}, f \in \{v, r\}$ denote visible or infrared features. C, H, W are the channel, height, and width dimensions, respectively. $\mathbf{h}^f \in \mathbb{R}^{C \times \frac{H}{\sigma} \times \sigma \times W}$ denotes a group of fine-grained horizontal patterns extracted from $\mathbf{Z}^f$. Similarly, the vertical patterns $\mathbf{v}^f \in \mathbb{R}^{C \times \frac{W}{\mu} \times H \times \mu}$ can also be obtained by:

$$\mathbf{v}^f = Concat[\underset{H \times \mu}{Window}(\{\mathbf{Z}^f\}_{m=0}^{C-1})], f \in \{v, r\}, \tag{6}$$

where μ and σ are used to control the slice granularity.

Further, the spatial information of $\mathbf{h}^f$ and $\mathbf{v}^f$ is condensed by the average pooling operation, which produces $\hat{\mathbf{h}}^f \in \mathbb{R}^{C \times \frac{H}{\sigma}}$ and $\hat{\mathbf{v}}^f \in \mathbb{R}^{C \times \frac{W}{\mu}}$. In the above manner, we refine the original spatial maps $\{\mathbf{Z}^f\}_{m=0}^{C-1} \in \mathbb{R}^{H \times W}$ into directional tensors $\hat{\mathbf{h}}^f$ and $\hat{\mathbf{v}}^f$ to cut off the original self-modality dependencies in spatial dimensions. Furthermore,

the cross-modality dependencies are established as follows.

$$\mathbf{Z}_h^f = \underset{(\frac{H}{\sigma}\times W)}{TransConv}(\hat{\mathbf{h}}^f), f \in \{v, r\}, \tag{7}$$

$$\mathbf{Z}_v^f = \underset{(H\times\frac{W}{\mu})}{TransConv}(\hat{\mathbf{v}}^f), f \in \{v, r\}. \tag{8}$$

In Eq. (7) and Eq. (8), we use two modality-shared transposed convolution layers to reconstruct the spatial maps of $\mathbf{h}^f$ and $\mathbf{v}^f$, respectively. This process aims to establish the cross-modality spatial dependencies. The outputs are $\mathbf{z}_h^f \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{z}_v^f \in \mathbb{R}^{C\times H\times W}$.

However, there exist "scheme stereotypes" in $\mathbf{Z}_h^f$ and $\mathbf{Z}_v^f$. The main reason is that these features are learned from two distinct directional schemes, which may lead to a skewed learning procedure. In contrast, we hope these two schemes can consistently provide diversified patterns that cover more valuable signals. Thus, the shuffle attention is designed to fuse $\mathbf{Z}_h^f$ and $\mathbf{Z}_v^f$ while decoupling them from two directions. Specifically, we concatenate them on the channel level and then shuffle the channel order to disarrange the spatial maps learned from two directions.

$$\mathbf{D}^f = Shuffle(Concat(\mathbf{Z}_h^f, \mathbf{Z}_v^f)), f \in \{v, r\}, \tag{9}$$

where $\mathbf{D}^f \in \mathbb{R}^{2C\times H\times W}$ denotes the concatenated feature with disarranged channel order. The new channel dependencies are established by weighting all the spatial maps.

$$\hat{\mathbf{D}}^f = \mathbf{W}_2 ReLU(\mathbf{W}_1 GAP(\mathbf{D}^f)), f \in \{v, r\}, \tag{10}$$

where $\mathbf{W}_1 \in \mathbb{R}^{2C\times\frac{C}{4}}$ and $\mathbf{W}_2 \in \mathbb{R}^{\frac{C}{4}\times C}$ indicate projection matrices to compute the channel attention [16]. Finally, we introduce $\hat{\mathbf{D}}^f$ to refine the original features.

$$\hat{\mathbf{Z}}^f = \hat{\mathbf{D}}^f \oplus \mathbf{Z}^f, f \in \{v, r\}, \tag{11}$$

where $\hat{\mathbf{Z}}^f$ is used as the input of the next block; $\oplus$ is the element-wise add operation. The proposed fine-grained dependency reconstruction module can effectively establish fine-grained and cross-modality dependencies in both visible and infrared features, which can help discover substantial small patterns that are vital in VI recognition.

