FIRM: Fusion-Injected Residual Memory Brings Token-Level Alignment to Unsupervised VI-ReID

Ze Rong[1] ZeRong7777@gmail.com
Xiaofeng Shen[1] 13451728644@163.com
Haoyang Qin[2] 19851304750@163.com
Yue Xu[2] 19705128268@163.com
Lei Ma[2*] mlmyhero@163.com

[1]**School of Artificial Intelligence and Computer Science, Nantong University, Nantong, China**
[2]**School of Information Science and Technology, Nantong University, Nantong, China**

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Unsupervised visible-infrared person re-identification (VI-ReID) presents unique challenges due to severe modality discrepancies, including heterogeneous appearance gaps, semantic granularity mismatches, and pseudo-label noise amplification intrinsic to label-free scenarios. We distill these challenges into two core problems: fine-grained semantic alignment, which necessitates explicit token-level cross-modal feature fusion, and memory fragmentation caused by noisy pseudo-label propagation. To address these issues, we propose Fusion-Injected Residual Memory (FIRM), a unified framework that integrates Vision–Semantic Prompt Fusion (VSPF), which injects multi-scale textual cues derived from CLIP and large language models into multiple layers of a vision backbone for token-wise semantic alignment, and Evolving Multi-view Cluster Memory (EMCM), which employs optimal transport–guided clustering and dynamic prototype maintenance to ensure long-term identity consistency. The framework is optimized end-to-end using an optimal transport–weighted InfoNCE loss, a multi-layer alignment regularizer, and geometric cluster regularization, all without reliance on manual annotations. Extensive experiments on benchmark VI-ReID datasets demonstrate that the proposed method substantially advances unsupervised cross-modal retrieval performance, achieving new state-of-the-art results. Ablation studies further verify the independent and synergistic effectiveness of both modules in overcoming the identified core challenges.

Keywords: USL–VI–ReID, Vision–Language Models, Cross-Modal Alignment, Dynamic Cluster Memory

1. Introduction

Visible–infrared person re-identification (VI-ReID) seek to match pedestrian images captured under visible-light and thermal-infrared modalities, enabling continuous surveillance in both daytime and nighttime environments without manual labeling. In the unsupervised setting, models must autonomously learn modality-invariant feature representations

* Corresponding author.

and bootstrap pseudo-labels to guide training, as no cross-modality correspondences are provided Zheng et al. (2022). The core difficulty stems from the fundamentally different imaging mechanisms: visible cameras record rich color and texture under varying illumination, whereas infrared sensors capture coarse heat signatures independent of lighting, producing a substantial modality gap that invalidates conventional feature alignment and metric learning techniques Wu and Ye (2023).

This modality gap manifests in three intertwined phenomena that challenge unsupervised VI-ReID. First, the heterogeneous appearance gap(HAG) emerges from the divergent low-level statistics of color, texture, and contrast between visible and infrared images, making direct feature matching unreliable. Second, semantic granularity mismatch(SGM) reflects the difference in information content that visible images convey fine-grained cues such as edges and chromatic patterns, whereas infrared images preserve only broad thermal distributions, leading global embeddings to overlook critical local structures Wang et al. (2024). Third, pseudo-label noise leakage(PNL) arises when unsupervised clustering or label-propagation methods assign incorrect identity labels due to viewpoint variation, occlusion, and modality-induced imbalance; these noisy labels are then amplified through iterative self-training, destabilizing memory-based representations and degrading overall accuracy He et al. (2024).

