Prototype Topology Consistency for Visible-Infrared Lifelong Person Re-Identification

Tan Shang[1†], Zhijie Lu[1†], Wuxuan Shi[1], He Li[1], Mang Ye[1*]
[1]School of Computer Science, Wuhan University
{shangtan, zhijielu, yemang}@whu.edu.cn

Abstract

Visible-Infrared Lifelong Person Re-Identification (VI-LReID) requires a model to acquire cross-modal person retrieval capabilities from sequentially arriving datasets while retaining retrieval performance on previously learned ones. We consider two challenges in VI-LReID. First, new-task updates can disrupt instance-to-prototype assignment rankings and weaken the retention of learned identity relations. Such disruption should be distinguished from representation changes that accompany adaptation, since a change in feature coordinates alone does not establish retrieval degradation. This motivates preserving assignment relations without explicitly fixing feature coordinates. Second, visible and near-infrared imaging rely on reflected light at different wavelengths, whereas thermal infrared imaging detects emitted radiation. Such heterogeneous cues motivate a consistency constraint on cross-modal inter-class relations. We propose S-PTC (Stage-wise Prototype Topology Consistency), which aligns the current and frozen previous-stage models' soft assignments for current-stage samples over modality-specific prototype banks. This encourages retention of learned identity relations while allowing adaptation. We further propose CM-PTC (Cross-Modal Prototype Topology Consistency), which aligns the inter-class topology matrices of RGB and IR prototypes, mitigating the structural inconsistency rooted in their imaging physics. Both modules require no stored exemplars. Extensive experiments on the VI-LReID benchmark verify the effectiveness and superiority of our approach against state-of-the-art methods. Code is available at https://github.com/Tan-Shang/PTC.

1 Introduction

Some surveillance cameras switch from RGB to infrared (IR) imaging in low-light conditions. Matching identities across RGB and IR images is therefore an important component of around-the-clock surveillance and is studied as **Visible-Infrared Person Re-Identification (VI-ReID)**[1, 18, 24, 26, 45]. In practice, surveillance data accumulates sequentially across scenes and retaining all historical person images is often infeasible due to privacy regulations, motivating **Lifelong Person Re-Identification (LReID)**[2, 13, 20, 3], where a model must continually learn from new scenes without forgetting the old ones. Their intersection, **Visible-Infrared Lifelong Person Re-Identification (VI-LReID)**, further demands that this sequential learning take place across both visible and infrared modalities simultaneously. Figure 1 illustrates the standard VI-LReID evaluation protocol: at each stage, the model is trained on the newly arrived dataset and then evaluated separately by cross-modal retrieval on each dataset encountered so far.

[†]Equal Contribution.
[*]Corresponding Author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

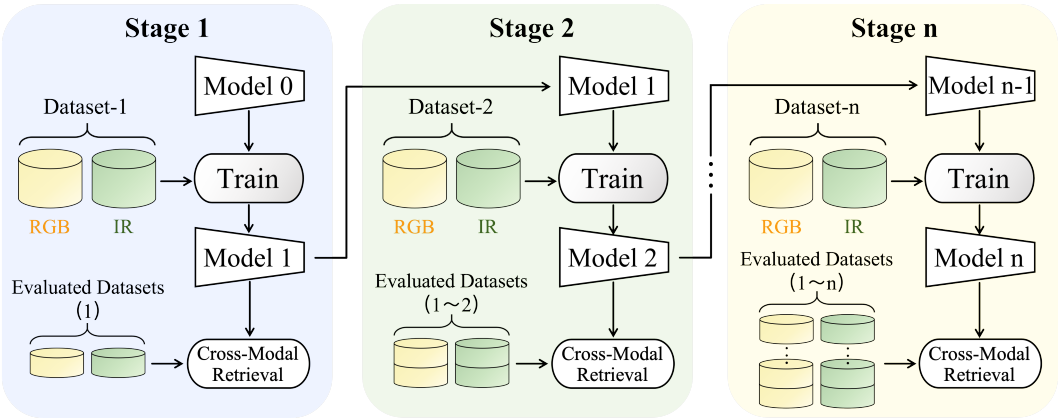


Figure 1: VI-LReID evaluation protocol. At each stage, the model trains only on the newly arrived RGB/IR dataset and is evaluated by cross-modal retrieval on all datasets seen so far.

Existing VI-LReID methods address forgetting by distilling knowledge from the frozen previous-stage model or by rehearsing a small buffer of historical person images [14, 32]. While these designs yield measurable improvements, objectives based on aggregate feature similarity or prediction consistency do not necessarily provide explicit constraints on instance-to-prototype assignment rankings or cross-modal inter-class relations [4].

We examine these two aspects of forgetting under a non-replay protocol. The first is **assignment ranking disruption**. New-task updates to the shared backbone can change samples' relative similarities to previous-stage prototypes and weaken the retention of learned identity relations. This motivates constraining the soft-assignment distributions of current-stage samples while allowing representations to adapt, without explicitly matching feature coordinates. The second concerns **cross-modal inter-class relations** [26]. Visible-light and near-infrared sensors capture reflected light at different wavelengths, whereas thermal infrared sensors detect emitted radiation. These differences in imaging mechanisms can produce modality-dependent identity cues and make cross-modal relations harder to preserve as models adapt to new scenes. The relative similarities among identity classes can consequently differ between modalities. Preserving historical assignments within each modality does not explicitly constrain agreement between these inter-class relations across modalities.

To address these challenges, we propose **Prototype Topology Consistency (PTC)** for VI-LReID. In PTC, prototype topology describes instance-to-prototype affinities and relative similarities among class prototypes. **Stage-wise Prototype Topology Consistency (S-PTC)** promotes retention across stages by constraining assignment relations over historical prototypes. **Cross-Modal Prototype Topology Consistency (CM-PTC)** encourages consistent inter-class relations between in-batch RGB and IR prototypes. Both modules operate on current-batch data only, require no stored exemplars, and are applied as lightweight loss terms on a standard ViT-B/16 backbone without introducing extra learnable parameters.

Our main contributions are:

- We analyze forgetting in VI-LReID from the perspectives of *assignment ranking disruption* and *cross-modal topology inconsistency*, motivating constraints on prototype relations across stages and modalities.
- We propose **S-PTC** to align per-instance soft-assignment distributions over frozen prototype banks and **CM-PTC** to encourage consistent inter-class relations between in-batch RGB and IR prototypes.
- Experiments on four VI-LReID datasets demonstrate that PTC mitigates forgetting and improves retrieval performance. Ablation studies further validate the effectiveness of both modules.