3.4. Loss Function

We utilize the hard-mining triplet loss (L_{tri}) [42] and cross-entropy loss (L_{id}) for metric learning and identity learning, respectively. The overall loss can be defined as:

$$L = L_{id} + \gamma * L_{tri}, \tag{12}$$

where γ is the parameter to balance each loss term.

4. Experimental Results

4.1. Datasets and Settings for VI-ReID

We first evaluate the proposed method for the visible-infrared person re-identification (VI-ReID) task on two public benchmarks: SYSU-MM01 [35] and RegDB [25].

SYSU-MM01 [35] is the largest VI-ReID dataset, which contains 491 persons captured from six different cameras across scenes (indoor and outdoor). There are a total of 287628 visible images and 15792 infrared images. Following the evaluation protocol in [26, 38], 395 persons are fixed for training and 96 persons for testing. The testing process includes all-search and indoor-search modes.

RegDB [25] contains 412 persons, and each person has 10 visible and 10 infrared images. "Visible to infrared" and "infrared to visible" are two testing modes. The former uses visible as the query and infrared as the gallery. The latter does the opposite. Following [26, 30, 38], the final performance on the RegDB is made with the average results of 10 times the training and testing procedure. Each time, we randomly sample 206 identities for training, and the remaining 206 identities are used for testing.

Evaluation Metrics. We use the standard cumulative matching characteristics (CMC), mean average precision (mAP), and mean inverse negative penalty (mINP) [42] as the evaluation metrics.

Implementation Details. For the ablation studies, we select the MobileNetV3-L [15] as the lightweight baseline to validate each proposed component. As shown in Table 1, we encapsulate all the layers of MobileNetV3-L as Block1-7 and then initialize the first four blocks twice to construct the dual-path network. During the TOP, detailed settings are

Table 1. Detailed structures for each Block. We package the entire MobileNetV3-Large into Block1-7 without overlap.

Block partitions on MobileNetV3-large		
Layer Name	Structures	Output Size
Block1	Conv(3×3, 2), Bneck(3×3, 1)	16×112^2
Block2	Bneck(3×3, 2)	24×56^2
Block3	Bneck(3×3, 1)	24×56^2
Block4	Bneck(5×5, 2), Bneck(5×5, 1)×2	40×28^2
Block5	Bneck(3×3, 2), Bneck(3×3, 1)×5	112×14^2
Block6	Bneck(5×5, 2), Bneck(5×5, 1)×2	160×7^2
Block7	Conv(1×1, 1)×3	1280×7^2

shown in Table 2(b). It is worth noticing that the "Random pick" means that for each image sample used in the Block1-1 path, we randomly pick one method from colour jitter, channel shuffle, or RGB shift to augment it.

During the VI-ReID training, detailed settings are presented in Table 2(a). Briefly, we randomly sample 6 identities with 6 visible and 6 infrared images per identity, which makes a total of 72 images in each batch. All images are resized to 288×144. Random erasing (RE) [49], flipping,

Table 2. Detailed training and augmentation settings.

(a): **Training settings**

Settings	On ImageNet-1K	On ImageNet-mini	On VI-ReID
Epochs	150	50	70
Batch size	512	36×2	36×2
Image size	224×224	224×224	288×144
Augs.	Stanard [15]	**Task-oriented**	RE , CE, Flip
Optimizer	SGD	SGD	AdamW [23]
Lr.	0.2	0.1	0.001
Decay	cosine	cosine	constant
Warm up	First 5 epoch	None	First 5 epoch
Weight decay	5E-4	5E-5	1E-2
Platform	Pytorch, FP16	Pytorch	Pytorch
GPU	Tesla V100	RTX 3060	RTX 3060
Cost	57 h	4 h	–

