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, USA

Paper tag :TUE-AM-337

Quick preview

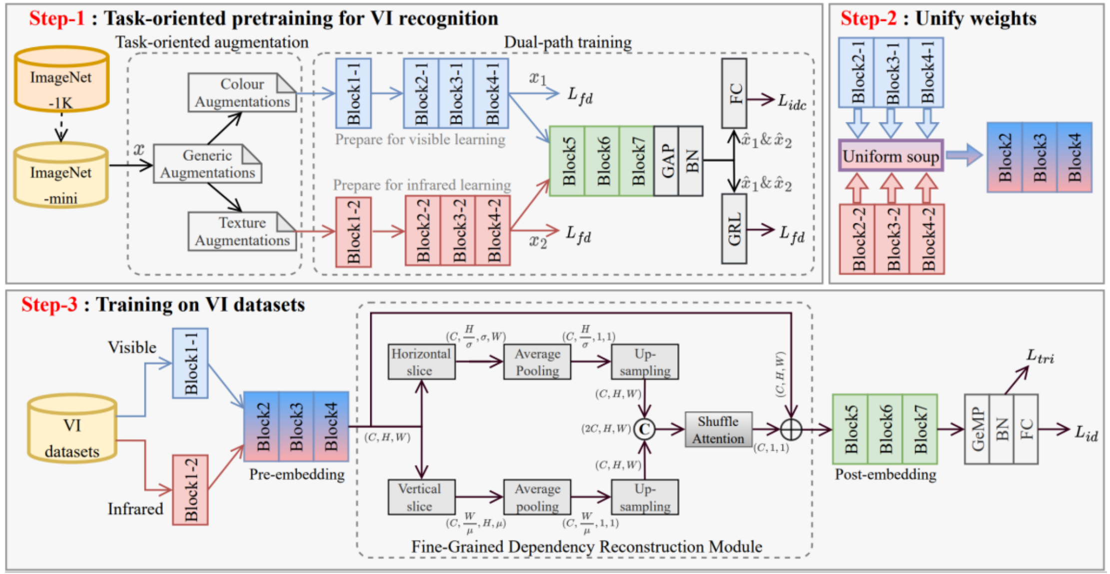
- Motivations

➢ Solve the visible-infrared recognition task efficiently and device friendly.
➢ Make the model easy and quick to train, finetune, and deploy.

- How to?

➢ Use lightweight backbones instead of the ResNet-50.
➢ Improve the pretraining strategy rather than pile up many modules.

- Three steps to understand our solutions



1. Prepare a lightweight network using the task-oriented pretraining strategy.
2. Use the uniform soup to make the structure best for VI training.
3. Training (finetuning) on VI datasets with the FDR module.

- Results

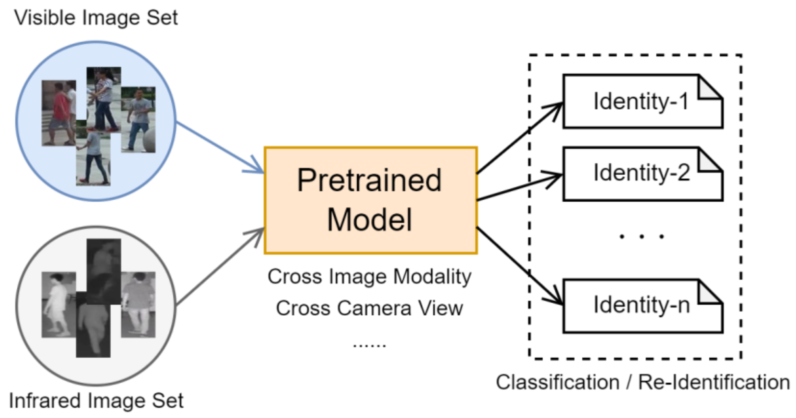
	Methods	FLOPs (M)	SYSU-MM01 r=1	SYSU-MM01 mAP	RegDB r=1	RegDB mAP
Convention	ResNet-50	3562	56.98	54.72	76.86	71.30
	ConvNeXt-Tiny	3620	58.72	55.31	78.25	72.64
	Vit-B	5689	52.17	51.81	75.31	70.37
	Swin-Tiny	3287	58.24	55.16	78.39	72.68
Lightweight	ShuffleNetV2-1.0×	139	41.88	41.94	67.83	64.85
	+TOP & 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

➢ Make lightweight networks better than conventional deep networks.

