Supplementary material on

Pose-guided Enriched Feature Learning for Federated-by-camera Person Re-identification

A. Overall Training Procedure of Pose-guided Enriched Feature Learning

As described in the main paper, each client sequentially performs three stages: local-expert model training, pose-extraction module training, and local-global model training. The overall training procedure of the proposed method is summarized in Algorithm 1.

Algorithm 1: Pose-guided Enriched Feature Learning (Round t, c-th client)

Input: Local dataset $\mathcal{D}_c$, local-expert model ϕ_c^e, local-global model ϕ_c^g, PEM π_c, pose detector φ.
Output: Updated ϕ_c^g, π_c.
for $t = 1$ *to* T **do**
  $\theta_{\phi_c^g}^{(t)}$ and $\theta_{\pi_c}^{(t)}$ broadcasted from the server.
  // 1. Local-expert model training
  Train ϕ_c^e on $\mathcal{D}_c$ using $\mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{tri}}$.
  // 2. Pose-Extraction Module Training
  Freeze ϕ_c^e, ϕ_c^g;
  for $B \sim \mathcal{D}_c$ **do**
    Obtain oracle pose $\boldsymbol{z}_i = \varphi(\boldsymbol{x}_i)$;
    Compute $\boldsymbol{S}_{\boldsymbol{z}}$ using $\{\boldsymbol{z}_i\}_{i=1}^{|B|}$;
    for $\phi \in \{\phi_c^e, \phi_c^g\}$ **do**
      $\boldsymbol{F}_i, \boldsymbol{F}_{\delta(i)} = \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_{\delta(i)})$;
      $(\boldsymbol{F}_i^+, \boldsymbol{F}_i^-) = \pi_c(\boldsymbol{F}_i)$;
      $(\boldsymbol{F}_{\delta(i)}^+, \boldsymbol{F}_{\delta(i)}^-) = \pi_c(\boldsymbol{F}_{\delta(i)})$;
      Compute $\boldsymbol{S}_{\boldsymbol{F}^+}$ using $\{\boldsymbol{F}_i^+\}_{i=1}^{|B|}$;
      Compute $\mathcal{L}_{\text{pose}}$ (Eq. 6) and $\mathcal{L}_{\text{cyc}}$ (Eq. 9);
      Update π_c by $\mathcal{L}_{\text{PEM}} = \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{cyc}}$.
  // 3. Local-Global model training
  Freeze π_c;
  for $B \sim \mathcal{D}_c$ **do**
    $\boldsymbol{F}_i, \boldsymbol{F}_{\delta(i)} = \phi_c^g(\boldsymbol{x}_i), \phi_c^g(\boldsymbol{x}_{\delta(i)})$;
    $(\boldsymbol{F}_i^+, \boldsymbol{F}_i^-) = \pi_c(\boldsymbol{F}_i)$;
    $(\boldsymbol{F}_{\delta(i)}^+, \boldsymbol{F}_{\delta(i)}^-) = \pi_c(\boldsymbol{F}_{\delta(i)})$;
    Synthesize $\boldsymbol{F}_i^{\delta(i)} = \boldsymbol{F}_i^- + \boldsymbol{F}_{\delta(i)}^+$;
    Train ϕ_c^g on $\{\boldsymbol{F}_i, \boldsymbol{F}_i^{\delta(i)}\}_{i=1}^{|B|}$ with $\mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{tri}}$.
return $\theta_{\phi_c^g}, \theta_{\pi_c}$

B. Pose Diversity Analysis

To further strengthen the motivation presented in the main paper, we quantitatively measure the pose diversity of each identity using pose features extracted by an off-the-shelf pose detector. Let $\{||\boldsymbol{z}_i||_2\}_{i=1}^{N_y}$ denote the L2-normalized pose features belonging to identity y. We compute diversity only when $N_y \geq 2$, since identities with a single sample do not provide a meaningful estimate of variation. For each valid identity, we compute the covariance matrix $\mathbf{C}_y$ from $\{||\boldsymbol{z}_i||_2\}_{i=1}^{N_y}$ and define the pose diversity as

$$D_{\text{pose}}(y) = \log\det(\mathbf{C}_y).$$

A larger value (less negative) indicates that the identity is observed under a broader range of poses, whereas more negative values correspond to highly limited pose variation.