(b): **Task-oriented augmentation settings**

Augmentations		Probability	Target
Generic (G)	Crop & Filp Deformation	0.5 0.2	All path
Colour (C)	Colour jitter Channel shuffle RGB shift	Random pick	Block1-1 path
Texture (T)	Compression Defocusing Sharpen gray	0.2 0.2 1	Block1-2 path

and channel erasing (CE) [40] are utilized against overfitting. The AdamW [23] optimizer with a learning rate (Lr) of 0.001, and a weight decay of 0.01 is adopted to train our network for 70 epochs. We use the Lr warm-up in the first 5 epochs and constant decay at the 20^{th} and 40^{th} epochs with a decay factor of 0.1. The parameters μ and σ are set to 3 and 2 based on the fine-tuning results, respectively. The balance parameter γ in Eq. (12) is empirically set to $1/2$.

4.2. Ablation Study

Effect of Task-oriented Pretraining. We first evaluate the effect of task-oriented pretraining on VI-ReID, as shown in Table 3. No.1 indicates the baseline performance obtained via direct fine-tuning of the ImageNet-1k pretrained model in VI-ReID datasets. When only using generic augmentations and identity loss to perform TOP on ImageNet-mini, the "catastrophic forgetting" makes No.2 perform worse than the baseline. However, when applying the proposed task-oriented augmentations, the network learns to extract stable features from visually discrepant images in advance, thus improving the downstream VI-ReID performance (No.3). Coupled with L_{idc} and L_{fd}, we create a conflicting learning process that enforces the network to learn how to represent and embed heterogeneous features. No.5 shows significant improvements in both the SYSU-MM01 and RegDB datasets. By only using the MobileNetV3-large backbone, the performance reported in No.5 has already exceeded many ResNet-based works [19, 26, 38].

Effect of Fine-grained Dependency Reconstruction Module. In Table 3, No.6-No.9 show the ablation experiments about the fine-grained dependency reconstruction module. We intend to construct the cross-modality feature dependencies, and the first issue is how to model the original feature relations. No.6 indicates directly applying the coarse-grained features to the shuffle attention module. The performance is slightly improved. When using the proposed directional scheme (H_s, V_s) to extract fine-grained patterns and reconstruct their spatial relation using transposed convolution layers, No.7 and No.8 show meaningful improvements. The full power of the FDR module is shown in No.9, which adopts both fine-grained horizontal and vertical patterns to discover cross-modality feature similarities, achieving 66.14% and 63.80% in terms of Rank-1 and mAP performance on SYSU-MM01 under the all-search mode.

Pointing at the difficulties in VI recognition tasks, we design the TOP strategy and FDR module to remedy the drawbacks of the lightweight network. Compared with the baseline (No.1), our method (No.9) considerably improves all metrics on two benchmarks. Meanwhile, the ablations reported in Table 3 demonstrate that all the proposed components bring advantages consistently to the accuracy gain.

Applicability with Lightweight Networks. We evaluate our methods on three mainstream lightweight networks (ShuffleNetV2 [24], GhostNet [11] and MobileNetV3 [15]). As shown in Table 4, our method can remarkably improve VI Re-ID performance with ignorable complexity growth. Compared with conventional deep networks, we surpass them significantly with minimal complexity.

Detailed Ablations on the FDR Module. We validate the rationality of each component in the FDR module. Table 5(a) shows the impacts of different spatial modelling methods. Except for ours, the HAP [38] achieves the best performance. However, we extract patterns from both the horizontal and vertical regions and use up-sampling (u.) to discover modality-shared cues from re-enlarged spatial areas, which can be regarded as a more effective solution.

Table 5(b) shows the impacts of different channel relation reasoning methods. Our shuffle attention (SA) wins both SE [16] and CBAM [33]. The recipe is the channel shuffle (cs.) operation that builds scheme-decoupled dependencies. This crucial step can evenly correlate visible and infrared features from each fine-grained pattern.

4.3. Feature Visualization

What Does TOP Do? We randomly track 10 identities during the TOP stage and visualize the feature distribution via T-SNE [28], as illustrated in Fig. 4.