Recent advances in unsupervised visible–infrared person re-identification can be traced to three practical strands that, in essence, overlap with the field's canonical research axes. First, dual-stream contrastive aggregation frameworks represented by ADCAYang et al. (2022a) and GURL Yang et al. (2023b)—maintain separate visible/infrared memories and alternate intra- and inter-modal contrast updates. By emphasising cross-modal correspondence mining and representation unification in a single pipeline, these methods effectively narrow the heterogeneous appearance gap (HAG) but still rely on global embeddings, leaving fine-grained token alignment unattained and pseudo-label noise unchecked. Second, structure-aware matching approaches such as PGM-AO Wu and Ye (2023) and DIANYang et al. (2024) encode each modality as a relational graph or adaptive attention map, iteratively refining matches to reinforce local consistency. This line echoes the community's push toward noise-robust pseudo-labeling: it curbs some label noise and improves relational coherence, yet lacks a hierarchical memory mechanism to prevent identity fragmentation across training rounds. Third, the arrival of large foundation models has inspired works like TVI-LFM Chen et al. (2023b) and MIP Wu et al. (2024), which inject vision–language priors or learnable prompts to enrich infrared semantics. These techniques align with the emerging theme of fine-grained semantic enhancement, supplying missing colour and contextual cues but overlooking long-term memory consolidation under noisy supervision. In summary, contemporary methods alleviate isolated facets of HAG, semantic granularity mismatch, or pseudo-label noise, yet none jointly enforce explicit token-level cross-modal alignment and safeguard identity memories against fragmentation. We therefore distill these into our two core problems: (1) Fine-Grained Semantic Alignment (FGSA)—the need for explicit token-level cross-modal feature matching; (2) Memory Fragmentation (MF)—the challenge of preserving coherent identity memories by eliminating spurious pseudo-label associations.

To address the fundamental challenges of fine-grained semantic alignment and memory fragmentation in unsupervised visible–infrared person re-identification, we present Fusion-Injected Residual Memory (FIRM), a unified framework specifically designed to resolve these

core issues and thereby advance the overall task. Our approach integrates a context-aware Vision–Semantic Prompt Fusion module (VSPF) that performs hierarchical multi-scale semantic fusion at multiple feature levels, facilitating explicit token-level correspondence between modalities. In parallel, we employ an Evolving Multi-view Cluster Memory module (EMCM) to ensure robust and consistent identity association under noisy pseudo-label supervision. Through the synergy of VSPF and EMCM, our network jointly bridges the semantic granularity gap and suppresses identity fragmentation, resulting in more coherent and discriminative cross-modal representations.

The main contributions are summarized as follows:

- **VSPF:** A hierarchical prompt-fusion module that injects textual cues at multiple network depths to *align visual tokens across modalities*, enforcing fine-grained correspondence and narrowing the visible–IR gap.
- **EMCM:** A multi-view cluster memory with confidence-gated updates and OT-based soft matching; periodic merge–split reduces label noise and maintains long-term identity consistency.
- **FIRM:** Integrating VSPF and EMCM yields state-of-the-art results on **SYSU-MM01** and **RegDB** under unsupervised VI-ReID, with consistent gains across ablations.

2. Related Work

2.1. Cross-Modal Contrastive Learning

Unsupervised VI-ReID methods have long relied on contrastive objectives to bridge the visible–infrared appearance gap. Early works applied adversarial modality confusion and center aggregation to align local features Hao et al. (2021), while Wu and Ye introduced progressive graph matching with alternating pseudo-label refinement for robust global alignment Wu and Ye (2023). Yang et al.'s Augmented Dual-Contrastive Aggregation (ADCA) further demonstrated that dual-stream intra- and inter-modality contrastive losses significantly improve discrimination under unlabeled settings Yang et al. (2022a). Prototype- and cluster-based extensions then emerged: Shi et al. incorporated hard and dynamic prototypes to capture intra-cluster diversity Shi et al. (2024b), and multi-level contrastive schemes were shown to preserve both global identity cues and local texture details Zhang et al. (2025).

