Supplementary Material of From Laboratory to Real World: A New Benchmark Towards Privacy-Preserved Visible-Infrared Person Re-Identification

Yan Jiang[1,2], Hao Yu[2], Xu Cheng[1]*, Haoyu Chen[2], Zhaodong Sun[1,2], Guoying Zhao[2]
[1] School of Computer Science, Nanjing University of Information Science and Technology
[2] Center for Machine Vision and Signal Analysis, University of Oulu
{jiangyan,xcheng,zhaodong.sun}@nuist.edu.cn
{hao.2.yu,chen.haoyu,guoying.zhao}@oulu.fi

A. Loss Functions

In this section, we will elaborate on the baseline loss functions omitted in the paper, including identity loss $\mathcal{L}_{id}$, circle loss $\mathcal{L}_{cir}$. We also provide details of two versions of our proposed memory rectification bank loss, which were omitted in the manuscript: $\mathcal{L}_{mrb(edu)}$ and $\mathcal{L}_{mrb(mixed)}$.

Identity Loss. The identity loss transforms the retrieval task of cross-modality ReID into an image classifier task, regarding the person ID as the category of the corresponding images. It is defined as follows:

$$\mathcal{L}_{id} = -\frac{1}{N_m}\sum_{i=1}^{N_m}\log P(y_i^m|l_i^m), \tag{1}$$

where N_m denotes the number of pedestrian samples in client m. l_i^m and y_i^m are the pedestrian logit and corresponding label, respectively.

Circle Loss. The purpose of the used circle loss is to reduce intra-class distances and increase inter-class distances, thereby effectively distinguishing different identities. It is defined as follows:

$$\begin{aligned}\mathcal{L}_{cir} = \log[1 + \sum_{i=1}^{N_n^m} exp(\gamma\beta_i^-(z_i^{m-} + \alpha)) \\ \cdot \sum_{j=1}^{N_p^m} exp(\gamma\beta_j^+(-z_j^{m+} + \alpha - 1))],\end{aligned} \tag{2}$$

where $\beta_i^- = [z_t^{m-} + \alpha]_+$ and $\beta_j^+ = [1 + \alpha - z_j^{m+}]_+$. $[\cdot]_+$ is the cut-off at zero operation to ensure β_i^- and β_j^+ are non-negative. z_i^{m-} and z_j^{m+} are the negative and positive samples of the pedestrian in client m. N_n^m and N_p^m denote the number of negative and positive samples, respectively. α and γ are hyperparameters to control the margin and loss scaling, which are set to 0.45 and 64, respectively.

*Corresponding Author

Table I. Ablation studies of circle loss on SYSU-MM01 and LLCM datasets. rank-1, mAP, and mINP are reported.

Settings	SYSU-MM01 [10]			LLCM [14]		
	r=1 ↑	mAP ↑	mINP ↑	r=1 ↑	mAP ↑	mINP ↑
DPPT	**51.27**	**49.29**	**34.47**	**34.69**	**41.91**	**38.48**
$\mathcal{L}_{cir} \rightarrow \mathcal{L}_{tri}$	45.73	41.86	26.19	32.67	39.31	35.98

Circle loss enables the model to focus more quickly on those hard-to-distinguish sample pairs by applying weighted treatment to all positive and negative samples, thereby reducing dependence on sample selection. This makes it more stable and efficient compared to triplet loss which is popular in ReID works. We conduct ablation experiments under the CI protocol on the SYSU-MM01 and LLCM datasets, as shown in Tab. I.

