Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-identification

Yukang Zhang[1,2], Hanzi Wang[1,2,3]*
zhangyk@stu.xmu.edu.cn, hanzi.wang@xmu.edu.cn

[1]Fujian Key Laboratory of Sensing and Computing for Smart City,
School of Informatics, Xiamen University, 361005, P.R. China.
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China.
[3]Shanghai Artificial Intelligence Laboratory, Shanghai, 200232, China.

https://github.com/ZYK100/LLCM

Poster Session: TUE-AM-206

Motivation

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Q. Why needs to generate diverse embedding features?

A: In VIReID and LL-VIReID task, the person samples are usually limited, while the modality gaps are too large. Thus, we generate diverse embedding features to effectively mine diverse cross-modality clues.

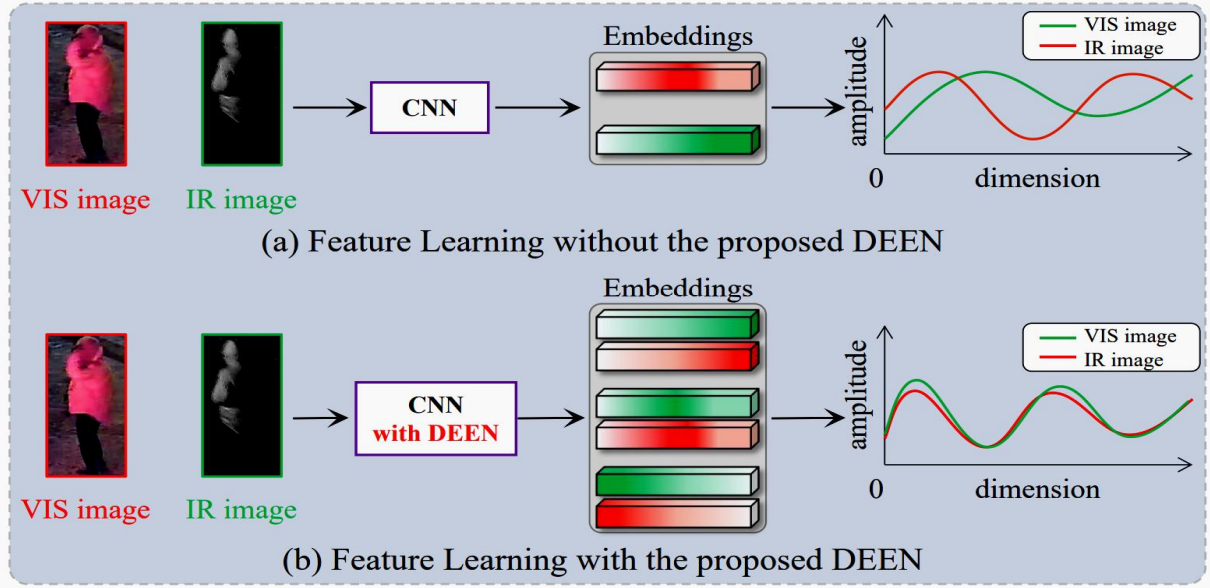

Figure 1: Motivation of the proposed DEEN, which aims to generate diverse embeddings to make the network focus on learning with the informative feature representations to reduce the modality gaps between the VIS and IR images.