➢ Around 10x faster than previous solutions.

Intro

- What is visible-infrared recognition?



- Shortcoming in existing work

Too heavy to deploy on edge devices

- Potential solutions

Use ImageNet pretrained lightweight backbones instead of the commonly used ResNet-50 for feature extraction.

- Issue

Huge performance gap

Model	Top-1 on ImageNet-1k	Rank-1 on SYSU-MM01
ResNet-50	78.8	56.98
MobileNetV3-L	75.2(↓ 3.6)	47.81(↓ 9.71)

- Why?

1. The ImageNet is a pure visible dataset.
2. Few learnable parameters, few learnt visual patterns.
3. Colour-related prior knowledge is dominant!

Our solution: Task-oriented pretraining

- **What is the difference between existing pretrain methods and ours ?**

P1 :

1. Existing: Prepare for all downstream tasks
2. **Ours: Prepare for VI recognition only.**

P2 :

1. Existing: Pretrain->Finetune
2. **Ours: (Pretrain->Adapt)->Finetune**

P3 :

1. Existing: From One-path to Dual-path
2. **Ours: From Dual-path to Dual-path**

- **What does TOP do?**

During the pretrain stage, we hope to

(1) Let the network learn prior knowledge related to infrared images.

(2) Disturb the color-prior knowledge to make the network pay more attention to the modality-shared patterns.

(3) Let the network know how to extract shared patterns from two groups of “heterogenous features” to identify them well.

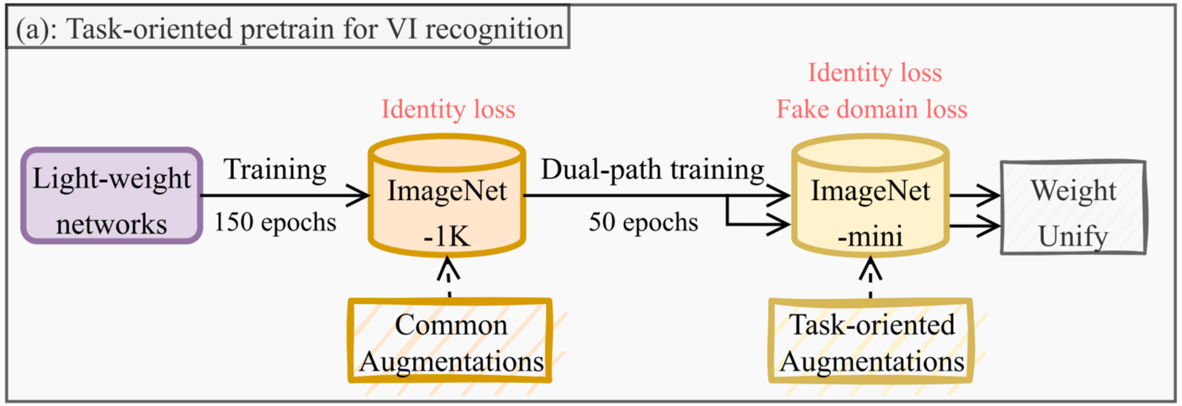
- **How to?**

1. Task-oriented Augmentation

→ Simulate the visual differences in VI scenes and disturb the colour information