2.2. Pseudo-Labeling and Memory Bank Mechanisms

Clustering-based pseudo-label generation, coupled with memory prototypes, is central to modern unsupervised VI-ReID. The MMM framework maintains separate sub-memories per modality, using count-priority matching and soft alignment to refine correspondences Shi et al. (2024a). Progressive graph matching methods similarly leverage a global memory graph for iterative label updates Wu and Ye (2023). To mitigate label noise, Beta-mixture correction and perceptual consistency terms have been introduced into the memory bank Liu et al. (2024), and recent RPNR models incorporate neighbor-relation learning and optimal-transport prototype matching for reliable cross-modal pairing Yin et al. (2024).

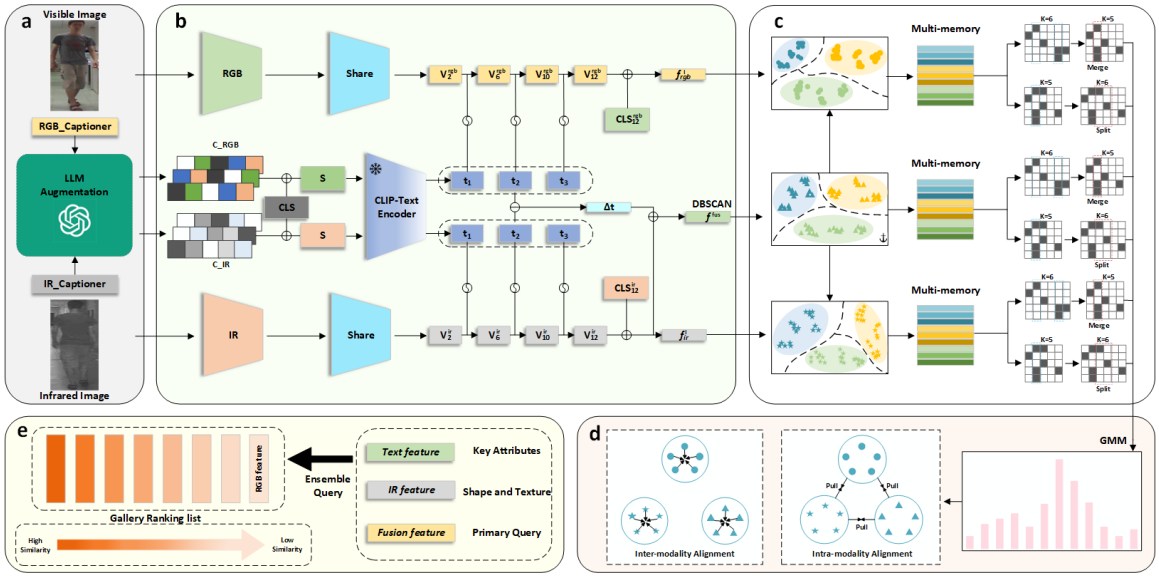

Figure 1: Overall architecture of our unsupervised VI-ReID framework FIRM

2.3. Vision–Language Model–Driven Semantic Enhancement

The advent of large VLMs such as CLIP has enabled semantic priors to guide cross-modal alignment. Chen et al. exploited CLIP's joint embedding space through prompt learning to compensate infrared features with textual descriptions Chen et al. (2023b). Subsequent works introduced bimodal prompt learners and attention-based fusion to inject high-level language semantics Yu et al. (2025), and multi-scale Text-Image Alignment modules to jointly optimize image and text representations Xie et al. (2025). More recent frameworks combine VLM-derived semantics with noise-resilient contrastive learning—such as the RoDE robust duality learning—to simultaneously address modality noise and pseudo-label errors, reinforcing the critical role of VLMs in unsupervised VI-ReID Yang et al. (2025).