2 Related Work

2.1 Lifelong Person Re-Identification

Lifelong Person Re-Identification (LReID) extends the conventional Re-ID paradigm to a sequential setting, where a model must continuously absorb knowledge from newly arriving deployment scenes while preserving its retrieval ability on all previously encountered scenarios. [13, 3, 19, 44, 23, 27, 9, 43] The fundamental difficulty is catastrophic forgetting: gradient updates that encode new-task representations tend to overwrite the discriminative structure responsible for earlier retrieval tasks [8].

To counter forgetting, *replay-based* methods retain a compact exemplar set drawn from previous tasks and interleave it with incoming data during training. This strategy bounds forgetting by keeping direct evidence of past distributions, at the cost of data storage overhead and serious privacy risks when dealing with biometric surveillance data.[23, 19] *Non-replay methods* instead rely on model-level knowledge transfer: a frozen snapshot of the previous-stage model serves as a teacher, guiding the current model to maintain consistent feature representations or output predictions without accessing stored samples [10, 19, 23, 34]. Recent exemplar-free LReID studies explore knowledge representation and consolidation across sequentially arriving domains. These studies aim to retain previously learned information while acquiring discriminative knowledge from new data, without storing historical image exemplars [33, 35, 36]. However, these methods focus almost exclusively on unimodal visible Re-ID, and their anti-forgetting objectives operate on aggregate feature similarity or per-class mean representations without attending to the structural inter-class ordering that governs retrieval ranking.

2.2 Visible-Infrared Person Re-Identification

Visible-Infrared Person Re-Identification (VI-ReID) aims to retrieve images of the same person across visible and infrared modalities[26]. It supports identity matching when observations are collected by different imaging devices under varying illumination conditions. Visible and near-infrared cameras capture reflected light at different wavelengths, whereas thermal infrared cameras detect emitted radiation. These differences affect the appearance cues available for recognition and introduce cross-modal discrepancies. Existing VI-ReID methods tackle this heterogeneity through either *modality-shared* representation learning, which maps both streams into a joint feature space to align cross-modal identity embeddings [40, 42, 7, 25], or *modality-specific* representation learning, which separately extracts and integrates cues unique to each imaging channel to produce more comprehensive identity descriptors [28, 41, 12, 46]. Beyond supervised settings, unsupervised VI-ReID methods [31, 37, 38] have also made significant progress, leveraging prototype-based contrastive learning [22, 50] or large vision-language model attribute supervision [17] to obtain reliable cross-modal pseudo-labels without manual identity annotations. The robustness of VI-ReID feature representations has also received attention, with feature-level adversarial attacks shown to effectively disrupt retrieval rankings [39], highlighting the sensitivity of cross-modal identity embeddings to structured perturbations.

The growing need for continuous all-day surveillance has motivated an emerging line of work that extends VI-ReID to the sequential learning setting of VI-LReID. TTQK [32] learns a dedicated inference token per deployment scene and matches the most appropriate token for each query at test time, preserving past knowledge through task-specific feature routing. PP-IPG [14] exploits a modality-aware prompt pool to capture instance-level cues from historical stages, enabling fine-grained knowledge recovery without raw data access. CKDA[4] further decomposes cross-modal knowledge into modality-common and modality-specific components through dedicated prompting modules, and aligns both knowledge streams with identity-prototype-based objectives in dual feature subspaces.

Despite these advances, existing VI-LReID methods still face assignment ranking disruption from unstructured gradient updates and asymmetric forgetting from the physical heterogeneity of RGB and IR imaging. We therefore propose S-PTC and CM-PTC to address these two failure modes respectively.

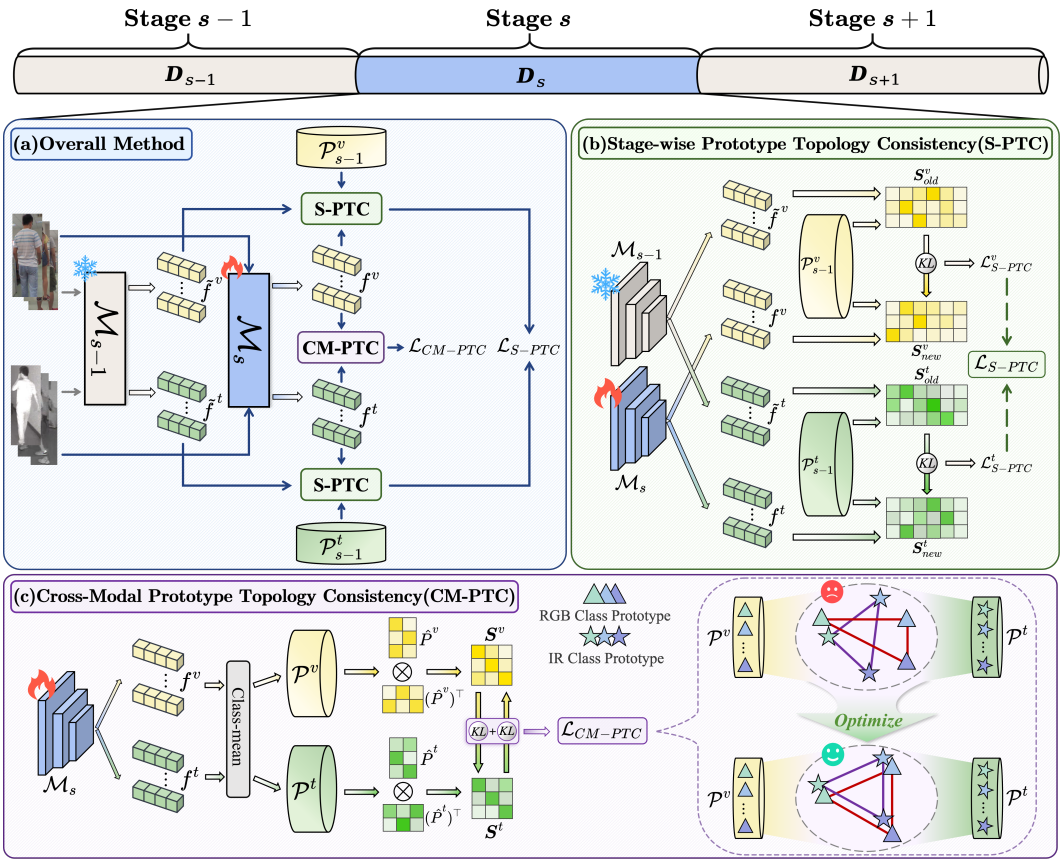


Figure 2: **Overview of our proposed PTC method. (a)** The current model M_s and frozen previous-stage model M_{s-1} encode the same mini-batch. Tildes denote frozen-model features. **(b) S-PTC** matches the two models' soft-assignment distributions over the same frozen previous-stage prototype banks within each modality. **(c) CM-PTC** uses symmetric KL divergence to align the inter-class topology of visible and infrared prototypes computed from the current mini-batch. All KL terms operate on row-wise softmax distributions.