Method

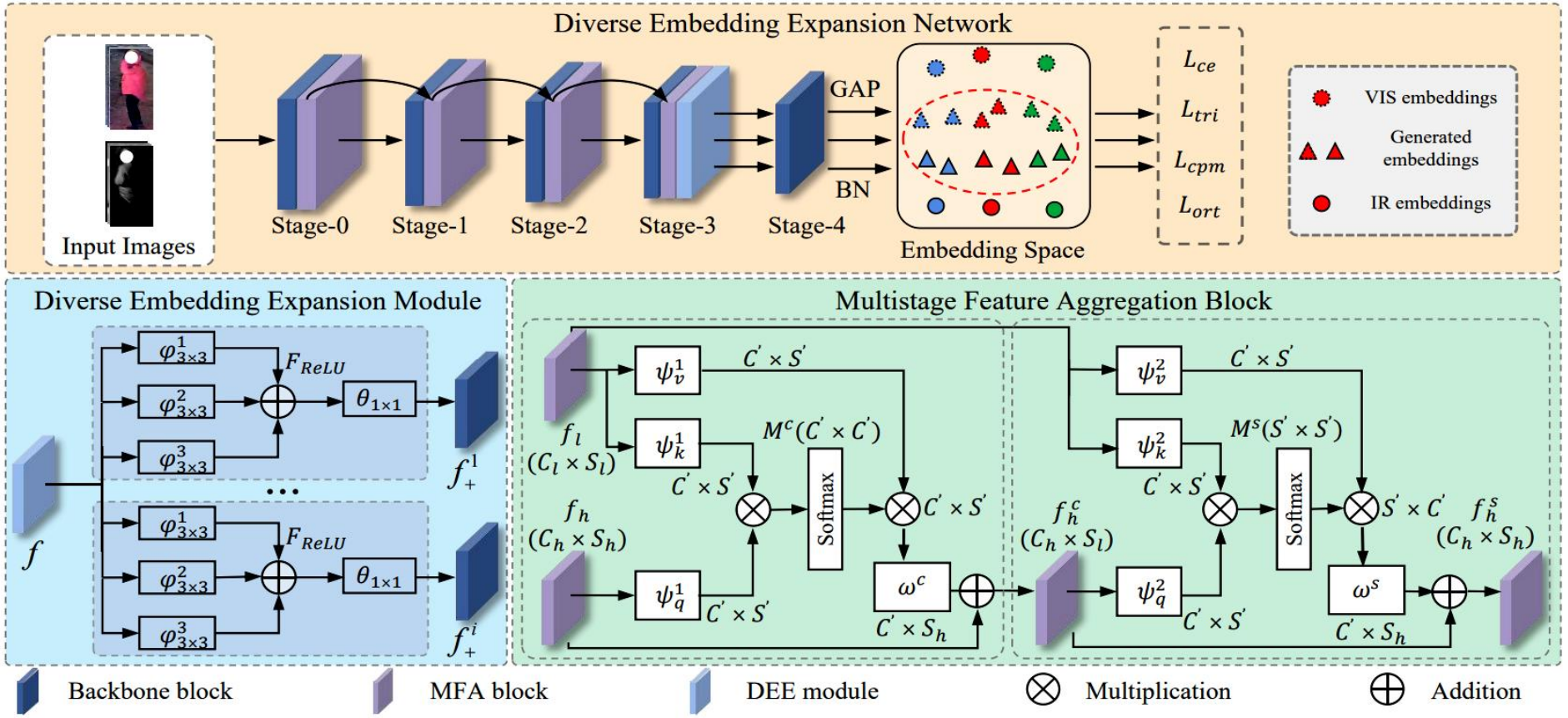

Figure 2: The pipeline of the proposed DEEN, which includes a DEE module and a MFA block.