More Details of MRB. In the manuscript, we have discussed the effect of different metrics in MRB, namely cosine similarity ($\mathcal{L}_{mrb(cos)}$), Euclidean distance ($\mathcal{L}_{mrb(edu)}$), and a mixture of both ($\mathcal{L}_{mrb(mixed)}$). The $\mathcal{L}_{mrb(cos)}$ is Eq.6 of the manuscript. The $\mathcal{L}_{mrb(edu)}$ is defined as follows:

$$\mathcal{L}_{mrb(edu)} = \frac{1}{N_m}\sum_{i=1}^{N_m}\|\langle z_i^m\rangle^{(t)} - \langle c_{y_i}^g\rangle^{(t-1)}\|^2, \tag{3}$$

where $\langle z_i^m\rangle^{(t)}$ is the *i-th* l_2-normalized feature embedding in the *t-th* epoch on the *m-th* client. $\langle c_{y_i}^g\rangle^{(t-1)}$ is the global center of identity y_i in the last epoch $t-1$. Similarity, The $\mathcal{L}_{mrb(mixed)}$ is defined as follows:

$$\mathcal{L}_{mrb(mixed)} = 0.5(\mathcal{L}_{mrb(cos)} + \mathcal{L}_{mrb(edu)}). \tag{4}$$

Under the EI protocol, our DPPT employs the mixed MRB. The ablation study for this choice is presented in Tab. II, showing the impact of using the mixed metric on model performance. This experiment highlights how combining cosine similarity and Euclidean distance contributes to improved generalization in the unseen entity.

Table II. Ablation studies of the MRB under EI protocol.

Settings	R [6]+L [14]→S [10]		L [14]+S [10]→R [6]		R [6]+S [10]→L [14]	
	r=1 ↑	mAP ↑	r=1 ↑	mAP ↑	r=1 ↑	mAP ↑
edu	10.49	11.31	18.29	19.64	13.81	18.19
cos	10.72	11.40	21.01	**20.93**	14.30	18.79
mixed	**11.27**	**11.86**	**21.51**	20.72	**14.63**	**19.15**

Table III. The dataset information of three protocols in our L2RW benchmark. ID is the pedestrian numbers in the training sets, and the samples denote the number of training images. Query and gallery are the infrared and visible images in the testing set.

(a) Camera Independence (CI)

Setting	Client	ID	Samples	Query	Gallery
SYSU-MM01 [10]	1	194	4721	3801	301
	2	195	5753		
	3	392	7909		
	4	382	5811		
	5	390	5973		
	6	200	4000		
RegDB [6]	1	206	2060	2060	2060
	2	206	2060		
LLCM [14]	1	21	343	7166	484
	2	20	137		
	3	345	5733		
	4	353	5843		
	5	661	8019		
	6	687	8450		
	7	72	1002		
	8	40	1089		
	9	24	305		

(b) Entity Sharing (ES)

Setting	Client	ID	Samples	Query	Gallery
R [6]+L [14]→S [10]	-	919	35041	3801	301
L [14]+S [10]→R [6]	-	1108	65088	2060	2060
R [6]+S [10]→L [14]	-	601	34577	7166	484

(c) Entity Independence (EI)

Setting	Client	ID	Samples	Query	Gallery
R [6]+L [14]→S	1	206	4120	3801	301
	2	713	30921		
L [14]+S [10]→R	1	713	30921	2060	2060
	2	395	34167		
R [6]+S [10]→L [14]	1	206	4120	7166	484
	2	395	34167		

B. Experiment Details

B.1. Datasets

We provide details on the training set, testing set, and client information under our designed protocols in L2RW, as shown in Tab.III. It is noted that the model adopts centralized training under the CS protocol, so there is no client information.

B.2. Training Details

The training settings under the three protocols in L2RW are shown in Tab. IV. In the CI protocol, the training set is further divided by cameras. In the ES protocol, the training sets of two datasets are directly concatenated. In the EI protocol, the training sets of two datasets are processed by two independent clients without data sharing. The test sets for all three protocols remain consistent with the official splits. We used a OneCycle learning scheduler, where the learning rate changes with each iteration, as shown in Fig. I.

Table IV. The training details of our proposed L2RW, CA is the channel augmentation.