3 Methodology

3.1 Preliminary

Task Formulation. Let $\{\mathcal{D}_1, \ldots, \mathcal{D}_S\}$ be a sequence of S VI-ReID datasets arriving incrementally. Each stage dataset $\mathcal{D}_s = \{(x_i, y_i, m_i)\}$ contains visible (m_i=v) and infrared (m_i=t) images with identity labels y_i shared across the two modalities. Under the non-replay setting adopted in this work, training at stage s uses only images from $\mathcal{D}_s$. Images from previous datasets are not retained. This setting addresses storage and privacy concerns. Current-stage training yields

$$\theta_s^* = \arg\min_{\theta} \mathcal{L}(\theta; \mathcal{D}_s). \tag{1}$$

Here θ denotes the trainable parameters, and θ_s^* is the result before stage-end EMA merging.

Method Overview. As illustrated in Fig. 2a, a shared backbone with trainable parameters θ encodes visible and infrared mini-batches. CM-PTC aligns the inter-class similarity distributions of prototypes computed within each mini-batch. From stage $s \geq 2$, S-PTC additionally matches current-instance soft assignments over the frozen previous-stage prototype banks $\mathcal{P}_{s-1}^v$ and $\mathcal{P}_{s-1}^t$ to those produced by the frozen model θ_{s-1}. These constraints aim to maintain learned identity relations while adapting to the current dataset. After the last training epoch of each stage $s \geq 2$, a one-time EMA parameter merge consolidates the newly learned knowledge into the model:

$$\theta_s = \alpha_s\, \theta_{s-1} + (1 - \alpha_s)\, \theta_s^*, \qquad \alpha_s = \frac{s-1}{s}, \tag{2}$$

where θ_{s-1} denotes the consolidated parameters from the previous stage, and θ_s denotes the parameters after the current merge. At the first stage, no merge is applied and $\theta_1 = \theta_1^*$. The coefficient α_s increases with the stage index so that every past stage contributes equal weight to the merged model. We then use the consolidated model θ_s to construct modality-specific prototype banks from $\mathcal{D}_s$ for the next stage.

3.2 Stage-wise Prototype Topology Consistency (S-PTC)

Learning a new task updates the shared model parameters and can alter samples' relative similarities to historical class prototypes. Such changes may weaken learned identity relations, motivating a constraint on prototype assignments during adaptation. We therefore propose **Stage-wise Prototype Topology Consistency (S-PTC)**, which matches the soft-assignment distributions of current-batch instances over frozen previous-stage prototype banks to the reference distributions produced by the previous-stage model (Fig. 2b). No historical exemplars are stored.

As illustrated in Fig. 2a, at the end of stage $s-1$ we use the consolidated model θ_{s-1} to perform a single inference pass over that stage's training data $\mathcal{D}_{s-1}$. We construct modality-specific prototype banks $\mathcal{P}_{s-1}^v$, $\mathcal{P}_{s-1}^t \in \mathbb{R}^{N_{s-1} \times d}$, where d is the feature dimension and N_{s-1} is the number of training identities in $\mathcal{D}_{s-1}$. Let $\mathcal{X}_k^v$ denote the visible training images of identity k in $\mathcal{D}_{s-1}$, and let $\mathbf{g}_i^v$ be the feature extracted from x_i by θ_{s-1}. The visible class mean is

$$\boldsymbol{\mu}_k^v = \frac{1}{|\mathcal{X}_k^v|} \sum_{x_i \in \mathcal{X}_k^v} \frac{\mathbf{g}_i^v}{\|\mathbf{g}_i^v\|_2}. \tag{3}$$

We then ℓ_2-normalise each class mean and stack the resulting vectors as rows of $\mathcal{P}_{s-1}^v$. The infrared bank is constructed analogously. These banks remain fixed throughout stage s. New banks are constructed from $\mathcal{D}_s$ at its completion for use in stage $s+1$. Earlier-stage information is carried forward through the consolidated teacher model.

At each training iteration of stage $s \geq 2$, both models process the same batch from the current-stage dataset $\mathcal{D}_s$. The current model with parameters θ produces visible features $\mathbf{f}_i^v$ and infrared features $\mathbf{f}_i^t$ before ℓ_2 normalisation. The frozen model θ_{s-1} provides the corresponding reference features $\tilde{\mathbf{f}}_i^v$ and $\tilde{\mathbf{f}}_i^t$. S-PTC ℓ_2-normalises these features before computing their soft-assignment distributions over the frozen prototype banks. The per-instance *assignment score vectors* over the frozen bank $\mathcal{P}_{s-1}^v$ at temperature τ_s are (Fig. 2b):

$$\mathbf{s}_{\text{new},i}^v = \text{softmax}\left(\frac{\mathbf{f}_i^v}{\|\mathbf{f}_i^v\|_2} (\mathcal{P}_{s-1}^v)^\top / \tau_s \right), \quad \mathbf{s}_{\text{old},i}^v = \text{softmax}\left(\frac{\tilde{\mathbf{f}}_i^v}{\|\tilde{\mathbf{f}}_i^v\|_2} (\mathcal{P}_{s-1}^v)^\top / \tau_s \right). \tag{4}$$