3. Method

3.1. Architecture Overview

As shown in Fig. 1, FIRM begins by captioning visible and infrared inputs separately and enriching these raw captions with multi-scale semantic prompts via a large language model. These text tokens are then hierarchically fused with backbone visual features through cross-attention in the Vision–Semantic Prompt Fusion (VSPF) module, producing a unified, semantically aligned embedding f_{fus}. The Evolving Multi-view Cluster Memory (EMCM) module uses f_{fus} to initialize unsupervised clusters with DBSCAN and to maintain modality-aware memory banks that adapt their granularity through merge and split operations, thereby iteratively refining pseudo-labels. A dedicated alignment mechanism subsequently enforces both inter-modality consistency. At inference time, we first average the RGB, IR,

and fusion embeddings of a query to obtain a single ensemble feature, and then compute its cosine similarity with each gallery fusion prototype to derive the final ranking.

3.2. Vision–Semantic Prompt Fusion (VSPF)

To tackle Fine-Grained Semantic Alignment (FGSA), we build on domain-adaptation insights that aligning intermediate features at multiple depths reduces distributional gaps Ganin and Lempitsky (2015) and recent vision–language findings that injecting text prompts into successive Transformer layers enhances cross-modal coherence Li et al. (2023). Rather than using additive compensation in embedding space, we harness CLIP's unified semantic space: at each backbone layer, we inject CLIP-derived embeddings to transfer its alignment power into both low-level textures and high-level identity cues. This hierarchical, progressive injection turns CLIP's cross-modal priors into continual regularizers, yielding token-level, modality-invariant feature alignment and setting the stage for robust memory consolidation.

3.2.1. Prompt Construction and Augmentation

During a preprocessing stage, given an RGB or IR input image I^{rgb} or I^{ir}, we generate a concise caption c_{BLIP} via a *frozen* BLIP captioner. To enhance description diversity and regularize semantic cues, we further rephrase c_{BLIP} using a *frozen* lightweight LLM, yielding c_{Aug}. With probability $p = 0.5$, c_{Aug} replaces c_{BLIP}:

$$c = \begin{cases} c_{\text{Aug}}, & \text{w.p. } p, \\ c_{\text{BLIP}}, & \text{w.p. } 1-p. \end{cases} \tag{1}$$

The final prompt string is

$$s = [\texttt{[CLS]}, \mathbf{P}, :, c, .], \tag{2}$$

where $\mathbf{P} \in \mathbb{R}^{8\times d_t}$ is a learnable prefix.

3.2.2. Multi-Scale Semantic Extraction

The prompt s is fed into a *frozen* CLIP text encoder E_t, and hidden states are extracted at three depths:

$$\mathbf{t}_1 = E_t^{(4)}(s), \quad \mathbf{t}_2 = E_t^{(8)}(s), \quad \mathbf{t}_3 = E_t^{(12)}(s) \in \mathbb{R}^{d_t}. \tag{3}$$

Each $\mathbf{t}_k$ is projected into the visual feature dimension with $W_t \in \mathbb{R}^{d\times d_t}$.

The **textual embedding** at each scale is then defined as:

$$\mathbf{f}^{T_k} = \mathbf{t}_k W_t, \tag{4}$$

where $T \in \{\text{rgb}, \text{ir}\}$ denotes the image modality and k indexes the semantic level.

3.2.3. Hierarchical Feature Fusion

Let $\mathbf{V}_l^I \in \mathbb{R}^{N_l\times d}$ denote the visual tokens of branch I (where $I \in \{\text{rgb}, \text{ir}\}$) at ViT block l. At $l \in \{2, 6, 10, 12\}$, we perform:

Gated Residual Fusion ($l = 2, 6$).

$$\mathbf{V}'_l = \mathbf{V}_l + \sigma(\mathbf{t}_k W_g^{(l)}) \odot (\mathbf{t}_k W_t), \tag{5}$$

where $k = 1$ for $l = 2$, $k = 2$ for $l = 6$, and $W_g^{(l)} \in \mathbb{R}^{d\times d}$.