We evaluate $D_{\text{pose}}(y)$ in two scenarios: (1) a centralized scenario, where all camera observations of an identity are combined into a single client, and (2) a federated-by-camera scenario, where each client contains images from only a single camera. Figure S-1 compares the resulting distributions of pose diversity. Identities in the centralized scenario exhibit substantially higher pose diversity, while those in the federated-by-camera scenario show much lower values. This restricted diversity limits the availability of hard samples, motivating the need for our mechanism.

C. Robustness Analysis under Occlusion

Since the proposed method relies on the pose detector φ, a natural concern is whether PEM can reliably function when input images contain severe occlusions. To address this, we evaluate on Occluded-Duke [6, 7], a challenging benchmark in which the majority of training images contain partially occluded pedestrians, making accurate pose estimation inherently difficult. As shown in Figure S-2, the proposed method achieves the best performance in terms of both mAP

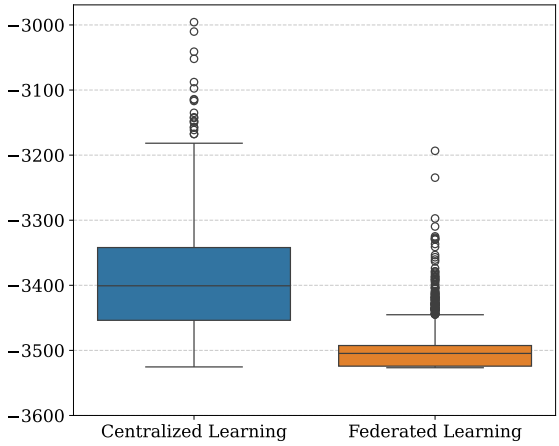


Figure S-1. Comparison of pose diversity between centralized learning and the federated-by-camera scenario. Higher values indicate higher pose diversity across images of the same identity.

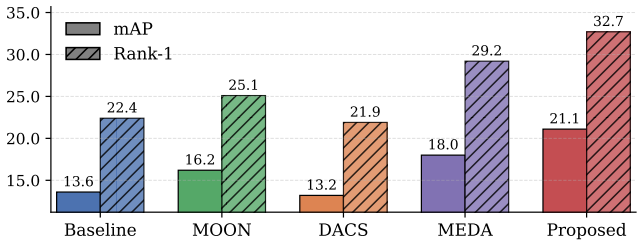


Figure S-2. Performance comparison on Occluded-Duke dataset to verify robustness to occluded training dataset.

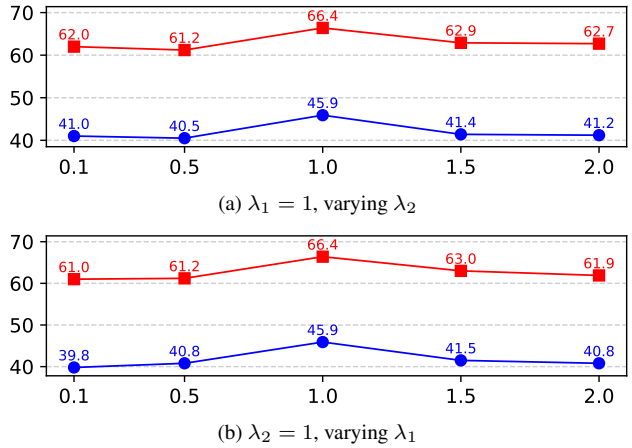
and Rank-1 accuracy, demonstrating its effectiveness even when training samples are heavily occluded.

D. Generalizability Analysis

Previous studies on FedReID [11, 13] have shown that the generalizability of individual clients is closely related to the performance of the aggregated global model. Following this observation, we evaluate whether our proposed method improves both local and global generalization. After the final communication round, we evaluate the local-global model of each client as well as the aggregated global model using the same test set. We report the mean and standard deviation of client performance and compare them with the performance of the global model. The results for Market1501 and MSMT17 are summarized in Table S-1. Our results indicate that the proposed method achieves the highest performance both in terms of average client mAP and global model performance. This suggests that our method successfully improves the generalizability of each client, which in turn leads to a more generalizable global model.