2. Dual-path training with fake domain loss

→ Improve the “heterogenous feature” represent and embed capacity



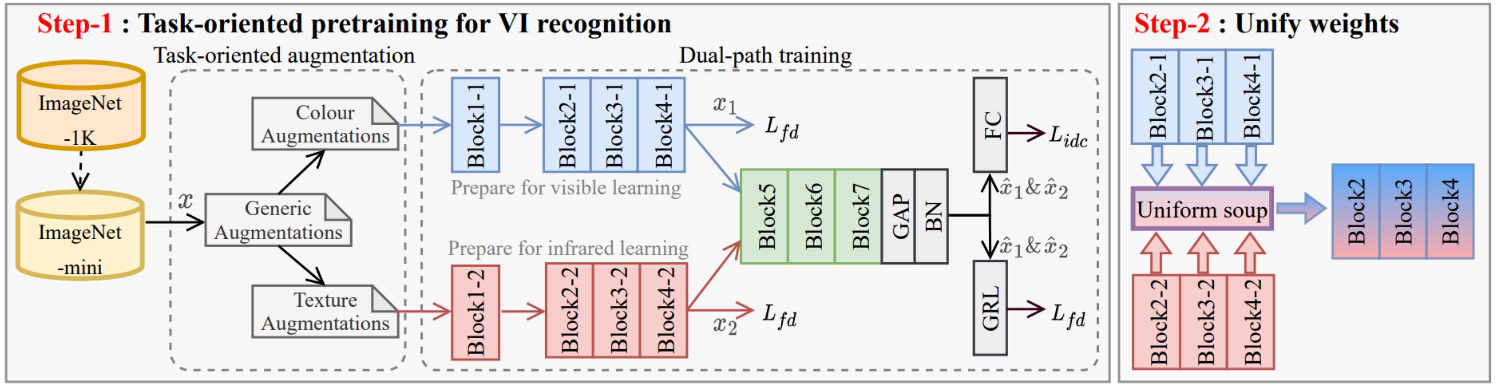
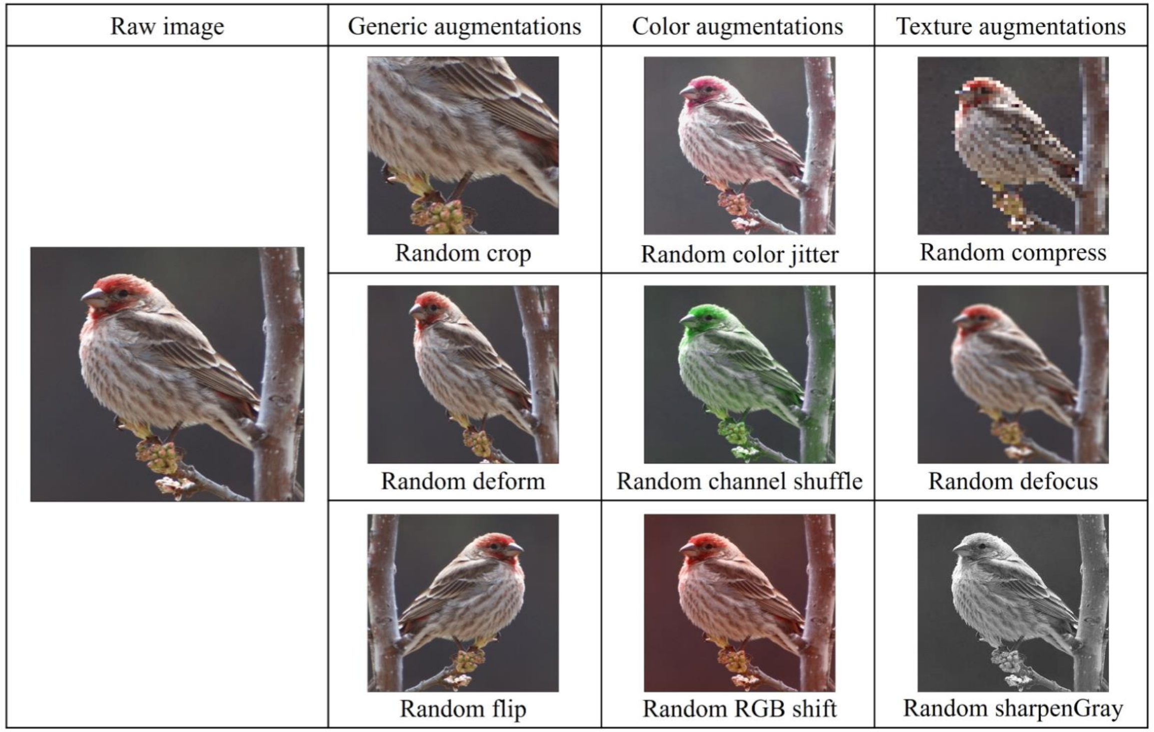