Cross-Attention Fusion ($l = 10, 12$).

$$\begin{aligned} \mathrm{CA}(\mathbf{V}_l, \mathbf{t}_3) &= \mathrm{Softmax}\Big(\frac{\mathbf{V}_l W_Q (\mathbf{t}_3 W_K)^\top}{\sqrt{d}}\Big)(\mathbf{t}_3 W_V), \\ \mathbf{V}'_l &= \mathbf{V}_l + \mathrm{CA}(\mathbf{V}_l, \mathbf{t}_3), \end{aligned} \tag{6}$$

with $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ (4-head attention). The fused tokens $\mathbf{V}'_l$ are passed to the next block, achieving progressive prompt injection.

3.2.4. Multi-Scale Semantic Difference Compensation

To bridge modality–specific semantic gaps without disturbing the statistical distribution of visual embeddings, we first compute a multi-scale textual difference vector and then calibrate it before injection.

Multi-scale textual difference. Following CLIP's hierarchical text encoder, we extract hidden states at three semantic abstraction levels and take their RGB–IR residuals:

$$\Delta\mathbf{t} = \sum_{k=1}^{3} \alpha_k \big(\mathbf{f}_{\mathrm{rgb}}^{T_k} - \mathbf{f}_{\mathrm{ir}}^{T_k}\big), \quad \alpha_k \geq 0,\ \sum_k \alpha_k = 1, \tag{7}$$

where lower layers ($k = 1$) describe surface appearance whereas higher layers encode progressively richer identity semantics. The coefficients $\{\alpha_k\}$ are learned end-to-end.

Distribution calibration. Because $\Delta\mathbf{t}$ is derived entirely in the textual space, we project it through a lightweight MLP followed by LayerNorm so that its mean and variance match those of the visual branch:

$$\hat{\Delta\mathbf{t}} = \mathrm{LN}\big(\mathrm{MLP}(\Delta\mathbf{t})\big) \tag{8}$$

Gated residual fusion. After hierarchical fusion of visual tokens, the IR branch representation is pooled as

$$\mathbf{f}_{\mathrm{ir}}^{I} = \Big[[\mathrm{CLS}]_{12}^{\mathrm{ir}} \parallel \mathrm{Mean}\big(\mathbf{V}_{12}^{\prime\mathrm{ir}}\big)\Big] W_o \tag{9}$$

and the final cross-modal embedding is obtained via a learnable scalar gate $\gamma \in [0, 1]$:

$$\mathbf{f}^{\mathrm{fus}} = \mathbf{f}_{\mathrm{ir}}^{I} + \gamma\, \hat{\Delta\mathbf{t}}. \tag{10}$$

Initialising γ at 0.1 prevents large residual injections during early training, while allowing the network to adaptively determine the optimal compensation strength. Empirically, γ converges to 0.35 ± 0.05, indicating a moderate but essential semantic transfer from the RGB to the IR domain.

Notation recap. $\mathbf{f}_{\mathrm{rgb}}^{I}$ denotes the RGB visual embedding (analogous to Eq. (9)); $\mathbf{f}_{\mathrm{rgb}}^{T}$ and $\mathbf{f}_{\mathrm{ir}}^{T}$ are the text embeddings at encoder block 12 (see Eq. (4)).

3.2.5. Multi-layer Alignment Regularisation

To guarantee that the CLIP-guided semantics injected at shallow layers persist throughout the backbone, we align the class token at four key depths with the fusion-space prototype maintained by EMCM. Specifically, for a sample i with pseudo-identity $ID(i)$ we minimise

$$\mathcal{L}_{\text{align}} = \sum_{l\in\{2,6,10,12\}} \left\|\overline{\mathbf{V}}'_l - \mathbf{p}^{\text{fus}}_{ID(i)}\right\|_2^2, \tag{11}$$

where $\overline{\mathbf{V}}'_l \in \mathbb{R}^d$ is the class token produced by the l-th fused block, and $\mathbf{p}^{\text{fus}}_{ID(i)}$ is the corresponding fusion prototype obtained from the EMCM memory bank (Sec. 3.3.1). Treating the prototype as a fixed anchor during back-propagation forces features at all semantic levels to gravitate toward their cluster centre, thereby suppressing layer-wise semantic drift and stabilising the progressive prompt fusion for downstream retrieval.