Method

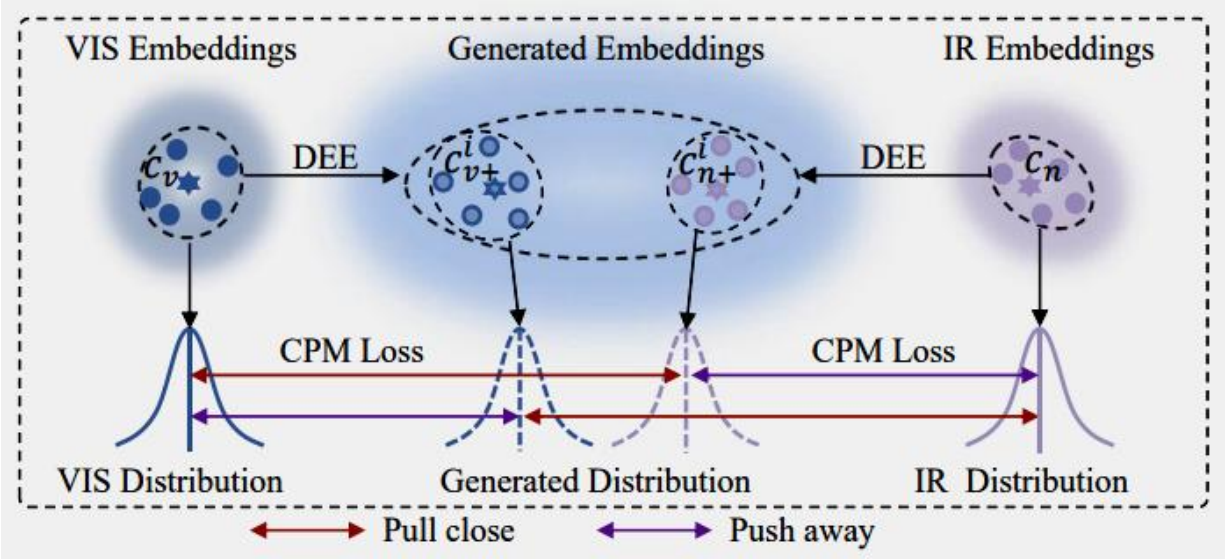

Figure 3: Illustration of the proposed CPM loss for DEE.

$$\mathcal{L}(\mathbf{c}_v, \mathbf{c}_n, \mathbf{c}_{v+}^i) = [\boldsymbol{D}(\mathbf{c}_n^j, \mathbf{c}_{v+}^{i,j}) - \boldsymbol{D}(\mathbf{c}_v^j, \mathbf{c}_{v+}^{i,j}) - \boldsymbol{D}(\mathbf{c}_v^j, \mathbf{c}_v^k) + \alpha]_+.$$

$$\mathcal{L}(\mathbf{c}_v, \mathbf{c}_n, \mathbf{c}_{n+}^i) = [\boldsymbol{D}(\mathbf{c}_v^j, \mathbf{c}_{n+}^{i,j}) - \boldsymbol{D}(\mathbf{c}_n^j, \mathbf{c}_{n+}^{i,j}) - \boldsymbol{D}(\mathbf{c}_n^j, \mathbf{c}_n^k) + \alpha]_+.$$

$$\mathcal{L}_{cpm} = \mathcal{L}(\mathbf{c}_v, \mathbf{c}_n, \mathbf{c}_{v+}^i) + \mathcal{L}(\mathbf{c}_v, \mathbf{c}_n, \mathbf{c}_{n+}^i).$$

Property 1: The generated embeddings should be as **diverse** as possible to effectively learn the informative feature representations.

Property 2: The generated embeddings should **facilitate reducing the modality discrepancy** between the VIS and IR images.

Property 3: **The intra-class distance should be less than the inter-class one.**

LLCM (Low-Light Cross-Modality) Dataset

(a) Different time under the same cameras

(b) Different light conditions under the same cameras

(c) Different cameras under the same time

Figure 4: The VIS and NIR images with different conditions under night.

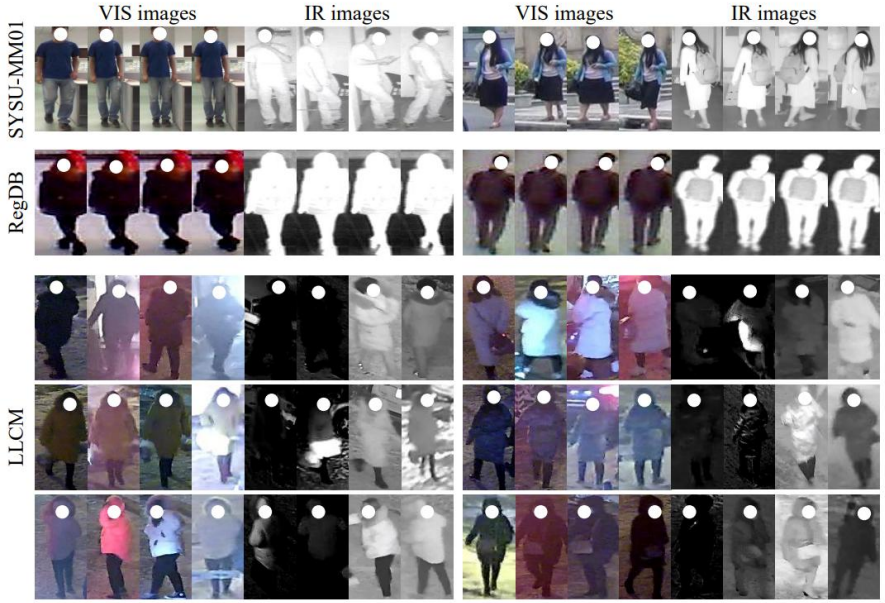

Figure 5: Comparison of person images on the SYSU-MM01, RegDB , and LLCM datasets.