Setting	CI	EN/ES
Total epochs	50	30
Batch size	64	64
Image size	288x144	288x144
Augmentation	RandomCrop RandomHorizontalFlip CA [11]	RandomCrop RandomHorizontalFlip CA [11]
Optimizer	SGD	SGD
LR	0.2	0.2
LR Scheduler	OneCycle	OneCycle
Weight decay	$5e^{-4}$	$5e^{-4}$
Momentum	0.9	0.9

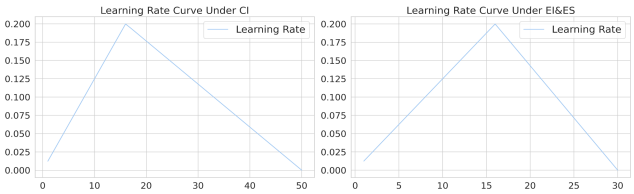
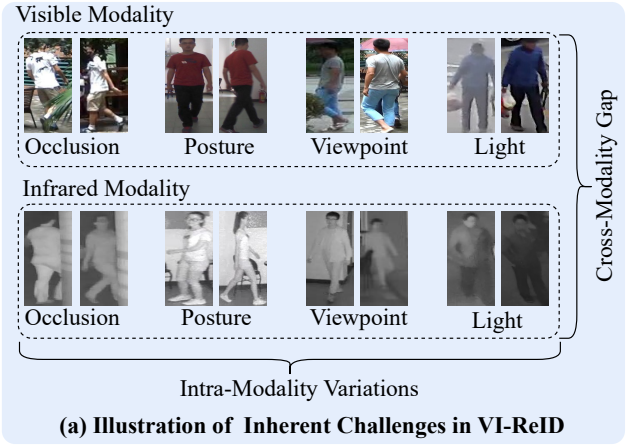


Figure I. Learning rate setting under our three protocols.



(a) Illustration of Inherent Challenges in VI-ReID

(b) Illustration of Challenges in Privacy-Preserved VI-ReID

Figure II. Illustration of challenges of (a) VI-ReID and (b) privacy-preserved VI-ReID in our L2RW benchmark.