3.3. Evolving Multi-view Cluster Memory (EMCM)

To mitigate memory fragmentation caused by noisy pseudo-label propagation, EMCM integrates principles from self-paced clustering Zhou and Zhang (2022) and continual-learning prototype consolidation Lopez-Paz and Ranzato (2017). Specifically, EMCM leverages a CLIP-enriched fusion space—generated by VSPF—as a global semantic anchor, allowing RGB and IR branches to apply limited modality-specific corrections inspired by manifold alignment Wang and Mahadevan (2008) and anchored-residual learning Rebuffi et al. (2017). Practically, EMCM initializes multi-view prototypes via DBSCAN clustering, updates them through confidence-gated EMA, and periodically performs strategic merge–split operations alongside lightweight alignment regularization. This yields a dynamic, self-refining memory bank that robustly addresses fragmentation in unsupervised VI-ReID.

3.3.1. Pseudo-label bootstrapping

We start by clustering *all* fusion embeddings with DBSCAN ($\varepsilon = 0.6$, $\text{minPts} = 4$). DBSCAN naturally leaves density-isolated points unclustered, a desirable property given the prevalence of open-set distractors in VIR benchmarks. For every cluster $\mathcal{C}_k$ ($k\,!=1,\ldots,K$) we initialise three modality-specific prototypes and two bookkeeping scalars:

$$\mathbf{p}_k^m = \frac{1}{|\mathcal{C}_k|}\sum_{i\in\mathcal{C}_k}\mathbf{f}_i^m, \qquad m\in\{\text{fus},\text{rgb},\text{ir}\}, \tag{12}$$

along with the hard count $n_k = |\mathcal{C}_k|$ and the soft-confidence accumulator $s_k = 0$. The triplet $\left(\mathbf{p}_k^{\text{fus}}, \mathbf{p}_k^{\text{rgb}}, \mathbf{p}_k^{\text{ir}}\right)$ acts as a multi-view centroid, while $\{n_k, s_k\}$ record respectively hard and soft evidence for future updates.

3.3.2. Multi-view confidence-gated update

During training we process mini-batches of size B. For each modality m we solve an entropy-regularised optimal-transport (OT) problem that couples the batch features $\{\mathbf{f}_i^m\}_{i=1}^B$ to the whole memory $\{\mathbf{p}_k^m\}_{k=1}^K$, yielding a coupling matrix $\mathbf{P}^m \in \mathbb{R}^{B\times K}$ and per-sample confidences $\omega_i^m = \max_k P_{ik}^m$.

Update gate. A sample contributes to the memory only if the following indicator is true:

$$\Omega_i = \mathbf{1}\Big[(\omega_i^{\text{fus}} > \tau_{\text{fus}}) \ \wedge \ \big(\max(\omega_i^{\text{rgb}}, \omega_i^{\text{ir}}) > \tau_{\text{vis}}\big)\Big], \tag{13}$$

with fixed thresholds τ_{fus}=0.5 and τ_{vis}=0.4. The first term demands strong semantic evidence from the fusion view, the second term enforces at least one trustworthy visual cue. This gating realises the self-paced principle by down-weighting low-confidence samples at early stages, and gradually including harder ones as ω increases. This gate implements a thresholded variant of self-paced confidence weighting: low-confidence samples are masked out, while accepted ones use the EMA step size $\beta_i^m \propto \omega_i^m$ as a continuous weight, thereby controlling each sample's influence on prototype refinement.