We first use task-oriented augmentation to create two types of visual discrepant samples from each image, as shown in Fig. 4(a) and Fig. 4(b). Then, the proposed fake domain loss forces the features after Block4 to be domain-specific, as shown in Fig. 4(c). Concurrently, it forces the network to learn to embed those domain-specific features and provide domain-shared predictions. In Fig. 4(d), after the training, we can observe that the network knows how to extract the domain-shared representations and successfully

Table 3. Evaluation of each proposed component on two VI-ReID datasets. "Augs." indicates the augmentations. G, C and T denote the generic, colour, and texture augmentations, respectively. In the FDR module, H_s and V_s denote the horizontal and vertical slices with up-sampling. SA is the shuffle attention module. Rank (r) (%), mAP (%) and mINP (%) are reported.

No.	Task-oriented pretraining stage					VI training stage			**SYSU-MM01** (all-search)				**RegDB** (visible-to-infrared)			
	Augs.		Loss functions			FDR module										
	G	C+T	L_{id}	L_{idc}	L_{fd}	H_s	V_s	SA	r=1	r=10	mAP	mINP	r=1	r=10	mAP	mINP
1									47.81	89.71	47.06	33.48	71.26	89.94	65.66	48.50
2	✓		✓						43.28	85.96	45.56	31.10	70.73	88.52	65.65	48.41
3	✓	✓	✓						49.85	89.74	47.56	35.52	71.32	89.91	65.67	48.49
4	✓	✓		✓					54.28	92.11	52.94	41.29	75.31	92.64	68.78	52.16
5	✓	✓		✓	✓				62.41	94.12	59.06	45.13	82.75	94.13	76.21	61.84
6	✓	✓		✓	✓			✓	62.89	94.26	59.79	45.84	82.88	94.19	76.24	61.92
7	✓	✓		✓	✓	✓		✓	63.95	95.28	60.09	46.80	83.07	94.48	76.55	62.20
8	✓	✓		✓	✓		✓	✓	64.04	95.41	61.12	46.92	83.22	94.69	77.01	63.16
9	✓	✓		✓	✓	✓	✓	✓	**66.14**	**96.03**	**63.80**	**49.76**	**84.15**	**94.98**	**79.26**	**63.86**

Table 4. Experimental results on different lightweight networks and conventional deep networks.

	Methods	FLOPs (M)	SYSU-MM01 r=1	SYSU-MM01 mAP	RegDB r=1	RegDB mAP
Convention	ResNet-50	3562	56.98	54.72	76.86	71.30
	ConvNeXt-Tiny	3620	58.72	55.31	78.25	72.64
	Vit-B	5689	52.17	51.81	75.31	70.37
	Swin-Tiny	3287	58.24	55.16	78.39	72.68
Lightweight	ShuffleNetV2-1.0×	139	41.88	41.94	67.83	64.85
	+TOP & FDR	177	55.71	52.63	79.82	66.36
	ShuffleNetV2-1.5×	265	47.39	47.81	70.15	65.28
	+TOP & FDR	371	63.35	60.81	84.13	76.98
	GhostNet-1.0×	150	42.53	42.94	71.28	64.40
	+TOP & FDR	189	58.54	55.19	83.26	77.16
	GhostNet-1.3×	281	50.89	47.92	72.51	65.98
	+TOP & FDR	395	**66.76**	**64.01**	**85.51**	**79.95**
	MobileNetV3-S	104	40.92	42.51	62.77	58.31
	+TOP & FDR	130	54.75	50.26	75.53	70.17
	MobileNetV3-L	250	47.81	47.06	71.26	65.66
	+TOP & FDR	362	66.14	63.80	84.15	79.26

Table 5. Evaluation of spatial modelling methods and channel relation reasoning methods in the FDR module. "u." and "cs." respectively denote the up-sampling and channel shuffle operations.