The historical prototype vectors have unit ℓ_2 norm, so these products give temperature-scaled cosine similarities before softmax. Both $\mathbf{s}_{\text{new},i}^v$ and $\mathbf{s}_{\text{old},i}^v$ lie in $\mathbb{R}^{N_{s-1}}$. The former encodes the current model's soft assignment of instance i over the identities in $\mathcal{D}_{s-1}$; the latter provides the reference assignment from the previous stage. Matching these distributions constrains changes in the current instances' relations to previous-stage identities, providing a reference for retaining learned identity structure. Let B be the number of samples from each modality in the mini-batch. The per-modality loss penalises such deviation:

$$\mathcal{L}_{\text{S-PTC}}^v = \frac{1}{B} \sum_{i=1}^{B} \sum_{k=1}^{N_{s-1}} s_{\text{old},ik}^v \log \frac{s_{\text{old},ik}^v}{s_{\text{new},ik}^v}, \tag{5}$$

where $s_{\text{old},ik}^v$ is the k-th element of $\mathbf{s}_{\text{old},i}^v$, the soft assignment score of instance i to the k-th identity in $\mathcal{D}_{s-1}$ under the frozen model, and $s_{\text{new},ik}^v$ is the corresponding score under the current model. An analogous loss $\mathcal{L}_{\text{S-PTC}}^t$ is applied to $\mathcal{P}_{s-1}^t$, and the final S-PTC objective is:

$$\mathcal{L}_{\text{S-PTC}} = \tfrac{1}{2} \left(\mathcal{L}_{\text{S-PTC}}^v + \mathcal{L}_{\text{S-PTC}}^t \right). \tag{6}$$

3.3 Cross-Modal Prototype Topology Consistency (CM-PTC)

Visible sensors capture reflected light, whereas thermal infrared sensors respond to emitted radiation. Infrared observations can also arise from near-infrared imaging, so modality-dependent appearance

cues vary with the sensing setup. These differences can lead to different inter-class similarity patterns and motivate consistency constraints on cross-modal identity relations. We propose **Cross-Modal Prototype Topology Consistency (CM-PTC)** to align these relations between visible and infrared prototypes within each mini-batch, without storing historical samples (Fig. 2c).

CM-PTC operates entirely on the current mini-batch, constructing modality-specific class-mean prototypes from identities co-present in both modalities without requiring any historical samples. Let $\mathcal{C}_{\text{batch}}$ denote the set of identities appearing in both modalities within the current mini-batch, $|\mathcal{C}_{\text{batch}}|{=}C$. For sample i, $\mathbf{f}_i = \mathbf{f}_i^{m_i}$ denotes the current-model feature before ℓ_2 normalisation. CM-PTC first averages these features within each identity and modality, then ℓ_2-normalises the resulting prototypes. For each identity $c \in \mathcal{C}_{\text{batch}}$, the class-mean prototype for each modality is:

$$\mathbf{p}_c^v = \frac{1}{n_c^v} \sum\nolimits_{i:\, y_i=c,\, m_i=v} \mathbf{f}_i, \qquad \mathbf{p}_c^t = \frac{1}{n_c^t} \sum\nolimits_{i:\, y_i=c,\, m_i=t} \mathbf{f}_i, \tag{7}$$

where n_c^m is the number of samples of identity c in modality $m \in \{v, t\}$ within the current mini-batch. All sums are taken over that mini-batch. We stack prototypes in the same identity order for both modalities to obtain $\mathcal{P}^v$, $\mathcal{P}^t \in \mathbb{R}^{C\times d}$.

Since the goal is to align *relational structure* rather than absolute feature positions, we ℓ_2-normalise each prototype matrix row-wise and compute the $C \times C$ intra-modal inter-class similarity matrices at temperature τ_{cm}:

$$\mathbf{S}^v = \hat{\mathcal{P}}^v(\hat{\mathcal{P}}^v)^\top/\tau_{cm}, \qquad \mathbf{S}^t = \hat{\mathcal{P}}^t(\hat{\mathcal{P}}^t)^\top/\tau_{cm}, \quad \mathbf{S}^v, \mathbf{S}^t \in \mathbb{R}^{C\times C}, \tag{8}$$

where $\hat{\mathcal{P}}^m = \ell_2\text{-norm}(\mathcal{P}^m)$ denotes the row-wise normalised prototype matrix. Each row of $\mathbf{S}^v$ (or $\mathbf{S}^t$) encodes the relational position of one class prototype relative to all others within its modality, capturing the intra-modal identity topology. CM-PTC minimises the symmetric KL divergence between the row-wise softmax distributions of the two topology matrices to encourage consistent inter-class relations among identities in the current mini-batch:

$$\mathcal{L}_{\text{CM-PTC}} = \frac{1}{2C} \sum_{j=1}^{C} \left[\sum_{k=1}^{C} q_{jk}^v \log \frac{q_{jk}^v}{q_{jk}^t} + \sum_{k=1}^{C} q_{jk}^t \log \frac{q_{jk}^t}{q_{jk}^v} \right], \tag{9}$$

where $q_{jk}^m = \text{softmax}(\mathbf{S}_{j,:}^m)_k$, $m \in \{v, t\}$, denotes the probability assigned to prototype k by the row-wise softmax distribution for prototype j. Gradients propagate through both modality-specific distributions, so the two representations are jointly updated under the consistency objective. The constraint regularises cross-modal relations through the shared backbone during sequential training.

3.4 Training Objective

The full training loss at stage s is:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_s \, \mathcal{L}_{\text{S-PTC}} + \lambda_{cm} \, \mathcal{L}_{\text{CM-PTC}}, \tag{10}$$

where the base loss combines standard identity supervision:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Tri}} + \tfrac{1}{2}\mathcal{L}_{\text{cent}} + \tfrac{1}{2}\mathcal{L}_{\text{cross}}. \tag{11}$$

Here $\mathcal{L}_{\text{cent}}$ and $\mathcal{L}_{\text{cross}}$ are the cross-modal centre-triplet and instance-centre-triplet losses [29] that align visible and infrared features in metric space. λ_s and λ_{cm} are balancing hyperparameters controlling the contributions of S-PTC and CM-PTC respectively, and $\mathcal{L}_{\text{S-PTC}}$ is activated only for $s \geq 2$.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on four publicly available VI-ReID benchmarks trained in sequence: **RegDB**[16] (412 identities, 4,120 image pairs, TIR imaging), **SYSU-MM01**[26] (491 identities, ≈303K RGB/NIR images), **LLCM**[47] (713 identities, 46,767 images), and **VCM**[11] (927 identities, 252,679 video frames). Unless otherwise stated, the training order is RegDB→SYSU-MM01→LLCM→VCM. Previous training data are not revisited at any stage. After each stage, we evaluate retrieval performance on all datasets encountered so far.