Datasets	IDs	Images	VIS / IR cam.	low-light
RegDB [19]	412	8,240	1 / 1	✗
SYSU-MM01 [36]	491	38,271	4 / 2	✗
LLCM	**1,064**	**46,767**	**9 / 9**	✓

Table 1. Comparison between the LLCM and other two popular VIReID datasets.

Experiments

Methods	SYSU-MM01								RegDB							
	All Search				Indoor Search				VIS to IR				IR to VIS			
	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
BDTR [46]	17.0	55.4	72.0	19.7	-	-	-	-	33.6	58.6	67.4	32.8	32.9	58.5	68.4	32.0
D^{2}RL [32]	28.9	70.6	82.4	29.2	-	-	-	-	43.4	66.1	76.3	44.1	-	-	-	-
Hi-CMD [4]	34.9	77.6	-	35.9	-	-	-	-	70.9	86.4	-	66.0	-	-	-	-
JSIA-ReID [29]	38.1	80.7	89.9	36.9	43.8	86.2	94.2	52.9	48.1	-	-	48.9	48.5	-	-	49.3
AlignGAN [28]	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.3	57.9	-	-	53.6	56.3	-	-	53.4
X-Modality [14]	49.9	89.8	96.0	50.7	-	-	-	-	62.2	83.1	91.7	60.2	-	-	-	-
DDAG [44]	54.8	90.4	95.8	53.0	61.0	94.1	98.4	68.0	69.3	86.2	91.5	63.5	68.1	85.2	90.3	61.8
LbA [20]	55.4	-	-	54.1	58.5	-	-	66.3	74.2	-	-	67.6	67.5	-	-	72.4
NFS [2]	56.9	91.3	96.5	55.5	62.8	96.5	99.1	69.8	80.5	91.6	95.1	72.1	78.0	90.5	93.6	69.8
CM-NAS [6]	60.8	92.1	96.8	58.9	68.0	94.8	97.9	52.4	82.8	95.1	97.7	79.3	81.7	94.1	96.9	77.6
MCLNet [10]	65.4	93.3	97.1	62.0	72.6	97.0	99.2	76.6	80.3	92.7	96.0	73.1	75.9	90.9	94.6	69.5
FMCNet [48]	66.3	-	-	62.5	68.2	-	-	74.1	89.1	-	-	84.4	88.4	-	-	83.9
SMCL [34]	67.4	92.9	96.8	61.8	68.8	96.6	98.8	75.6	83.9	-	-	79.8	83.1	-	-	78.6
DART [41]	68.7	96.4	99.0	66.3	72.5	97.8	99.5	78.2	83.6	-	-	75.7	82.0	-	-	73.8
CAJ [43]	69.9	95.7	98.5	66.9	76.3	97.9	99.5	80.4	85.0	95.5	97.5	79.1	84.8	95.3	97.5	77.8
MPANet [37]	70.6	96.2	98.8	68.2	76.7	98.2	99.6	81.0	82.8	-	-	80.7	83.7	-	-	80.9
MMN [49]	70.6	96.2	99.0	66.9	76.2	97.2	99.3	79.6	**91.6**	97.7	**98.9**	84.1	87.5	96.0	98.1	80.5
DCLNet [23]	70.8	-	-	65.3	73.5	-	-	76.8	81.2	-	-	74.3	78.0	-	-	70.6
MAUM [15]	71.7	-	-	68.8	77.0	-	-	81.9	87.9	-	-	**85.1**	87.0	-	-	**84.3**
DEEN (ours)	**74.7**	**97.6**	**99.2**	**71.8**	**80.3**	**99.0**	**99.8**	**83.3**	91.1	**97.8**	**98.9**	**85.1**	**89.5**	**96.8**	**98.4**	83.4

Table 2. Comparisons between the proposed DEEN and some state-of-the-art methods on the SYSU-MM01 and RegDB datasets.