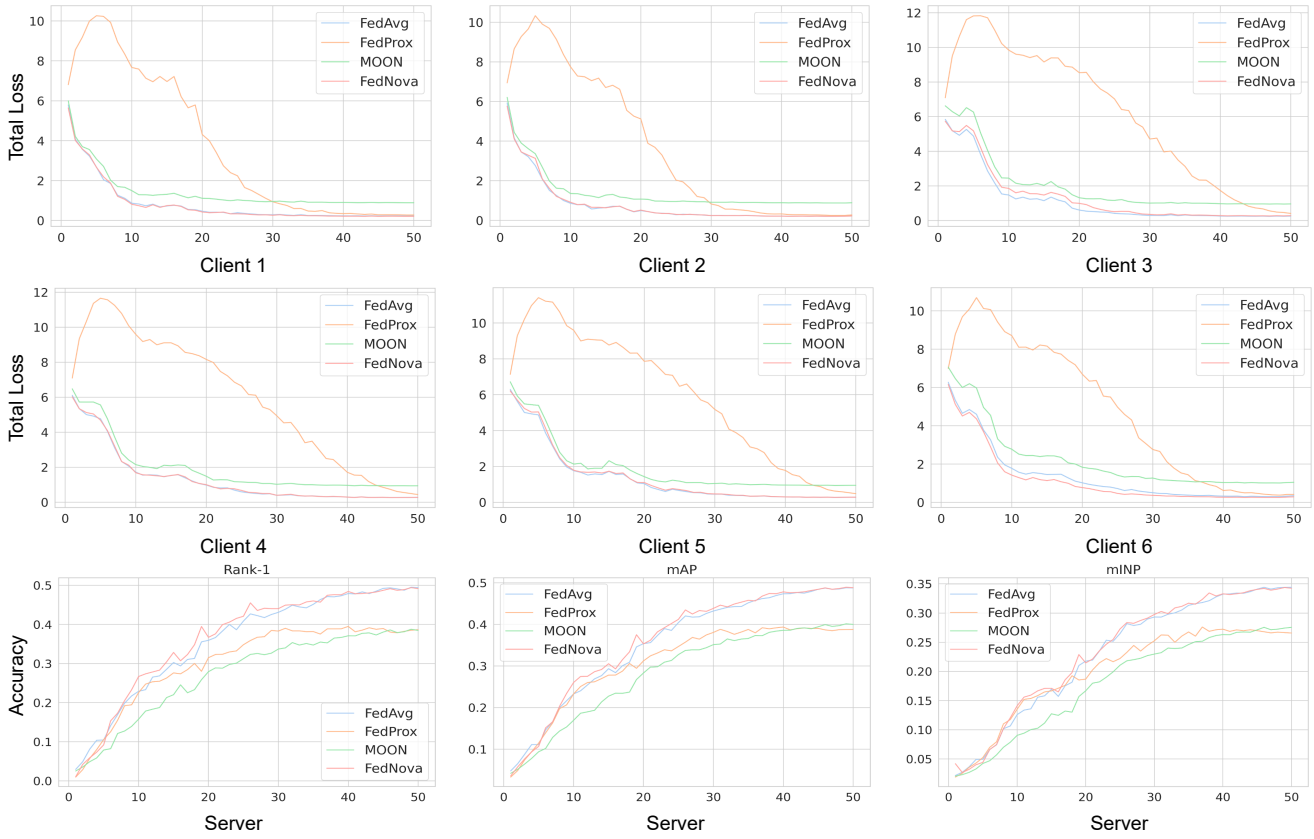


Figure III. Training details of four federated learning algorithms on the SYSU-MM01 dataset. The total loss of clients and the accuracy of the server are shown.

B.3. Problem Illustration

As shown in Fig. II (a), VI-ReID is more challenging due to the large discrepancy between visible and infrared modalities. To be specific, visible images have three channels and contain abundant color information, while infrared images only have one channel of invisible electromagnetic radiation. This leads to the lack of color information in infrared images, making it challenging (even for humans) to distinguish identities between visible and infrared modalities. In addition, VI-ReID encounters intra-modality variations in posture, viewpoint, light, etc., making VI-ReID more challenging compared to traditional single-modality ReID task.

We also show the challenges encountered with privacy-preserved VI-ReID, i.e., modality incomplete, identity missing, and domain shift, as shown in Fig. II (b). Specifically, the modalities available on different clients may vary, with some having only RGB, others only IR, or both, namely *modality incomplete*. Some pedestrians are not captured by specific cameras, resulting in the absence of their images on the corresponding clients. This is referred to as *identity missing*. Moreover, the intra-modality variations or cross-modality gap across clients inevitably bring *domain shift*.

B.4. Reproduced Method

CI. In the CI protocol, we reproduced the two VI-ReID methods, i.e., AGW [12] and DNS [1]. Note that these two methods cannot directly be applied to the CI protocol as their frameworks are designed for data-shared learning. So we modify their two-stream architecture into one-stream architecture and adopt our sampling strategy. Moreover, the module or loss that needs modality information is removed. Specifically, the AGW† is trained by the ResNet-50 with Non-local modules, supervised by identity loss and weighted regularization triplet loss. The DNS† is trained by the ResNet-50 with heterogeneous space shifting (HSS) modules, supervised by the identity loss and circle loss.

ES. In the ES protocol, existing VI-ReID methods can be seamlessly applied. Therefore, apart from the training epoch, all other settings in our reported methods [1, 7, 11, 12, 14] in EI remain consistent with the official configurations.

Federated Learning Algorithm. In the CI, we reproduce four federated learning algorithms, including FedAvg [5], FedProx [4], Moon [3], and Fednova [9]. Different fed-