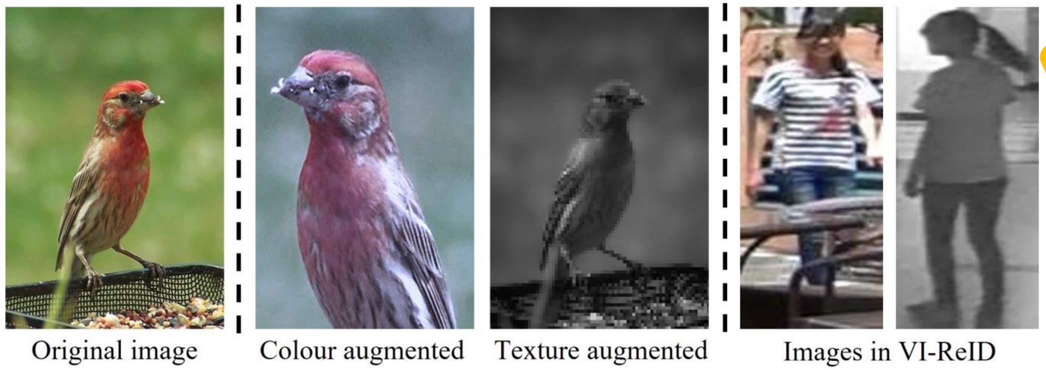
Table 1. Detailed structures of each Block. We package the entire mobileNetV3-Large into Block1-7 without overlap.

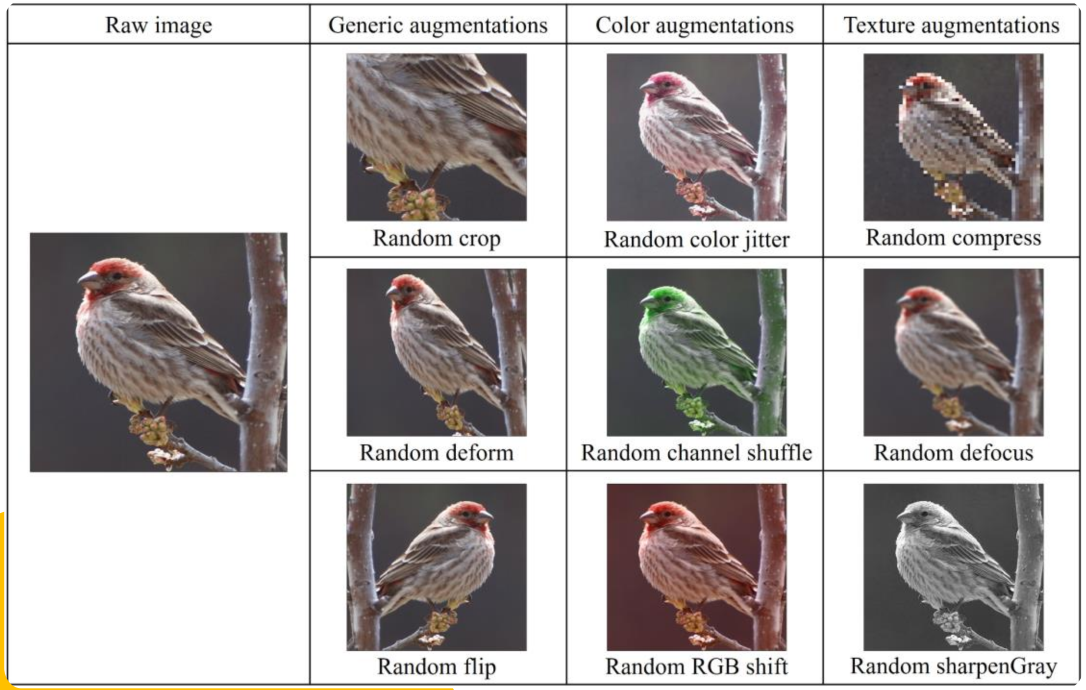
Block partitions on MobileNetV3-large		
Layer name	Structures	Output size
Block1	conv(3×3, 2), bneck(3×3, 1)	16×112^2
Block2	bneck(3×3, 2)	24×56^2
Block3	bneck(3×3, 1)	24×56^2
Block4	bneck(5×5, 2), bneck(5×5, 1)*2	40×28^2
Block5	bneck(3×3, 2), bneck(3×3, 1)*5	112×14^2
Block6	bneck(5×5, 2), bneck(5×5, 1)*2	160×7^2
Block7	conv(1×1, 1)*3	1280×7^2