Experiments

Model	LLCM							
	IR to VIS				VIS to IR			
	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
DDAG [44]	40.3	71.4	79.6	48.4	48.0	79.2	86.1	52.3
DDAG* [44]	41.0	73.4	81.9	49.6	48.5	81.0	87.8	53.0
AGW [45]	43.6	74.6	82.4	51.8	51.5	81.5	87.9	55.3
LbA [20]	43.8	78.2	86.6	53.1	50.8	84.3	91.1	55.6
LbA* [20]	44.6	78.2	86.8	53.8	50.8	84.6	91.1	55.9
AGW* [45]	46.4	77.8	85.2	54.8	56.0	84.9	90.6	59.1
CAJ [43]	48.8	79.5	85.3	56.6	56.5	85.3	90.9	59.8
DART [41]	52.2	80.7	87.0	59.8	60.4	87.1	91.9	63.2
MMN [49]	52.5	81.6	88.4	58.9	59.9	88.5	93.6	62.7
DEEN (ours)	**54.9**	**84.9**	**90.9**	**62.9**	**62.5**	**90.3**	**94.7**	**65.8**

Table 3. Performance obtained by the competing methods on our LLCM dataset. The symbol of "*" represents the methods that we reproduced with the random erasing technique.

Settings				LLCM		SYSU-MM01	
DEE	$\mathcal{L}_{cpm}$	$\mathcal{L}_{ort}$	MFA	R-1	mAP	R-1	mAP
				45.4	53.6	60.7	57.7
✓				50.5	59.0	64.7	62.0
✓	✓			53.1	61.1	69.2	66.2
✓		✓		51.5	60.1	65.3	63.2
✓	✓	✓		53.9	62.3	69.8	66.7
			✓	51.2	59.6	64.7	62.0
✓	✓	✓	✓	**54.9**	**62.9**	**74.7**	**71.8**

Table 4. The influence of each component on the performance of the proposed DEEN.