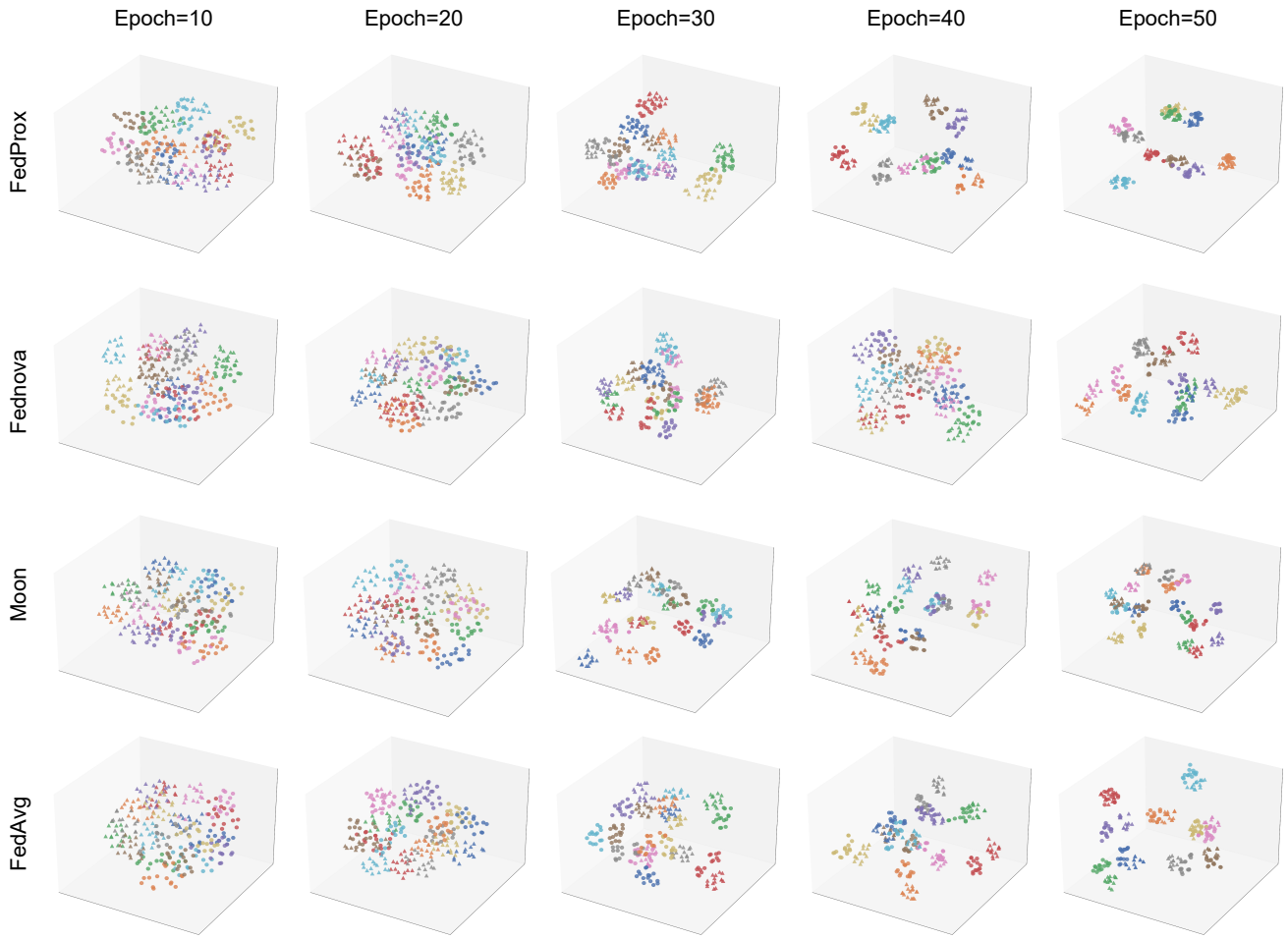


Figure IV. t-SNE visualization of four federated algorithms during training on the SYSU-MM01 dataset under the CI protocol.

erated learning algorithms only change step three ③ and step four ④. We present the information of clients and the server for different federated learning algorithms on the SYSU dataset under the CI protocol, as shown in Fig. III.

To further investigate their roles in VI-ReID, we adopt t-SNE visualizations of the training processes for four algorithms, as shown in Fig. IV. During the training process, modality differences gradually decrease across all methods. Although FedProx [4] significantly reduces modality differences by the end, it struggles to effectively distinguish between different identities, making it challenging for the model to differentiate between pedestrians. Fednova [9] does not effectively reduce modality differences but achieves slightly better identity differentiation compared to FedProx. Moon [3] performs well in distinguishing pedestrians within the same modality but faces challenges in handling modality differences and inter-class issues. Finally, FedAvg [5] outperforms the other three algorithms but still exhibits noticeable modality differences and struggles with distinguishing certain identities. These observations highlight the limitations of current federated learning algorithms in cross-modality scenarios, making this an important topic for future research.