(a): Impact of different spatial modelling methods

Methods	SYSU-MM01 r=1	SYSU-MM01 mAP	SYSU-MM01 mINP	RegDB r=1	RegDB mAP	RegDB mINP
GAP	62.89	59.79	45.84	82.88	76.24	61.92
Context [2]	61.37	57.52	46.08	81.56	74.49	62.11
HAP [38]	62.92	58.13	47.84	82.98	76.20	62.77
$H_s + V_s$ (w/o u.)	63.45	59.75	49.61	83.94	78.82	63.21
$H_s + V_s$	**66.14**	**63.80**	**49.76**	**84.15**	**79.26**	**63.86**

(b): Impact of different channel relation reasoning methods

Methods	SYSU-MM01 r=1	SYSU-MM01 mAP	SYSU-MM01 mINP	RegDB r=1	RegDB mAP	RegDB mINP
SE [16]	62.91	59.70	45.78	82.79	77.45	62.34
CBAM [33]	61.79	56.88	45.25	83.41	77.28	62.96
SA (w/o cs.)	62.81	58.10	45.76	82.75	76.12	62.56
SA	**66.14**	**63.80**	**49.76**	**84.15**	**79.26**	**63.86**

embed all the features into their respective identity groups. Thus, the proposed task-oriented pretraining strategy makes lightweight networks fit cross-modality tasks.

Pattern Visualization. As demonstrated in Fig. 5, we visualize the learned patterns using Grad-CAM [27] to explain the effect of the FDR module. Clearly, the original patterns extracted by MobileNetV3-L are concentrated and overfit on the modality-specific parts. To remedy this issue, we use the FDR module to discover modality-shared patterns from fine-grained regions, which separates learned patterns to focus on considerable modality-shared cues. A good case is shown in Fig. 5. The visible patterns initially learned are concentrated in the high chest area, while the infrared patterns are all in the right arm. After re-weighting by the FDR module, modality-shared small regions like the face, arms, neck, and shoes are all taken into account, thus

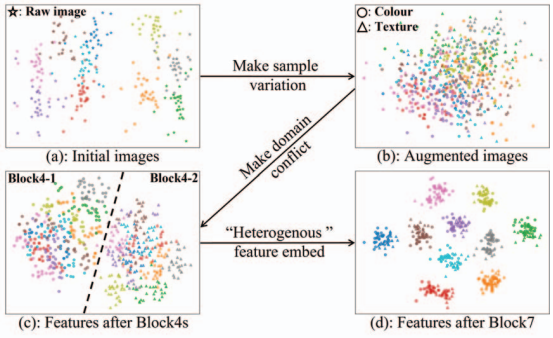

Figure 4. Feature distributions during task-oriented pretraining. Each colour denotes an identity. In (a), the star markers represent raw samples. In (b), (c) and (d), the circle and triangle markers represent colour and texture-augmented samples, respectively.

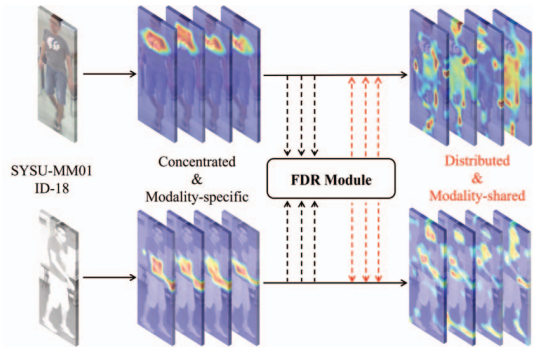

Figure 5. Visualization of learned patterns via Grad-CAM [27]. For a clear presentation, we average all the spatial maps into four.

Table 6. Comparison with the state-of-the-arts on SYSU-MM01 [35]. Metrics of Rank at r (%), mAP (%) and mINP (%) are reported.