- Firstly, we pretrain our network on ImageNet-1k with identity-loss and common augmentations (Crop+Flip).
- Secondly, we package the entire trained network (example shows the MobileNetV3-L) into Block1-Block7. Then, for the Block1-Block4, we initialize them twice with the same pretrain weights to make the dual-path network.
- Thirdly, we retrain the dual path network on ImageNet-mini, with task-oriented augmentation to create visual differences between each path. During training, the identity-consistency loss (L_{idc}) and fake domain loss (L_{fd}) are adopted to supervise the overall network.
- Finally, after the retrain on ImageNet-mini, we unify the weights of Block(2,3,4)-1 and Block(2,3,4)-2 via Uniform Soup. That makes the final dual-path network used in VI datasets only have two stem blocks (Block1-1, Block1-2).



Why task-oriented augmentation?





- The **generic DAs** aim to increase the broad diversity
- The **colour DAs** are designed to disturb the regularity of colour-prior information and are only imposed on the branch prepared for visible learning.
- The **texture DAs** are designed to to remove the colour information and change the texture styles only in the branch prepared for infrared learning.
- **Combining them all, we aim to simulate the visual differences in the real VI recognition scene and train the dual-path network to handle them during the pretraining stage. E.g., in this manner, two stem blocks are trained for extracting the modality-prior irrelevant patterns, like global shapes and the relative position of local objects.**

Why we need fake domain loss ?

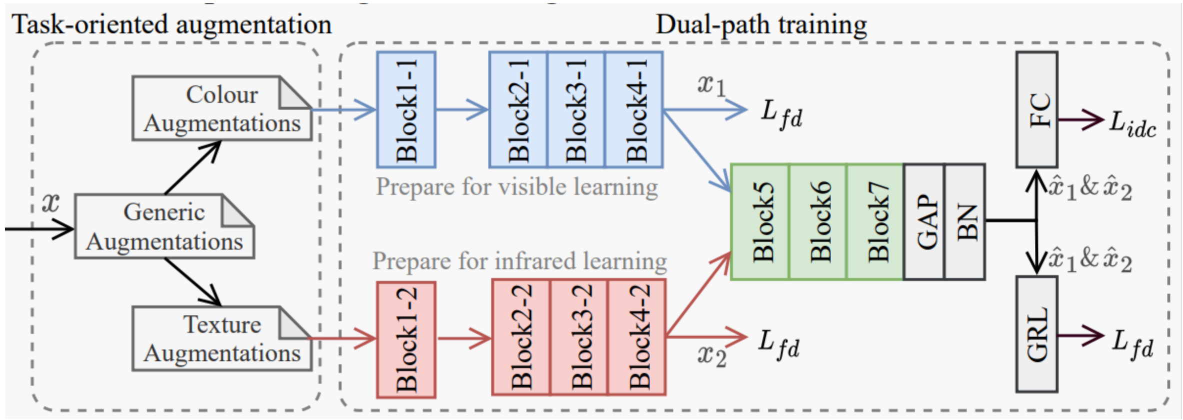
- Just in the above manner, the network may still lazily learn from one path to avoid feature embedding. Meanwhile, the visual differences made via augmentations are scanty to simulate the actual domain conflict during training. Thus, we proposed the fake domain loss to perform the self-against learning, which impels the network to learn domain knowledge from both paths.