Table V. Efficiency comparison of reported methods.

Method	Throughput (bps) ↑	Lantency (ms) ↓	Method	Throughput (bps) ↑	Lantency (ms) ↓
DEEN [14]	425.58	112.79	AGW [12]	1174.08	40.88
LBA [7]	634.99	50.40	CAJ [11]	1184.36	54.04
DNS [1]	909.51	52.78	Ours	**2528.57**	**18.98**

B.5. Effiency Analysis

To show the effiency of decentralized training, we have evaluated the throughput and latency of all methods under ES and EI protocols on a RTX 3090 GPU, as shown in Tab. V. It is evident that our proposed DPPT achieves the highest throughput and the lowest latency, demonstrating its effectiveness.

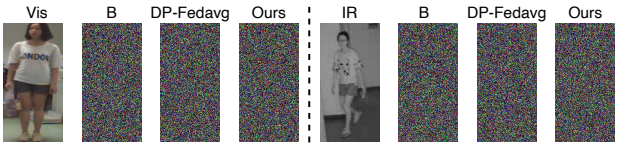


Figure V. Visualization of gradient inversion.

C. Discussion

C.1. Decentralized Training

Some existing methods [8, 15, 16] leverage decentralized training to address single-modality ReID tasks in a privacy-preserved way. We will discuss the distinctions between our L2RW and theirs from two perspectives: technical challenges and social benefits. **Technical Challenges:** Unlike traditional single-modality ReID, VI-ReID faces significant cross-modality discrepancies, exacerbating the domain shift problem between clients. As shown in the visualization in Fig. IV, while visible modality features (circle) exhibit minimal intra-class distances, the cross-modality distances remain substantial. This fundamental difference and additional difficulty underscore the unique challenges of VI-ReID compared to existing single-modality ReID approaches. **Social Benefits:** VI-ReID ensures all-day retrieval capability, whereas traditional single-modality ReID operates only under favorable lighting conditions. From the perspective of contributing to intelligent surveillance systems, VI-ReID offers greater societal value and is more deserving of research focus.

Table VI. Evaluation on various attacks.

Attack	r=1 ↑	r=10 ↑	mAP ↑	mINP ↑
image(100%)	46.36$^{\downarrow 4.91}$	87.32$^{\downarrow 1.23}$	46.53$^{\downarrow 2.76}$	33.21$^{\downarrow 1.26}$
gradient(33%)	42.82$^{\downarrow 8.45}$	84.99$^{\downarrow 3.56}$	42.57$^{\downarrow 6.72}$	29.10$^{\downarrow 5.37}$
gradient(66%)	32.20$^{\downarrow 19.07}$	75.00$^{\downarrow 13.55}$	32.74$^{\downarrow 16.55}$	20.95$^{\downarrow 13.52}$
gradient(83%)	24.00$^{\downarrow 27.27}$	65.35$^{\downarrow 23.20}$	25.32$^{\downarrow 23.97}$	15.12$^{\downarrow 19.35}$
gradient(100%)	4.90$^{\downarrow 46.37}$	22.76$^{\downarrow 65.79}$	6.10$^{\downarrow 43.19}$	3.16$^{\downarrow 31.31}$

C.2. Privacy Security

To discuss the privacy protection of our proposed DPPT, we evaluated it under the CI protocol on the SYSU-MM01 dataset with various attacks, as shown in Tab. VI. Under image attacks (adding Gaussian noise), our method maintained good performance. For gradient attacks, we multiplied client gradients by a random factor in [0.5, 2]. Even with 83% gradient attack, our method achieved 24% Rank-1. Performance only degraded significantly when all client gradients were attacked, highlighting its robustness and privacy-preserving properties. To further validate the privacy of our method, we assumed gradient leakage and reconstructed pedestrian images by using gradient inversion, as shown in Fig. V. The reconstructed results revealed no identifiable pedestrian information to the naked eye, demonstrating that our method effectively preserves privacy.