Method	Market1501		MSMT17	
	$\overline{\text{mAP}}^c$	mAP	$\overline{\text{mAP}}^c$	mAP
BDB [1]	30.8 (±1.05)	32.5	9.7 (±1.15)	10.5
ISE [12]	33.0 (±0.83)	34.8	9.3 (±1.11)	10.0
SCAFFOLD [3]	26.6 (±0.71)	28.6	6.8 (±0.69)	7.5
FedProx [5]	28.1 (±1.23)	31.7	9.5 (±1.00)	10.4
MOON [4]	32.2 (±0.95)	33.2	12.2 (±0.69)	13.2
FedRCL [8]	25.6 (±1.51)	29.6	6.1 (±1.11)	7.3
DACS [11]	33.5 (±1.05)	40.0	10.1 (±0.58)	11.4
Proposed	**45.0** (±0.79)	**45.9**	**13.7** (±0.73)	**14.5**

Table S-1. Performance measure between mAP averaged across clients ($\overline{\text{mAP}}^c$) and global (mAP) to validate the effectiveness for model generalizability.



(a) $\lambda_1 = 1$, varying λ_2

(b) $\lambda_2 = 1$, varying λ_1

Figure S-3. Sensitivity analysis of the hyperparameters in $\mathcal{L}_{\text{PEM}}$. The blue curve denotes mAP, and the red curve denotes Rank-1 accuracy.

E. Hyperparameter Analysis

We analyze the sensitivity of the hyperparameters in $\mathcal{L}_{\text{PEM}} = \lambda_1 \mathcal{L}_{\text{pose}} + \lambda_2 \mathcal{L}_{\text{cyc}}$. As shown in Figure S-3, $\lambda_1 = \lambda_2 = 1$ yields the best performance in terms of both mAP and Rank-1 accuracy. When $\mathcal{L}_{\text{pose}}$ dominates, the model enforces pose alignment too strongly, which can distort the semantic consistency of the augmented features and reduce their identity-preserving faithfulness. In contrast, when $\mathcal{L}_{\text{cyc}}$ dominates, the model excessively emphasizes on reconstruction of the original features, making the augmented features less capable of reflecting pose variation and thereby weakening the intended pose guidance. These results indicate that a proper balance between pose guidance and consistent reconstruction is crucial for effective feature augmentation.

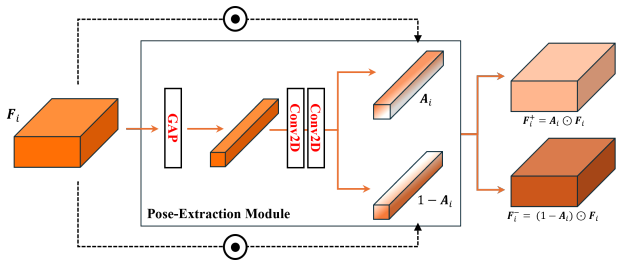


Figure S-4. Architecture of the PEM.

F. More Details on Pose-Extraction Module

Figure S-4 illustrates the architecture of the Pose-Extraction Module (PEM), which follows the channel-gating design introduced in [2]. The PEM takes a feature map as input and predicts a channel-wise gate that measures how much each channel reflects pose-related information. Using a lightweight two-layer convolutional network, the PEM generates an attention vector that decomposes the input feature into pose-relevant and -irrelevant components. This simple gating mechanism enables effective feature disentanglement and supports recombining these components to generate pose-augmented hard samples.

G. Additional T-SNE Visualizations

In the main paper, we visualized the feature distributions of original and augmented features across different communication rounds to demonstrate that the proposed method continuously supplies hard samples during training. In this section, we provide additional T-SNE [10] visualizations on the Market1501 dataset to further analyze how the augmented features distribute in the feature space under the federated-by-camera scenario. As shown in Figure S-5, the augmented features consistently appear along the outer boundary of each identity cluster across all examined rounds (100, 200, and 300). This indicates that the PEM keeps generating samples that differ meaningfully in pose, even in the later stages of training. As a result, the proposed method provides hard samples throughout the 300 communication rounds, helping the model avoid overfitting to each client's limited pose range and supporting stable contrastive learning.