Experiments

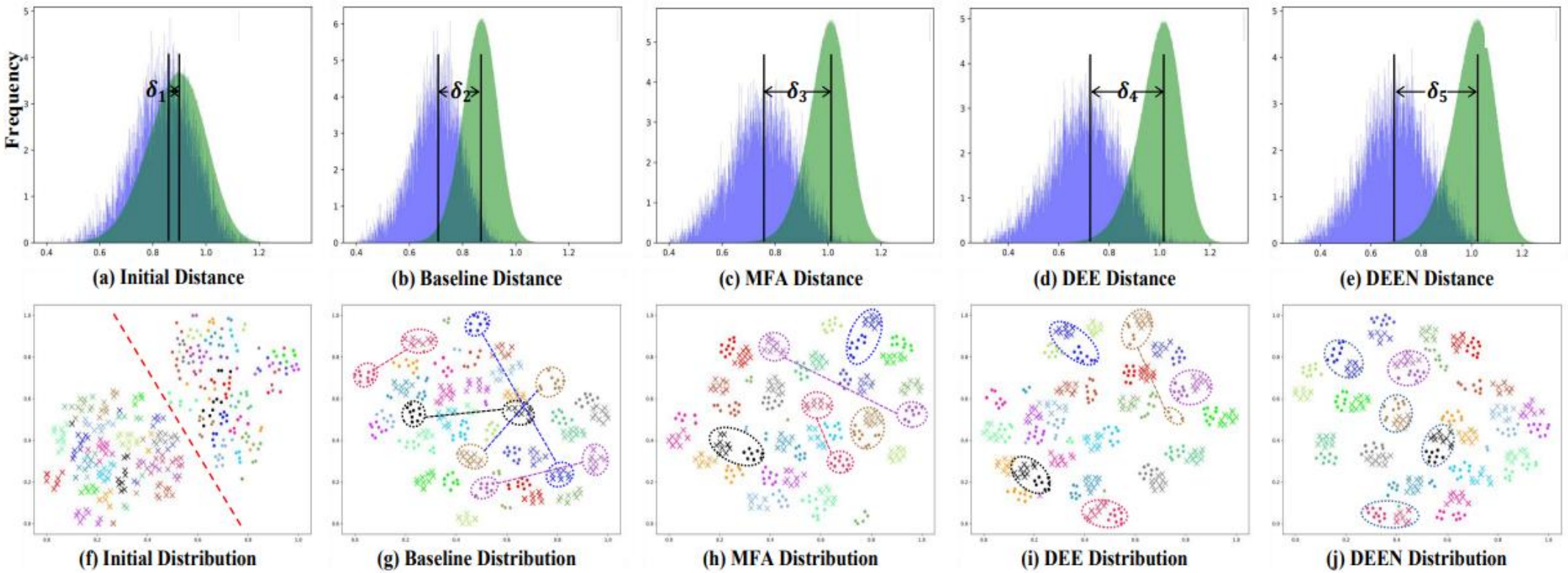

Figure 5: (a-e) show the intra-class and inter-class distances of cross-modality features. The intra-class and inter-class distances are indicated in blue and green colors, respectively. (f-j) show the distribution of feature embeddings in the 2D feature space, where circles and triangles in different colors denote visible and infrared modalities. A total of 20 persons are selected from the test set. The samples with the same color are from the same person. The "dot" and "cross" markers denote the images from the VIS and IR modalities, respectively.

Experiments

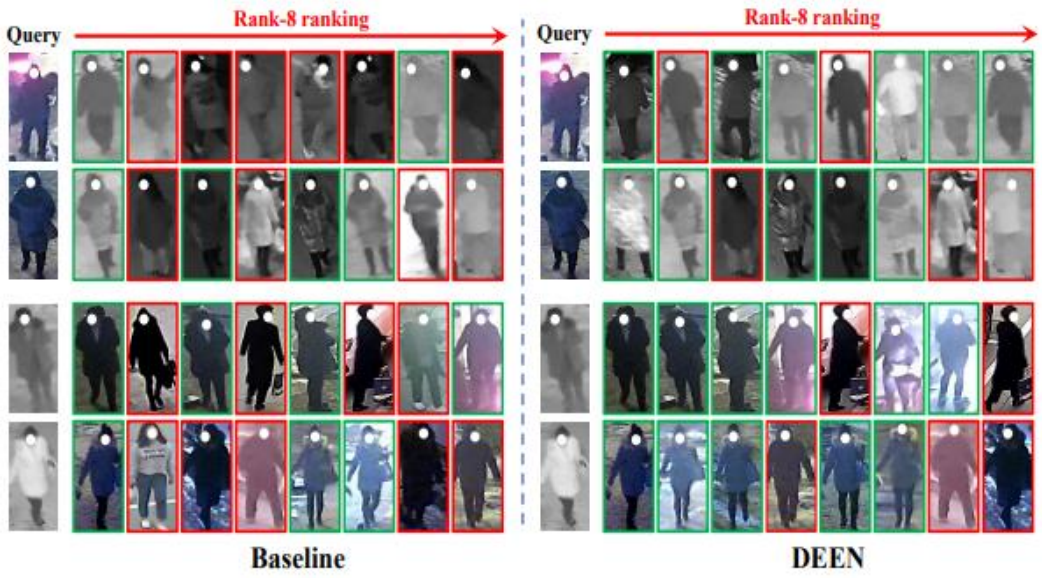

Figure 6: Some Rank-8 retrieval results obtained by the baseline and the proposed DEEN on our LLCM dataset.

Conclusion

- We propose a novel diverse embedding expansion network (DEEN), which can generate diverse embeddings to learn the informative feature representations for reducing the modality discrepancy between the VIS and IR images.
- We provide a challenging low-light cross-modality (LLCM) dataset, which has more new and important features and can further facilitate the research of VIReID towards practical applications.
- Extensive experiments on the SYSU-MM01, RegDB and LLCM datasets show the superiority of the proposed DEEN over several other state-of-the-art methods.

Thanks!

LLCM&DEEN