Details			All-search					Indoor-search				
Methods	Backbone	FLOPs (M)	r=1	r=10	r=20	mAP	mINP	r=1	r=10	r=20	mAP	mINP
Zero-pad [35]	ResNet50	>3562	14.80	54.12	71.33	15.95	–	20.58	68.38	85.79	26.92	–
JSIA [30]	ResNet50+GAN	>4133	38.10	80.70	89.90	36.90	–	43.80	86.20	94.20	52.90	–
AGW [42]	ResNet50	>3562	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23
X-Modal [19]	ResNet50	>3562	49.90	89.80	96.00	50.70	–	–	–	–	–	–
DMiR [38]	ResNet50	>3562	50.54	88.12	94.86	49.29	–	53.92	92.50	97.09	62.49	–
FBP-AL [32]	ResNet50	>3562	54.14	86.04	93.03	50.20	–	–	–	–	–	–
DDAG [41]	ResNet50	>3562	54.75	90.39	95.81	55.02	39.62	61.02	94.06	98.41	67.98	62.61
HAT [43]	ResNet50	>3562	55.29	92.14	97.36	53.89	–	62.10	95.75	99.20	70.84	–
LBA [26]	ResNet50	>3562	55.41	–	–	54.14	–	58.46	–	–	66.33	–
TSME [21]	ResNet50	>3562	64.23	95.19	**98.73**	61.21	–	64.80	96.92	**99.31**	71.53	–
SPOT [4]	ResNet50+ViT	>4810	65.34	92.73	97.04	62.25	48.86	69.42	96.22	99.12	74.63	70.48
TOPLight (Ours)	MobileNetV3-L	= 362	66.14	96.03	97.68	63.80	49.76	72.41	97.54	99.23	76.11	71.43
TOPLight (Ours)	GhostNet-1.3×	= 395	**66.76**	**96.23**	98.70	**64.01**	**50.18**	**72.89**	**97.93**	99.28	**76.70**	**71.95**

Table 7. Comparison with the state-of-the-arts on RegDB [25]. Metrics of Rank at r (%), mAP (%) and mINP (%) are reported.

Details			Visible-to-Infrared					Infrared-to-Visible				
Methods	Backbone	FLOPs (M)	r=1	r=10	r=20	mAP	mINP	r=1	r=10	r=20	mAP	mINP
Zero-pad [35]	ResNet50	>3562	17.75	34.21	44.35	18.90	–	16.63	34.68	44.25	17.82	–
JSIA [30]	ResNet50+GAN	>4133	48.50	–	–	49.30	–	48.10	–	–	48.90	–
AGW [42]	ResNet50	>3562	70.05	86.21	91.55	66.37	50.19	70.49	87.21	91.84	65.90	51.24
X-Modal [19]	ResNet50	>3562	62.21	83.13	91.72	60.18	–	–	–	–	–	–
DMiR [38]	ResNet50	>3562	75.79	89.86	94.18	69.97	–	73.93	89.87	93.98	68.22	–
FBP-AL [32]	ResNet50	>3562	73.98	89.71	93.69	68.24	–	70.05	89.22	93.88	66.61	–
DDAG [41]	ResNet50	>3562	69.34	86.19	91.49	63.46	49.24	68.06	85.15	90.31	61.80	48.62
HAT [43]	ResNet50	>3562	71.83	87.16	92.16	67.56	–	70.02	68.45	91.61	66.30	–
LBA [26]	ResNet50	>3562	74.17	–	–	67.64	–	72.43	–	–	65.46	–
SPOT [4]	ResNet50+ViT	>4810	80.35	93.48	96.44	72.46	56.19	79.37	92.79	96.01	72.26	56.06
GECNet [48]	ResNet50+GAN	>4350	82.33	92.72	95.49	78.45	–	78.93	91.99	95.44	75.58	–
TOPLight (Ours)	MobileNetV3-L	=362	84.15	94.98	96.58	79.26	**63.86**	**80.94**	**92.85**	**96.37**	**76.10**	**59.33**
TOPLight (Ours)	GhostNet-1.3×	=395	**85.51**	**94.99**	**96.70**	**79.95**	63.85	80.65	92.81	96.32	75.91	59.26

improving the accuracy of the cross-modality matching.

4.4. Comparison with State of The Arts

We compare our method with recently proposed state-of-the-art VI-ReID methods, as shown in Table 6 and Table 7. They illustrate that we remarkably and efficiently exceed all the compared SOTAs under diverse evaluation settings. Specifically, based on the tiny GhostNet-1.3× backbone, we attain the rank-1 of 66.76% and mAP of 64.01% on the all-search mode of the large-scale SYSU-MM01 dataset with only 395 FLOPs. The proposed method also achieves excellent results on the RegDB dataset, surpassing the previous SOTAs by a significant margin with lower FLOPs.