H. Top-10 Ranked Retrieval Visualization

To further investigate whether the proposed method can identify a person regardless of query pose, we visualize the top-10 ranked gallery images for query examples. As shown in Figure S-6, the baseline model (a) retrieves gallery images that closely match the pose of the query, indicating a reliance on pose similarity. In contrast, our method (b) successfully identifies the correct identity regardless of pose, demonstrating robustness to pose discrepancies. These results suggest that the proposed framework enables pose-invariant feature learning, which is crucial for person re-identification.

Dataset	Client ID	# Samples	# IDs
Market1501	Centralized	12,936	751
	Cam-1	2,017	652
	Cam-2	1,709	541
	Cam-3	2,707	694
	Cam-4	920	241
	Cam-5	2,338	576
	Cam-6	3,245	558
MSMT17	Centralized	32621	1,041
	Cam-1	4,910	669
	Cam-2	203	14
	Cam-3	454	81
	Cam-4	1,614	173
	Cam-5	4,296	679
	Cam-6	1,678	207
	Cam-7	3,453	670
	Cam-8	795	137
	Cam-9	1,396	154
	Cam-10	655	58
	Cam-11	3,154	389
	Cam-12	1,364	308
	Cam-13	3,635	379
	Cam-14	3,876	649
	Cam-15	1,138	254

Table S-4. Data statistics of the Market1501 and MSMT17 datasets under the federated-by-camera scenario.

I. Data Statistics

In our federated-by-camera scenario, each camera is treated as a single client. Accordingly, the Market1501 dataset is partitioned into 6 clients, and the MSMT17 dataset into 15 clients. The details of client-wise data statistics for each client are summarized in Table S-4. Unlike standard federated learning benchmarks, where non-IID levels are manually controlled using a Dirichlet distribution, the federated-by-camera scenario naturally results in highly heterogeneous data distributions. Each camera captures different identities under different viewpoints and provides a varying number of samples per identity. As a result, both the class distribution and the number of samples per identity vary widely across clients, creating a realistic and challenging non-IID environment without any manipulation.

References

[1] Zuozhuo Dai, Mingqiang Chen, Xiaodong Gu, Siyu Zhu, and Ping Tan. Batch dropblock network for person re-identification and beyond. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3691–3701, 2019. 2

[2] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3143–3152, 2020. 3

[3] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020. 2

[4] 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. 2

[5] 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. 2

[6] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In *ICCV*, 2019. 1

[7] Jiaxu Miao, Yu Wu, and Yi Yang. Identifying visible parts via pose estimation for occluded person re-identification. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 1

[8] Seonguk Seo, Jinkyu Kim, Geeho Kim, and Bohyung Han. Relaxed contrastive learning for federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12279–12288, 2024. 2