Table 1: Comparison with state-of-the-art lifelong VI-ReID methods. Red: best result; Blue: second best among non-replay methods.

	Method	Publication	Backbone	RegDB		SYSU-MM01		LLCM		VCM		Average	
				mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1
	JointTrain	—	ViT-B/16	74.8	79.4	55.8	59.4	60.2	56.3	19.1	33.4	52.5	57.1
	SFT	—	ViT-B/16	6.8	5.7	21.0	21.5	26.4	21.3	17.1	29.0	17.8	19.4
Replay	LwF[10]	*T-PAMI* (2017)	ViT-B/16	26.9	26.1	30.9	30.3	39.1	34.5	18.2	30.0	28.8	30.2
Replay	iCaRL[21]	*CVPR* (2017)	ViT-B/16	42.4	41.3	32.4	32.4	41.9	36.8	18.5	30.7	33.8	35.3
Replay	BiC [30]	*CVPR* (2019)	ViT-B/16	40.9	39.3	22.8	21.9	43.3	37.5	16.4	27.4	30.9	31.5
Replay	WA [48]	*CVPR* (2020)	ViT-B/16	44.9	43.9	34.6	34.4	44.1	38.7	18.7	29.3	35.6	36.6
Replay	PTKP [6]	*AAAI* (2022)	ViT-B/16	50.3	47.9	26.6	25.8	43.2	36.5	17.9	27.4	34.5	34.4
Replay	TTQK [32]	*ICMR* (2024)	ViT-B/16	60.3	58.3	33.6	30.9	45.8	39.6	19.4	32.2	39.8	40.2
Replay	PP-IPG [14]	*ICMR* (2025)	ViT-B/16	59.0	58.8	32.7	35.2	44.4	50.8	35.3	23.3	42.8	42.0
Non-Replay	LwF[10]	*T-PAMI* (2017)	ResNet	23.4	23.0	9.5	9.7	17.8	13.6	6.8	11.6	14.4	14.6
Non-Replay	AKA[19]	*CVPR* (2021)	ResNet	26.2	23.9	11.0	10.0	17.2	12.9	7.4	13.0	15.5	15.1
Non-Replay	PatchKD[23]	*MM* (2022)	ResNet	52.2	55.2	9.6	7.9	18.5	13.7	3.9	5.9	21.1	20.8
Non-Replay	LSTKC [34]	*AAAI* (2024)	ResNet	15.9	13.5	15.1	13.4	22.1	17.4	16.0	26.8	17.3	17.8
Non-Replay	DKP [33]	*CVPR* (2024)	ResNet	34.1	33.5	26.4	27.1	32.8	27.9	15.3	26.4	27.2	28.7
Non-Replay	DASK [35]	*AAAI* (2025)	ResNet	17.3	15.4	23.6	22.0	29.1	24.1	18.1	25.9	22.0	21.8
Non-Replay	LSTKC++ [36]	*T-PAMI* (2025)	ResNet	36.1	38.1	35.7	37.0	29.3	24.9	16.3	26.6	29.4	31.6
Non-Replay	LSTKC [34]	*AAAI* (2024)	ViT-B/16	35.0	35.4	32.7	34.3	39.6	35.1	15.1	25.3	30.6	32.5
Non-Replay	DKP [33]	*CVPR* (2024)	ViT-B/16	30.3	32.7	30.9	32.8	37.3	32.2	15.2	26.3	28.4	31.0
Non-Replay	TTQK [32]	*ICMR* (2024)	ViT-B/16	8.1	6.5	20.8	18.6	19.2	14.9	20.8	32.6	17.2	18.2
Non-Replay	DASK [35]	*AAAI* (2025)	ViT-B/16	21.6	25.6	28.1	32.3	39.6	36.2	14.1	27.4	25.9	30.4
Non-Replay	LSTKC++ [36]	*T-PAMI* (2025)	ViT-B/16	19.2	19.0	27.1	30.3	32.9	29.2	16.1	29.2	23.8	26.9
Non-Replay	CKDA [4]	*AAAI* (2026)	ViT-B/16	57.0	60.4	34.7	37.3	37.8	31.5	15.6	28.2	36.3	39.4
	Ours	This Paper (2026)	ViT-B/16	58.86	62.33	41.90	43.60	38.38	31.44	14.85	25.81	38.50	40.79

Metrics. We report mean Average Precision (mAP)[49] and Rank-1 accuracy (R1)[15, 51] per dataset, as well as their arithmetic averages over all four datasets:

$$\text{AVG} = \frac{1}{S}\sum_{s=1}^{S} \text{metric}(\theta_S;\, \mathcal{D}_s^{\text{test}}). \tag{12}$$

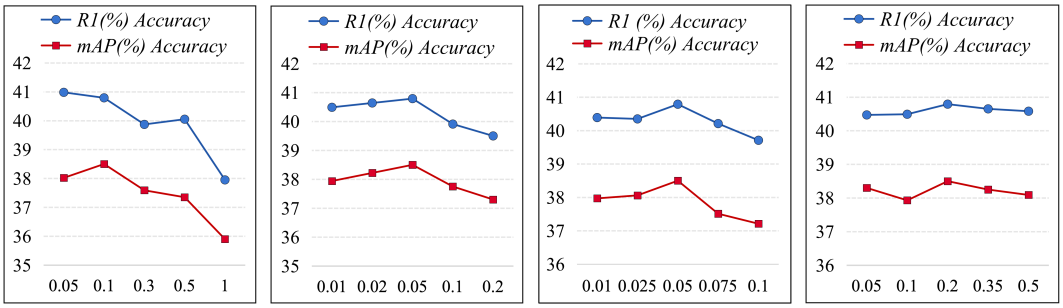
θ_S denotes the consolidated model after all four training stages, and $\mathcal{D}_s^{\text{test}}$ is the test set of dataset s. AVG is computed separately for mAP and Rank-1, with equal weight assigned to each dataset. RegDB is evaluated under the visible-to-infrared protocol; SYSU-MM01, LLCM, and VCM use the infrared-to-visible protocol.

Implementation Details. We adopt ViT-Base/16 [5] initialised with ImageNet-21K weights. Each stage trains for 50 epochs with a batch size of 64, an initial learning rate of 3×10^{-4} with cosine decay, and an image size of 256×128. The stage-end EMA merge uses the task-count-dependent coefficient α_s defined in Eq. 2.