Table 8. Evaluation on two VI-FR datasets. CA is channel augmentation [40]. B is the LightCNN-29 baseline. Rank at 1 accuracy (%) and false acceptance rate (F: %) are reported.

Methods	Oulu [5]			BUAA [17]		
	r=1	F:1%	F:0.1%	r=1	F:1%	F:0.1%
IDR [13]	94.3	73.4	46.2	94.3	93.4	84.7
VSA [44]	99.9	96.8	82.3	98.0	98.2	92.5
PACH [8]	100	97.9	88.2	98.6	98.0	93.5
B [36]	100	97.9	87.0	98.0	97.7	93.7
B+CA [40]	100	98.9	91.7	98.3	98.2	94.5
B+TOP	100	98.8	91.5	98.3	98.1	94.5
B+TOP+FDR (Ours)	100	98.9	91.7	98.3	98.2	94.6

4.5. Visible-Infrared Face Recognition

We follow [40] to examine our method on Oulu-CASIA NIR-VIS [5] and BUAA-VisNir face databases [17]. Following previous studies, we also adopt the LightCNN-29 [36] as our baseline network (B). Except for IDR [13], VSA [44], and PACH [8], we also compare recently proposed channel augmentation (CA) [40] with our method on the same baseline. Results are displayed in Table 8.

Evidently, the proposed TOP and FDR also enhance the performance of VI face recognition. These results further validate the effectiveness and generalization of our methods, which can solve VI recognition tasks effectively.

5. Conclusion

This work presents an effective task-oriented pretrained lightweight neural network (TOPLight) to solve visible-infrared recognition problems. First, the task-oriented pretraining strategy significantly improves the ability of lightweight networks to understand heterogeneous images by introducing domain conflict and sample variation during the pretraining phase. Second, the fine-grained dependency reconstruction module is utilized to destroy the modality-specific pattern dependencies and construct the fine-grained and modality-shared pattern dependencies between visible and infrared features. Extensive experiments demonstrate the superiority and effectiveness of the proposed method as well as the efficacy of each component of our framework.

Acknowledgements. This work is funded by the National Natural Science Foundation of China (Grant No. 61802058, 61911530397).

References

[1] Sun , $et.al$. Learning part-based convolutional features for person re-identification. TPAMI, 2019. 2

[2] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 7

[3] Srinivasa Rao Chalamala, Krishna Rao Kakkirala, and Jami Santosh Kumar. Face recognition using spatial pyramid matching and lrbp. In *2014 IEEE 10th International Colloquium on Signal Processing and its Applications*, pages 67–70. IEEE, 2014. 2

[4] Cuiqun Chen, Mang Ye, Meibin Qi, Jingjing Wu, Jianguo Jiang, and Chia-Wen Lin. Structure-aware positional transformer for visible-infrared person re-identification. *IEEE Transactions on Image Processing*, 31:2352–2364, 2022. 1, 8

[5] Jie Chen, Dong Yi, Jimei Yang, Guoying Zhao, Stan Z Li, and Matti Pietikainen. Learning mappings for face synthesis from near infrared to visual light images. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–163. IEEE, 2009. 8

[6] Changxing Ding and Dacheng Tao. A comprehensive survey on pose-invariant face recognition. *ACM Transactions on intelligent systems and technology (TIST)*, 7(3):1–42, 2016. 1

[7] Hang Du, Hailin Shi, Dan Zeng, Xiao-Ping Zhang, and Tao Mei. The elements of end-to-end deep face recognition: A survey of recent advances. *ACM Computing Surveys (CSUR)*, 54(10s):1–42, 2022. 1

[8] Boyan Duan, Chaoyou Fu, Yi Li, Xingguang Song, and Ran He. Cross-spectral face hallucination via disentangling independent factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7930–7938, 2020. 1, 2, 8

[9] Hamed Kiani Galoogahi and Terence Sim. Face sketch recognition by local radon binary pattern: Lrbp. In *2012 19th IEEE International Conference on Image Processing*, pages 1837–1840. IEEE, 2012. 2