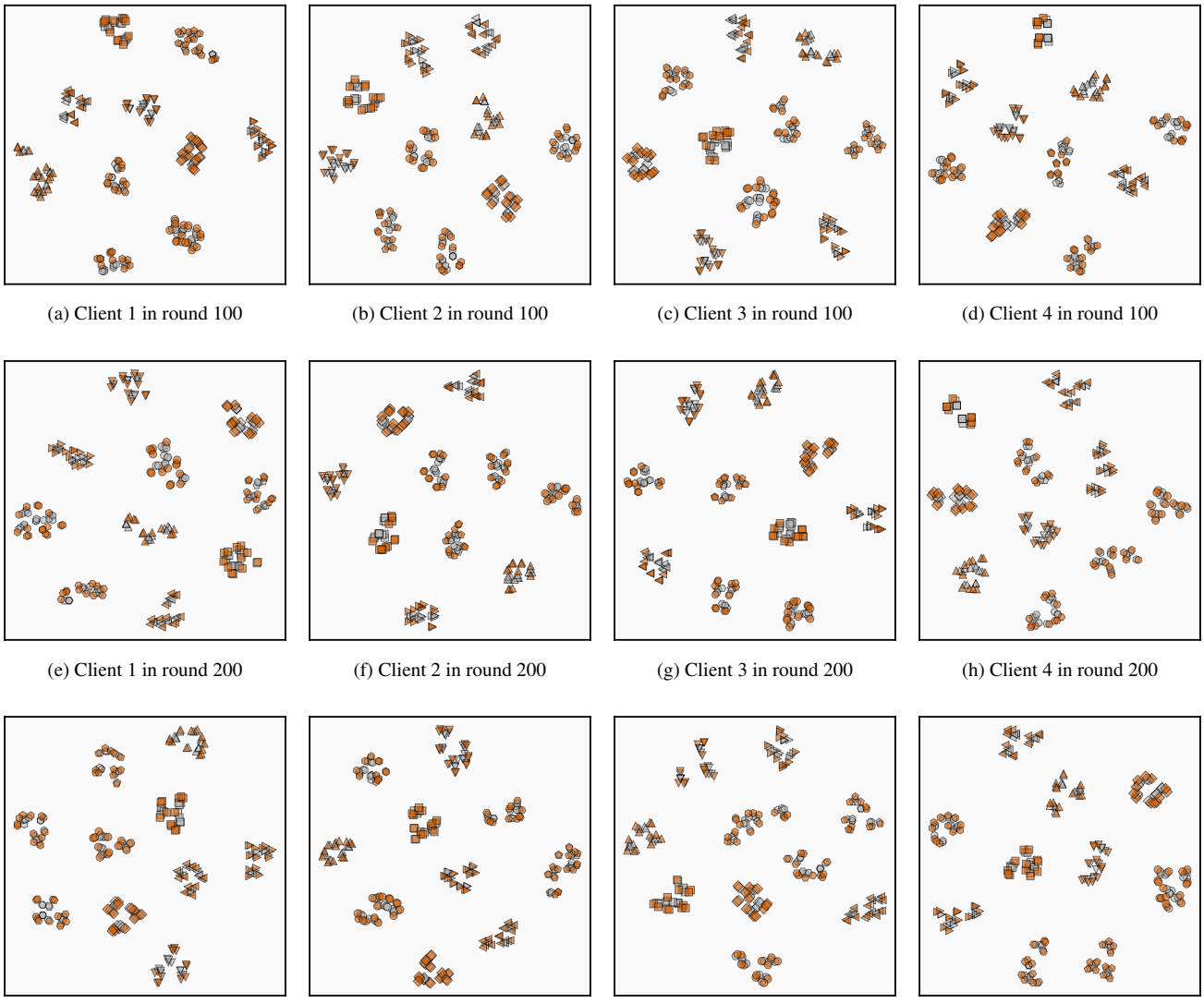
[9] Myungseo Song, Jin-Woo Park, and Jong-Seok Lee. Exploring the camera bias of person re-identification. *arXiv preprint arXiv:2502.10195*, 2025.

[10] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 3

[11] Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yifan He, Shaozi Li, and Nicu Sebe. Diversity-authenticity co-constrained stylization for federated domain generalization in person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6477–6485, 2024. 2

[12] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit sample extension for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7369–7378, 2022. 2