Hyperparameters λ_s=0.1, τ_s=0.05, λ_{cm}=0.05, τ_{cm}=0.2 are fixed across all stages and datasets without dataset-specific tuning. All experiments are conducted on 2× NVIDIA RTX 4090 GPUs with mixed-precision (AMP) training.

4.2 Comparison with State-of-the-Art

As shown in Table 1, our method outperforms the compared non-replay ViT-B/16 methods in average mAP (**38.50**%) and Rank-1 (**40.79**%) across the four datasets. The method combines two topology constraints. S-PTC matches the current model's soft assignments of current-stage samples to those of the frozen previous-stage model over fixed historical prototype banks. CM-PTC encourages consistent inter-class relations between in-batch RGB and IR prototypes. These constraints support identity-structure retention during adaptation to new datasets without storing historical exemplars or introducing additional learnable parameters.



(a) λ_s (b) λ_{cm} (c) τ_s (d) τ_{cm}

Figure 3: Effects of loss weights and softmax temperatures on average mAP and Rank-1.

Random-Seed Variability. We repeat the full four-stage training sequence with random seeds 0, 1, and 2 under the standard order, keeping all other settings fixed. Table 2 reports the final average retrieval scores. Across the three runs, AVG mAP is 38.59 ± 0.41 and AVG Rank-1 is 40.85 ± 0.41 (mean $\pm$ sample standard deviation). These results indicate limited variation across the tested seeds. Table 1 reports the seed-0 result for our method.

Table 2: Results over three random seeds.

Seed	AVG mAP	AVG Rank-1
0	38.50	40.79
1	38.23	40.47
2	39.04	41.29
Mean $\pm$ std	38.59 ± 0.41	40.85 ± 0.41

Robustness to Task Order. We further evaluate task-order robustness under the reversed sequence VCM→LLCM→SYSU-MM01→RegDB. All methods retain the other experimental settings used in the main comparison. Table 3 reports test results after the full four-stage sequence. Our method achieves the highest average mAP (**36.23**%) and Rank-1 (**38.72**%) among the four compared methods. This result shows that its average-performance advantage also holds under a substantially different arrival order.

Table 3: Final results under the reversed task order. Entries show mAP (top) and Rank-1 (bottom).

Method	VCM	LLCM	SYSU-MM01	RegDB	AVG
LSTKC	19.18	30.08	37.20	38.50	31.24
	34.62	24.24	41.78	42.80	35.86
DKP	19.90	31.28	37.89	34.20	30.82
	34.34	24.77	41.36	39.60	35.02
CKDA	20.23	30.78	39.70	49.43	35.04
	33.70	24.25	41.76	51.46	37.79
Ours	20.87	35.04	40.50	48.50	36.23
	33.94	27.94	41.50	51.50	38.72

Table 4: Component and EMA-merge ablation. Results are AVG mAP / AVG R1 over four datasets.

Base	S-PTC	CM-PTC	EMA merge	AVG mAP	AVG R1
✓	–	✓	–	13.72	14.41
✓	✓	✓	–	17.92	18.48
✓	–	–	✓	31.8	33.9
✓	–	✓	✓	36.97	39.08
✓	✓	–	✓	38.18	39.96
✓	✓	✓	✓	**38.50**	**40.79**

4.3 Ablation Study

Table 4 evaluates S-PTC, CM-PTC, and end-of-stage EMA merging. Base denotes the underlying training losses ($\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Tri}} + \mathcal{L}_{\text{cent}} + \mathcal{L}_{\text{cross}}$). The baseline retains EMA merging without either PTC constraint. The final row reports the full method with all components enabled.

Effectiveness of S-PTC. With EMA merging enabled, adding S-PTC to the baseline improves AVG mAP and AVG R1 by **6.38** and **6.06** points, respectively. S-PTC constrains per-instance soft-assignment distributions over historical prototype banks to remain consistent across stages, supporting the retention of learned identity structure.

Effectiveness of CM-PTC. Under the same EMA setting, CM-PTC improves AVG mAP and AVG R1 by **5.17** and **5.18** points over the baseline, respectively. Adding CM-PTC to S-PTC further improves these metrics by **0.32** and **0.83** points. These results support the benefit of aligning in-batch RGB and IR inter-class topology alongside historical assignment preservation.

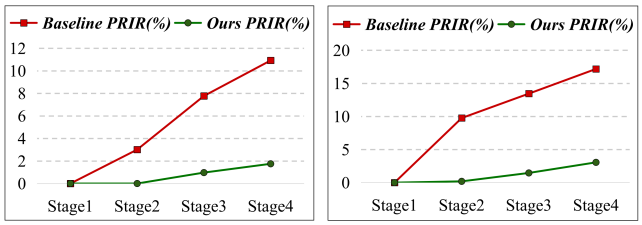
Effectiveness of EMA Merge. Removing end-of-stage EMA merging lowers AVG mAP from 38.50 to 17.92 for the full method and from 36.97 to 13.72 for CM-PTC only. With CM-PTC enabled, S-PTC improves AVG mAP by 4.20 points without merging and by 1.53 points with merging. The large drops without merging show its substantial contribution to final average performance. S-PTC provides additional gains in both settings, indicating that its benefit persists when EMA merging is enabled.

4.4 Hyperparameter Analysis

PTC has two loss weights, λ_s and λ_{cm}, and two softmax temperatures, τ_s and τ_{cm}. Figure 3 shows their effects on average mAP and Rank-1 across the four datasets. For S-PTC, λ_s=0.1 achieves the highest average mAP among the plotted settings. The decline at larger weights is consistent with an overly strong historical-assignment constraint limiting adaptation to new datasets. Performance changes relatively little as τ_s varies from 0.05 to 0.5, suggesting that the constraint remains useful across this range. For CM-PTC, λ_{cm}=0.05 and τ_{cm}=0.2 give the highest average mAP and Rank-1 on their respective curves. The lower scores at other settings suggest that cross-modal topology alignment benefits from a suitable balance of constraint strength and distribution sharpness. We use λ_s=0.1, τ_s=0.05, λ_{cm}=0.05, and τ_{cm}=0.2 for the main experiments.