C.3. Potential Privacy-Preserved Way

It is worth noting that decentralized training is not the only approach to ensuring privacy protection. Anonymizing surveillance images is another viable solution [13]. However, this approach still requires transmitting the anonymized images to a central server for training, which imposes significant demands on both time and storage resources. In contrast, decentralized training offers a more efficient way to address these challenges.

C.4. Generalization in Unseen Domain

To evaluate the generalization of current VI-ReID methods on unseen domains, we randomly sample 10 identities from three datasets under the L+S→R setting and visualize feature distributions using t-SNE, shown in Fig. VI. It shows that while these methods alleviate the cross-modality gap in the seen domain, whether the source domain is shared (ES) or independent (EI), they fail to do so for unseen domains. This highlights a heavy reliance on visible knowledge and a lack of adaptability to unseen domains, severely limiting the real-world deployment of VI-ReID. Therefore, we argue that *enhancing the generalization ability of VI-ReID methods to adapt to unseen environments is a pressing issue that requires immediate attention.*

D. Limitations and Future Works

D.1. Limitations

This work revisits VI-ReID and introduces a decentralized training approach for privacy-preserved VI-ReID. However, under the CI protocol, our method still exhibits a performance gap compared to methods with fully shared data [1, 2, 14]. Furthermore, the overall rank-1 accuracy remains relatively low under both the ES and EI protocols, indicating that current methods, whether centralized or decentralized, struggle to handle unseen domains effectively.

D.2. Future Works

In future work, we will explore methods to enhance the model's ability to address cross-modal discrepancies under the CI protocol and improve the generalization of VI-ReID to unseen domains. Additionally, we plan to investigate more complex scenarios, such as designing client-specific models based on data volume to minimize resource waste. Furthermore, we aim to explore the feasibility of achieving privacy-preserved VI-ReID without label information.

E. Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. U22B2062), the Research Council of Finland (former Academy of Finland)