[13] 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. 2

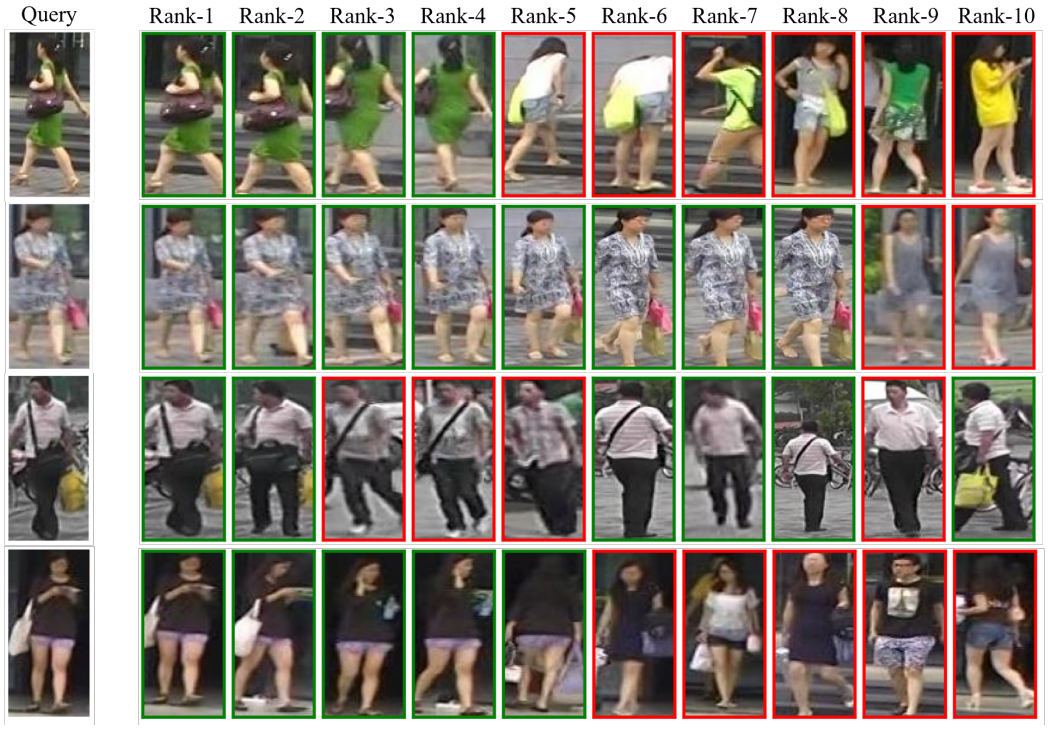


(a) Client 1 in round 100 (b) Client 2 in round 100 (c) Client 3 in round 100 (d) Client 4 in round 100

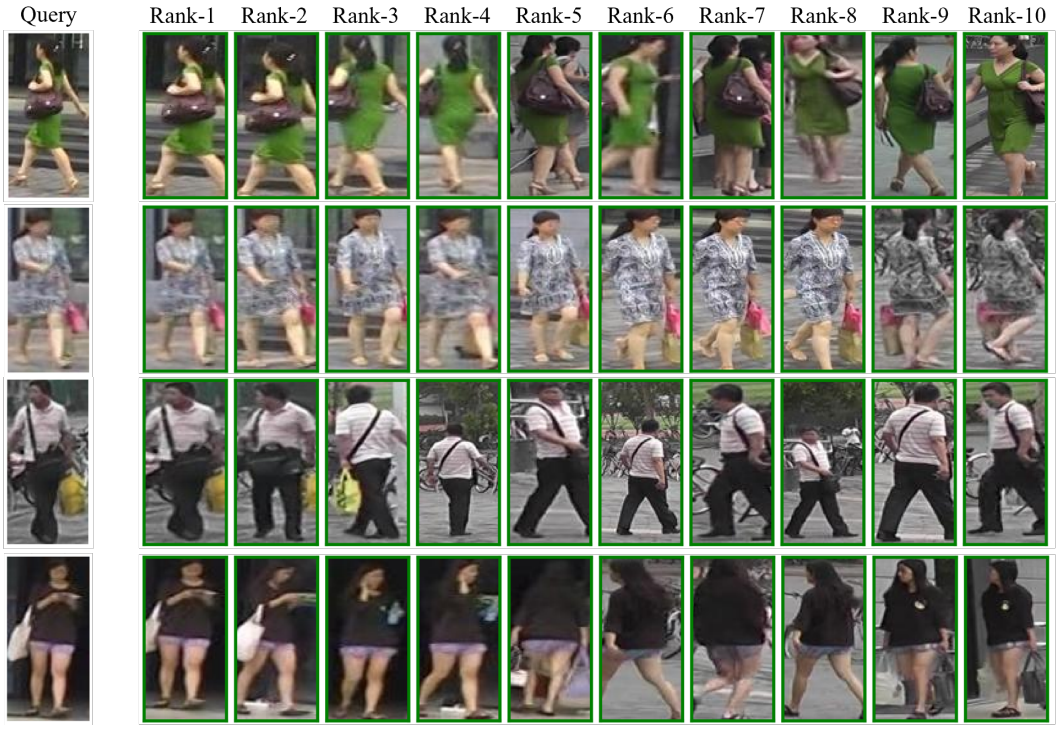
(e) Client 1 in round 200 (f) Client 2 in round 200 (g) Client 3 in round 200 (h) Client 4 in round 200

(i) Client 1 in round 300 (j) Client 2 in round 300 (k) Client 3 in round 300 (l) Client 4 in round 300

Figure S-5. Visualization of feature embeddings for original (gray) and augmented (orange) samples from the Market1501 dataset across different clients.



(a) Baseline method



(b) Proposed method with PEM

Figure S-6. Comparison of the Top-10 retrieval results obtained by using (a) the baseline method and (b) the proposed method. Correct matches are shown in green boxes, and incorrect matches are shown in red.