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

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

$\theta_1 = \{0,0,0,\dots,0\}$, pretend to be domain A

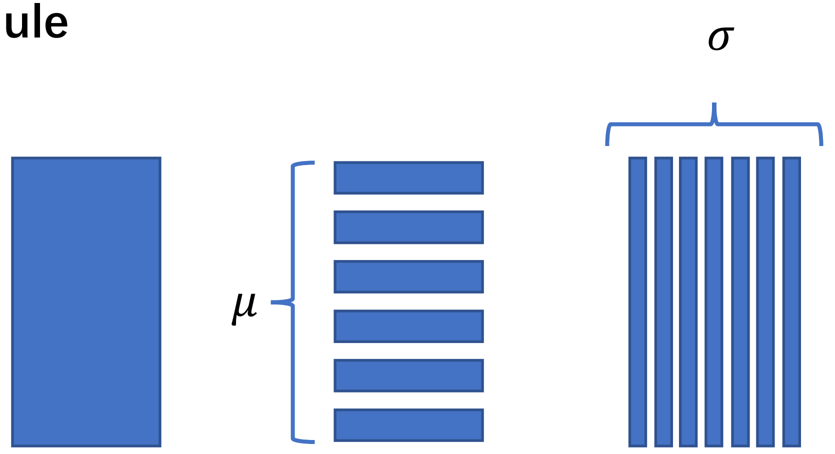
$\theta_2 = \{1,1,1,\dots,1\}$, pretend to be domain B



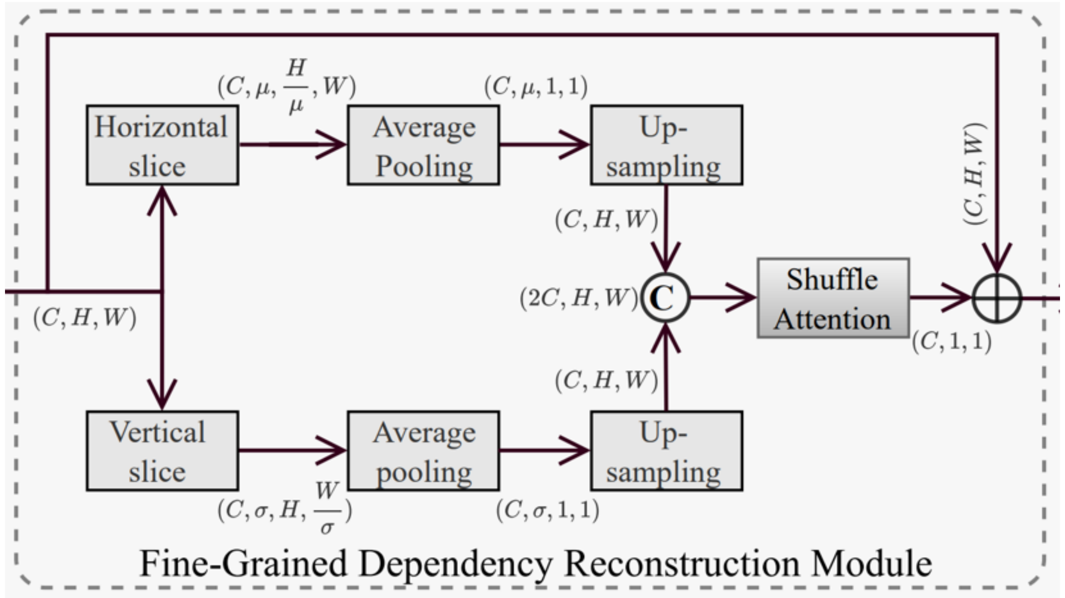
- In this manner, we create two contradictory learning procedures: the positive domain constraints are set on $x_1 \& x_2$, which force them to be representative for the fake domain we pretended. Meanwhile, with reserved gradients, inversed domain constraints are set on $\hat{x}_1 \& \hat{x}_2$, which encourage the final features after Block7 to be domain-shared.