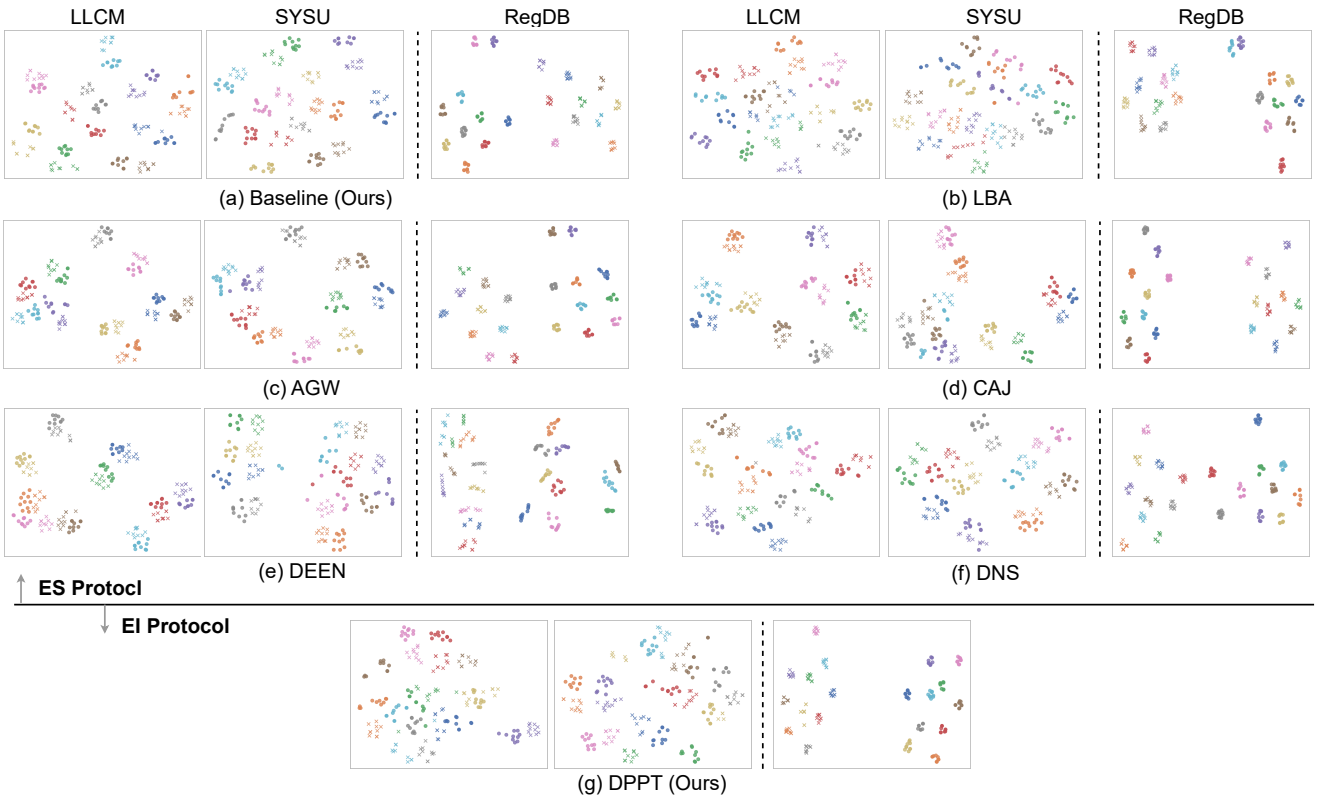


Figure VI. t-SNE visualization of reported methods during training on the SYSU-MM01 dataset under the **L+S→R** setting. Each color denotes an identity. The upper part is under the ES protocol and the lower is under the EI protocol. The circle and cross mask represent the visible and infrared features, respectively.

Academy Professor project EmotionAI (grants 336116, 345122, 359854), the University of Oulu & Research Council of Finland Profi 7 (grant 352788), EU HORIZON-MSCA-SE-2022 project ACMod (grant 101130271), the Finnish Doctoral Program Network in Artificial Intelligence, AI-DOC (decision number VN/3137/2024-OKM-6), and the Startup Foundation for Introducing Talent of NUIST.

References

[1] Yan Jiang, Xu Cheng, Hao Yu, Xingyu Liu, Haoyu Chen, and Guoying Zhao. Domain shifting: A generalized solution for heterogeneous cross-modality person re-identification. In *European Conference on Computer Vision*, pages 289–306. Springer, 2025. 3, 4, 5

[2] Yan Jiang, Xu Cheng, Hao Yu, Xingyu Liu, Haoyu Chen, and Guoying Zhao. Dsaf: Dual space alignment framework for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 2025. 5

[3] Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10713–10722, 2021. 3, 4

[4] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 3, 4

[5] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 3, 4

[6] 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

[7] 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. 3, 4

[8] Chunli Song, Xiaohua Chen, Wenqiu Zhu, Yucan Zhou, Xiaoyan Gu, and Bo Li. Meta-knowledge enhanced data augmentation for federated person re-identification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8901–8905. IEEE, 2024. 5

[9] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020. 3, 4

[10] 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

[11] 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. 2, 3, 4

[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. 3, 4

[13] Mang Ye, Wei Shen, Junwu Zhang, Yao Yang, and Bo Du. Securereid: Privacy-preserving anonymization for person re-identification. *IEEE Transactions on Information Forensics and Security*, 2024. 5

[14] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2153–2162, 2023. 1, 2, 3, 4, 5

[15] Weiming Zhuang, Yonggang Wen, Xuesen Zhang, Xin Gan, Daiying Yin, Dongzhan Zhou, Shuai Zhang, and Shuai Yi. Performance optimization of federated person re-identification via benchmark analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 955–963, 2020. 5

[16] Weiming Zhuang, Xin Gan, Yonggang Wen, and Shuai Zhang. Optimizing performance of federated person re-identification: Benchmarking and analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(1s):1–18, 2023. 5