4.5 Diagnostic Analysis

Prototype Ranking-Inversion Rate (PRIR). We use PRIR to quantify the retention of historical prototype assignments. For each configuration and modality, we evaluate RegDB training samples offline against the corresponding prototype bank fixed at S1. A sample is correctly assigned when its true-class prototype has the highest cosine similarity to its feature. PRIR is the fraction of samples correctly assigned at S1 that become incorrectly assigned at a later stage. The fixed S1 bank serves only as a diagnostic reference; the banks used by S-PTC during training are updated at stage boundaries. Figure 4 compares CM-PTC only with the full method. At S4, adding S-PTC reduces VIS PRIR from 10.92% to 1.75% and IR PRIR from 17.18% to 3.06%, indicating better retention of historical assignments. S-PTC constrains prototype-assignment distributions without explicitly matching feature coordinates.



(a) Visible (VIS) (b) Infrared (IR)

Figure 4: PRIR of RegDB training samples across stages, measured against the fixed S1 prototype bank for each modality. The curves compare CM-PTC only (baseline) with the full method (S-PTC + CM-PTC). Lower values indicate better retention of historical prototype assignments.

Association with Retrieval Performance. We relate PRIR to retrieval performance at nine configuration-stage checkpoints from CM-PTC only, S-PTC only, and the full method at S2–S4. At each checkpoint, training-set PRIR is paired with RegDB test-set mAP and Rank-1. Table 5 reports strong negative Pearson and Spearman correlations across these checkpoints for both modalities. The association supports PRIR as a diagnostic of historical assignment retention relevant to test retrieval performance.

Table 5: Correlation between PRIR and RegDB test retrieval.

PRIR modality	Test metric	Pearson r	Spearman ρ
IR	mAP	−0.940	−0.967
IR	Rank-1	−0.941	−0.950
VIS	mAP	−0.892	−0.979
VIS	Rank-1	−0.897	−0.996

TIR/NIR Forgetting Patterns. We examine forgetting under two training orders using the full method. Under the standard order RegDB→SYSU-

MM01→LLCM→VCM, RegDB (TIR) loses **17.12 mAP** points from S1 to S4. Under SYSU-MM01→LLCM→VCM→RegDB, the three historical NIR datasets lose an average of **1.04 mAP** points by the end of training. Each loss is measured from the completion of training on the corresponding dataset to the end of the sequence, including the final RegDB stage in the second order. These contrasting patterns highlight the relevance of task and imaging heterogeneity to forgetting and motivate attention to cross-modal structure.

5 Broader Impact

This work advances lifelong person re-identification under cross-modal surveillance conditions, with potential societal benefits including enhanced around-the-clock security monitoring and improved missing-person search under low-light or nighttime conditions. As with all person re-identification research, responsible deployment requires compliance with applicable privacy regulations and institutional oversight to prevent misuse.

6 Conclusion

We identify two orthogonal causes of forgetting in VI-LReID: assignment ranking disruption over historical prototypes and cross-modal topology inconsistency induced by heterogeneous imaging physics. To address them, we propose S-PTC, which preserves per-instance assignment distributions over frozen prototype banks to maintain retrieval rankings, and CM-PTC, which aligns inter-class topology between RGB and IR modalities. Both modules require no stored exemplars and introduce no additional parameters. Extensive experiments demonstrate that enforcing prototype topology consistency effectively mitigates forgetting and establishes new state-of-the-art performance among non-replay methods.

Limitations. Our evaluation is conducted on the standard Visible–Infrared LReID benchmark; generalisation to other cross-modal settings (*e.g.*, RGB-Depth) remains to be explored. Repeated-run evaluation is limited to the full method under the standard training order. Variability across seeds for the ablations and other task orders remains to be evaluated.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant No. 62506269.

References

[1] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[2] Zhenyu Cui, Jiahuan Zhou, Xun Wang, Manyu Zhu, and Yuxin Peng. Learning continual compatible representation for re-indexing free lifelong person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16614–16623, 2024.

[3] Zhenyu Cui, Jiahuan Zhou, and Yuxin Peng. Dkc: Differentiated knowledge consolidation for cloth-hybrid lifelong person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3573–3582, 2025.

[4] Zhenyu Cui, Jiahuan Zhou, and Yuxin Peng. Ckda: Cross-modality knowledge disentanglement and alignment for visible-infrared lifelong person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 3452–3460, 2026.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[6] Wenhang Ge, Junlong Du, Ancong Wu, Yuqiao Xian, Ke Yan, Feiyue Huang, and Wei-Shi Zheng. Lifelong person re-identification by pseudo task knowledge preservation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 688–696, 2022.

[7] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16383–16392. IEEE, 2021.

[8] Jinze Huang, Xiaohan Yu, Dong An, Yaoguang Wei, Xiao Bai, Jin Zheng, Chen Wang, and Jun Zhou. Learning consistent region features for lifelong person re-identification. *Pattern Recognition*, 144:109837, 2023.

[9] Qiwei Li, Kunlun Xu, Yuxin Peng, and Jiahuan Zhou. Exemplar-free lifelong person re-identification via prompt-guided adaptive knowledge consolidation. *International Journal of Computer Vision*, 132(11):4850–4865, 2024.

[10] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[11] Xinyu Lin, Jinxing Li, Zeyu Ma, Huafeng Li, Shuang Li, Kaixiong Xu, Guangming Lu, and David Zhang. Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20973–20982, 2022.

[12] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13376–13386. IEEE, 2020.

[13] Zhijie Lu, Wuxuan Shi, He Li, and Mang Ye. Prototype-guided knowledge propagation with adaptive learning for lifelong person re-identification. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 5851–5859, 2025.

[14] Zhenxi Luo, Guoqiang Xiao, Michael S Lew, and Song Wu. Lifelong visible-infrared person re-identification with prompt pool and instance-level prompt generator. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 953–962, 2025.

[15] Hyeonjoon Moon and P Jonathon Phillips. Computational and performance aspects of pca-based face-recognition algorithms. *Perception*, 30(3):303–321, 2001.

[16] 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.

[17] Zhiqi Pang, Lingling Zhao, Junjie Wang, and Chunyu Wang. Lvlm-driven attribute-aware modeling for visible-infrared person re-identification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[18] Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12026–12035, 2021. doi: 10.1109/ICCV48922.2021.01183.

[19] Nan Pu, Wei Chen, Yu Liu, Erwin M Bakker, and Michael S Lew. Lifelong person re-identification via adaptive knowledge accumulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7901–7910, 2021.

[20] Nan Pu, Zhun Zhong, Nicu Sebe, and Michael S Lew. A memorizing and generalizing framework for lifelong person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13567–13585, 2023.