- During this procedure, Block(2,3,4)-1 and Block(2,3,4)-2 are trained to extract two types of strongly distinguished features. In comparison, Block(5,6,7) are trained to embed these two types of "heterogenous features" and find their common ground.

The icing on the cake: Fine-grained dependency reconstruction module



- This module intends to help lightweight networks build cross-modality correlations effectively. The core motivation is to break the original dependencies, and then build the modality-shared one. It can be summarized as two parts: The spatial modelling part (before shuffle attention) and channel relation reasoning part (shuffle attention).



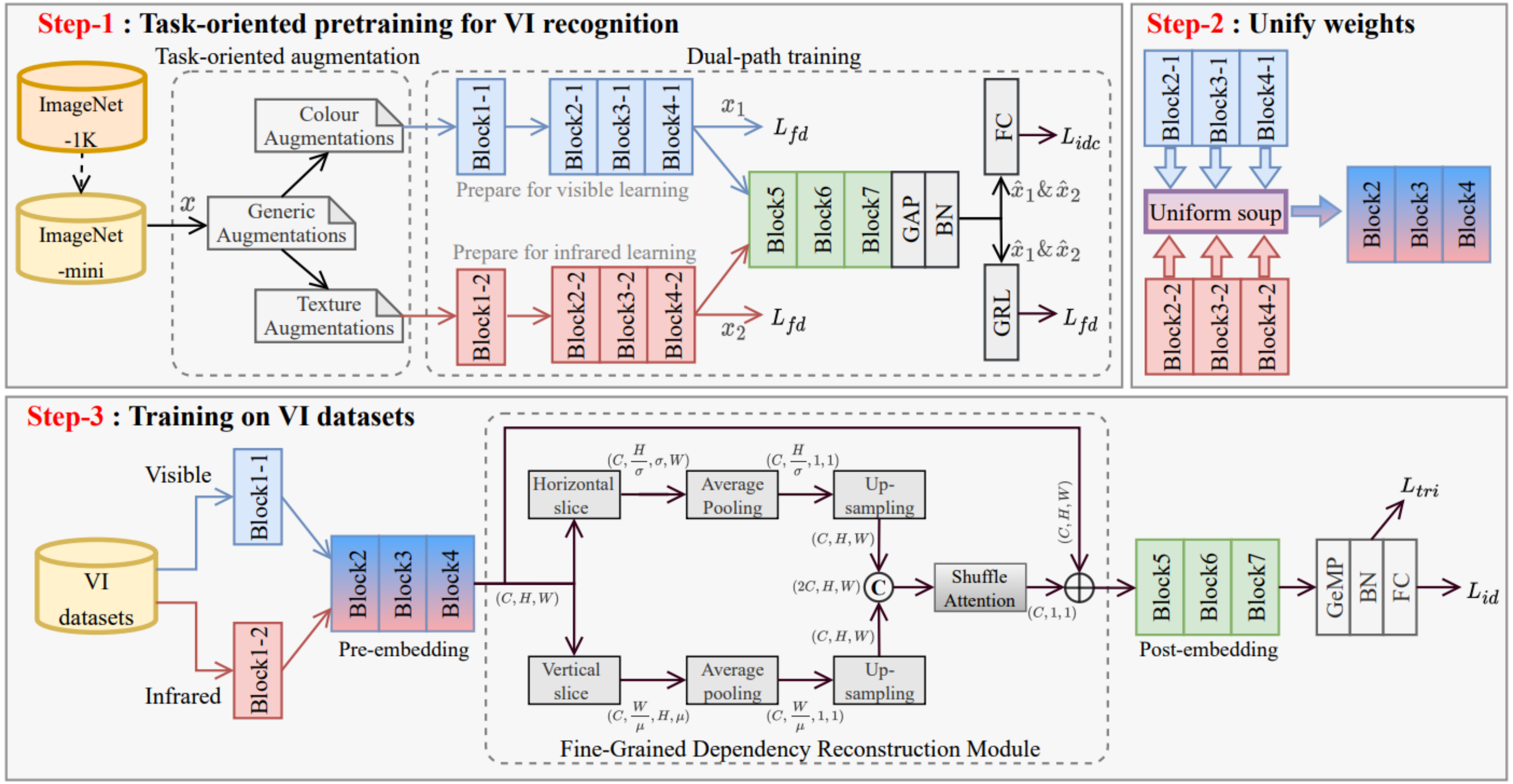
- For the first part, we first slice the original features into two types of fine-grained features : horizontal and vertical. They are respectively concated in the second dimension. Then, we use the average pooling operation to concentrate the sliced spatial information, which converts the original spatial maps into vectors. The procedure breaks the original spatial dependencies among each fine-grained regions.