[10] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 4

[11] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020. 1, 6

[12] 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

[13] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Learning invariant deep representation for nir-vis face recognition. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 1, 2, 8

[14] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1761–1773, 2018. 2

[15] 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. 1, 5, 6

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5, 6, 7

[17] D Huang, J Sun, and Y Wang. The buaa-visnir face database instructions. *School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001*, 3:3, 2012. 8

[18] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10657–10665, 2019. 2, 3

[19] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4610–4617, 2020. 2, 6, 8

[20] Shengcai Liao, Dong Yi, Zhen Lei, Rui Qin, and Stan Z Li. Heterogeneous face recognition from local structures of normalized appearance. In *International Conference on Biometrics*, pages 209–218. Springer, 2009. 2

[21] Jianan Liu, Jialiang Wang, Nianchang Huang, Qiang Zhang, and Jungong Han. Revisiting modality-specific feature compensation for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 8

[22] Xiaoxiang Liu, Lingxiao Song, Xiang Wu, and Tieniu Tan. Transferring deep representation for nir-vis heterogeneous face recognition. In *2016 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2016. 2

[23] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 6

[24] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018. 1, 6

[25] 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. 5, 8

[26] 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, 3, 5, 6, 8

[27] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra.

Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[28] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[29] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3623–3632, 2019. 1, 2

[30] Guan-An Wang, Tianzhu Zhang, Yang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zeng-Guang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12144–12151, 2020. 1, 2, 5, 8

[31] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 618–626, 2019. 2

[32] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Flexible body partition-based adversarial learning for visible infrared person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 8

[33] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 6, 7

[34] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, pages 23965–23998. PMLR, 2022. 3

[35] 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. 1, 2, 5, 8

[36] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 8

[37] Xiang Wu, Huaibo Huang, Vishal M Patel, Ran He, and Zhenan Sun. Disentangled variational representation for heterogeneous face recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9005–9012, 2019. 2

[38] Mang Ye, Cuiqun Chen, Jianbing Shen, and Ling Shao. Dynamic tri-level relation mining with attentive graph for visible infrared re-identification. *IEEE Transactions on Information Forensics and Security*, 17:386–398, 2021. 1, 2, 5, 6, 7, 8

[39] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Transactions on Information Forensics and Security*, 15:407–419, 2019. 2

[40] Mang Ye, Weijian Ruan, Bo Du, and Mike Zheng Shou. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13567–13576, 2021. 6, 8

[41] Mang Ye, Jianbing Shen, David J Crandall, Ling Shao, and Jiebo Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 229–247. Springer, 2020. 8

[42] 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, 5, 8

[43] Mang Ye, Jianbing Shen, and Ling Shao. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security*, 16:728–739, 2020. 8

[44] Aijing Yu, Haoxue Wu, Huaibo Huang, Zhen Lei, and Ran He. Lamp-hq: A large-scale multi-pose high-quality database and benchmark for nir-vis face recognition. *International Journal of Computer Vision*, 129(5):1467–1483, 2021. 1, 8

[45] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7349–7358, 2022. 2

[46] Teng Zhang, Arnold Wiliem, Siqi Yang, and Brian Lovell. Tv-gan: Generative adversarial network based thermal to visible face recognition. In *2018 international conference on biometrics (ICB)*, pages 174–181. IEEE, 2018. 2

[47] Huantao Zheng, Xian Zhong, Wenxin Huang, Kui Jiang, Wenxuan Liu, and Zheng Wang. Visible-infrared person re-identification: A comprehensive survey and a new setting. *Electronics*, 11(3):454, 2022. 2

[48] Xian Zhong, Tianyou Lu, Wenxin Huang, Mang Ye, Xuemei Jia, and Chia-Wen Lin. Grayscale enhancement colorization network for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1418–1430, 2021. 8

[49] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 5

[50] Jun-Yong Zhu, Wei-Shi Zheng, Jian-Huang Lai, and Stan Z Li. Matching nir face to vis face using transduction. *IEEE Transactions on Information Forensics and Security*, 9(3):501–514, 2014. 2