[21] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[22] Jiangming Shi, Xiangbo Yin, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. *Advances in Neural Information Processing Systems*, 37:99715–99734, 2024.

[23] Zhicheng Sun and Yadong Mu. Patch-based knowledge distillation for lifelong person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 696–707, 2022.

[24] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ich Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 618–626, 2019. doi: 10.1109/CVPR.2019.00071.

[25] Ziyu Wei, Xi Yang, Nannan Wang, and Xinbo Gao. Syncretic modality collaborative learning for visible infrared person re-identification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 225–234. IEEE, 2021.

[26] 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.

[27] Guile Wu and Shaogang Gong. Generalising without forgetting for lifelong person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2889–2897, 2021.

[28] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4330–4339, 2021.

[29] Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, Meng Liu, and Peng Wang. Enhancing visible-infrared person re-identification with modality-and instance-aware visual prompt learning. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 579–588, 2024.

[30] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382, 2019.

[31] Zesen Wu and Mang Ye. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9548–9558. IEEE, 2023.

[32] Yitong Xing, Guoqiang Xiao, Michael S Lew, and Song Wu. Lifelong visible-infrared person re-identification via a tri-token transformer with a query-key mechanism. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 988–997, 2024.

[33] Kunlun Xu, Xu Zou, Yuxin Peng, and Jiahuan Zhou. Distribution-aware knowledge prototyping for non-exemplar lifelong person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16604–16613, 2024.

[34] Kunlun Xu, Xu Zou, and Jiahuan Zhou. Lstkc: Long short-term knowledge consolidation for lifelong person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16202–16210, 2024.

[35] Kunlun Xu, Chenghao Jiang, Peixi Xiong, Yuxin Peng, and Jiahuan Zhou. Dask: Distribution rehearsing via adaptive style kernel learning for exemplar-free lifelong person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8915–8923, 2025.

[36] Kunlun Xu, Zichen Liu, Xu Zou, Yuxin Peng, and Jiahuan Zhou. Long short-term knowledge decomposition and consolidation for lifelong person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[37] Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11035–11045. IEEE, 2023.

[38] Bin Yang, Jun Chen, and Mang Ye. Shallow-deep collaborative learning for unsupervised visible-infrared person re-identification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16870–16879. IEEE, 2024.

[39] Xi Yang, De Cheng, Nannan Wang, Xinbo Gao, et al. Feature-level adversarial attacks and ranking disruption for visible-infrared person re-identification. *Advances in Neural Information Processing Systems*, 37:135043–135061, 2024.

[40] 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.

[41] 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.

[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.

[43] Mang Ye, Shipeng Zhang, Wuxuan Shi, Zhijie Lu, Xingyu Gao, and He Li. Camera-incremental person re-identification via geometric representation evolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[44] Chunlin Yu, Ye Shi, Zimo Liu, Shenghua Gao, and Jingya Wang. Lifelong person re-identification via knowledge refreshing and consolidation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 3295–3303, 2023.

[45] Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11185–11195, 2023.

[46] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7339–7348. IEEE, 2022.

[47] 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.

[48] Bowen Zhao, Xi Xiao, Guojun Gan, Bin Zhang, and Shu-Tao Xia. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13208–13217, 2020.

[49] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[50] Zhongao Zhou, Bin Yang, Wenke Huang, Jun Chen, and Mang Ye. Unbiased prototype consistency learning for multi-modal and multi-task object re-identification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[51] Jialong Zuo, Hanyu Zhou, Dongyue Wu, Yongtai Deng, Mengdan Tan, Nong Sang, Changxin Gao, and Xiang Bai. Achieving text-based person retrieval with any granularity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly identify two independent forgetting mechanisms (assignment ranking disruption and cross-modal topology inconsistency) and claim that the proposed S-PTC and CM-PTC modules address them without stored exemplars. These claims are directly supported by the ablation study (Table 4) and the SOTA comparison (Table 1) in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in a dedicated paragraph at the end of the Conclusion section, covering generalisation to other cross-modal settings and the scope of repeated-run evaluation. Multiple seeds are evaluated only for the full method under the standard task order; variability across seeds for ablations and other task orders remains to be evaluated.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper presents no formal theorems or mathematical proofs; all contributions are algorithmic and empirically validated.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 defines the method and stage-end EMA merge. Section 4.1 reports the backbone (ViT-Base/16 with ImageNet-21K initialisation), learning rate schedule (cosine decay from 3×10^{-4}), batch size (64), training epochs (50 per stage), image resolution (256×128), and the main hyperparameter settings (λ_s=0.1, τ_s=0.05, λ_{cm}=0.05, τ_{cm}=0.2).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is publicly available at `https://github.com/Tan-Shang/PTC`. The datasets are accessible through the sources cited in Section 4.1.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 describes the datasets, training and evaluation protocols, evaluation metrics, main training settings, and hardware. Section 4.4 analyses sensitivity to the loss weights and softmax temperatures and reports the settings used for the main experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Table 2 reports the full method under the standard task order with random seeds 0, 1, and 2, including the mean and sample standard deviation of the final average mAP and Rank-1. The standard deviation captures variation across complete training runs with different seeds and is computed with a denominator of $n - 1$, where $n = 3$. Ablations and other task orders are reported from single runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.1 specifies the hardware configuration: 2× NVIDIA RTX 4090 GPUs with mixed-precision (AMP) training.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research relies solely on publicly available datasets; no new human data were collected. The study introduces no deceptive practices and raises no privacy concerns beyond those acknowledged in the Broader Impact section.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: A dedicated Broader Impact section discusses positive applications (24-hour security monitoring, missing-person search) and calls for responsible deployment in compliance with privacy regulations.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new datasets, generative models, or scraped data. No assets with elevated misuse risk are introduced.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All four datasets (RegDB, SYSU-MM01, LLCM, VCM) and the ViT-Base/16 pretrained backbone are cited with their original publications in Section 4.1. These assets are used strictly within their publicly stated research-use terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper introduces no new datasets, benchmarks, or pretrained models. The sole contribution is the algorithmic method (S-PTC and CM-PTC), whose implementation details are fully described in Section 3.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper involves no crowdsourcing and no experiments with human subjects. All experiments are conducted on pre-existing publicly available datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research was conducted. IRB approval is therefore not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as any component of the proposed method. Any LLM assistance was limited to writing and editing and does not affect the core methodology or scientific results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.