- After that, two independent Up-sampling layers are adopted to reconstruct the spatial maps according to two types of directional vectors. The "embedding with up-sampling" scheme intends to fully discover the modality-shared patterns from the re-enlarged regions fully.

- Finally, we concat these two types of features in the channel dimension and fed them into the shuffle attention module to perform channel relation reasoning.

The Overall Pipeline of This Paper

- **Contributions**
  - We proposed a task-oriented pretraining strategy for VI recognition.
  - We proposed a fine-grained dependency reconstruction module for VI recognition.
  - **We make the lightweight networks competitive with the regulars for VI recognition.**
  - **Our methods reach the current SOTA level with nearly 1/10 FLOPs using common identity and triplet loss.**



Ablation Experiments

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 on 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 on 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**

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**

Comparison with SOTAs in VI ReID

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**

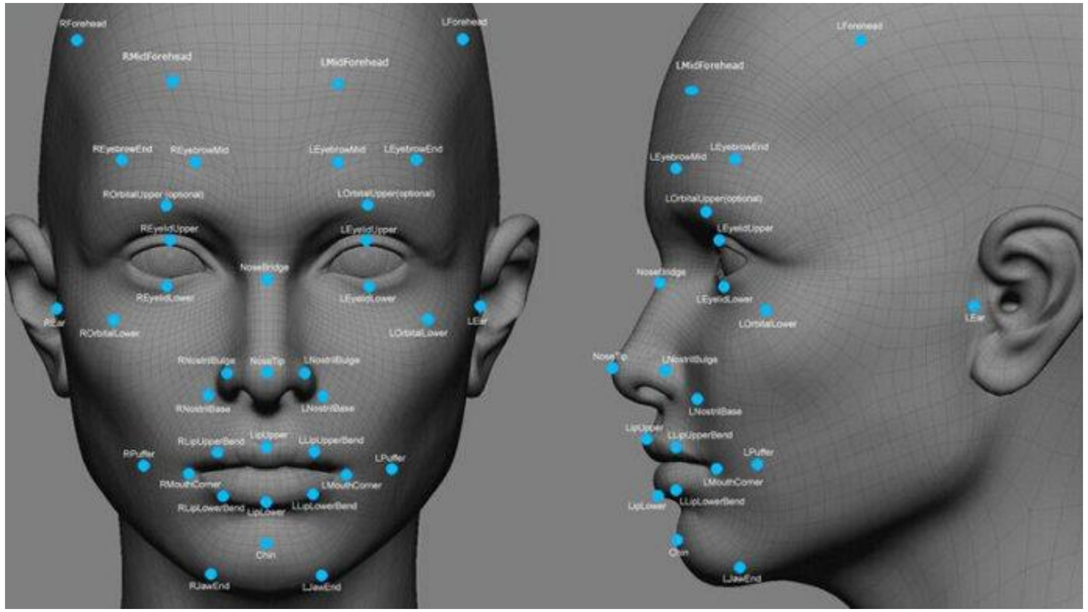
Table 7. Comparison with the state-of-the-arts on RegDB [25]. Metrics of Rank at r (%), mAP (%) and mINP (%) are reported.

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

Comparison with SOTAs in VI Face Recognition

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



Thanks!