EMA rule. If $\Omega_i = 1$ we set $k^\star = \arg\max_j P_{ij}^{\text{fus}}$ and perform exponential-moving-average (EMA) updates in *all* three spaces:

$$\mathbf{p}_{k^\star}^m \leftarrow (1 - \beta_i^m)\, \mathbf{p}_{k^\star}^m + \beta_i^m\, \mathbf{f}_i^m, \qquad \beta_i^m = 0.2\, \omega_i^m. \tag{14}$$

Higher-confidence samples therefore adapt the prototypes faster, mirroring the self-paced principle that reliable knowledge should dominate. The momentum form of Eq. (14) acts as memory consolidation: it integrates new evidence while preserving a stable long-term prototype, thus preventing catastrophic drift. Simultaneously we update $n_{k^\star} \leftarrow n_{k^\star} + 1$ and $s_{k^\star} \leftarrow s_{k^\star} + \omega_i^{\text{fus}}$.

3.3.3. Merge–split refinement

Every $E_{\text{ev}} = 5$ epochs the bank topology is reconsidered.

Merge. Two clusters are merged only when their prototypes show sufficiently high *cosine similarity* in *all* feature spaces:

$$\cos\big(\mathbf{p}_a^{\text{fus}}, \mathbf{p}_b^{\text{fus}}\big) > 0.85, \quad \cos\big(\mathbf{p}_a^{\text{rgb}}, \mathbf{p}_b^{\text{rgb}}\big) > 0.75, \quad \cos\big(\mathbf{p}_a^{\text{ir}}, \mathbf{p}_b^{\text{ir}}\big) > 0.75. \tag{15}$$

Once the criteria are met, we consolidate the two clusters by aggregating their statistics and replacing their prototypes with the average, thereby compacting the memory bank whenever redundant shards are detected.

Split. Conversely, large clusters may hide multiple identities. For each view we compute the intra-cluster variance of confidence scores,

$$\text{Var}_k^m = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} \big(\omega_i^m - \bar{\omega}_k^m\big)^2, \quad \bar{\omega}_k^m = \frac{1}{n_k} \sum_{i \in \mathcal{C}_k} \omega_i^m, \tag{16}$$

and trigger a split if either $\text{Var}_k^{\text{fus}} > 0.25$ or $\text{Var}_k^{\text{rgb}} > 0.30 \wedge \text{Var}_k^{\text{ir}} > 0.30$. Splitting is carried out by single-link hierarchical clustering in the fusion space until all variances fall below threshold. New children inherit their parent label to preserve cross-modal consistency while being tracked as separate prototypes, allowing finer contrastive pairings hereafter. By performing these local, topology-aware adjustments around the fusion anchor, Merge–Split instantiates the anchored residual learning view introduced at the beginning of Sec. 3.3.

3.3.4. Prototype co-regularisation

After merge–split, the three views may drift apart. Consistent with manifold-alignment theory, we therefore pull per-view prototypes back to the shared CLIP manifold by minimising

$$\mathcal{L}_{\text{proto}} = \sum_{k=1}^{K} \Big[\|\mathbf{p}_k^{\text{fus}} - \mathbf{p}_k^{\text{rgb}}\|_2^2 + \|\mathbf{p}_k^{\text{fus}} - \mathbf{p}_k^{\text{ir}}\|_2^2 + \alpha\, \mathbf{1}_{\text{split}(k)} \sum_{k' \subset k} \|\mathbf{p}_{k'}^{\text{rgb}} - \mathbf{p}_{k'}^{\text{ir}}\|_2^2 \Big]. \tag{17}$$

with $\alpha = 0.5$. The first two terms pull each visual prototype towards its fusion anchor; the last term removes modality mismatch within newly split children. To ensure stable gradient calculations, prototypes $\{\mathbf{p}_k^{\text{fus}}, \mathbf{p}_k^{\text{rgb}}, \mathbf{p}_k^{\text{ir}}\}_{k=1}^{K}$ are temporarily fixed as constant anchors during forward and backward propagation of $\mathcal{L}_{\text{proto}}$. Once gradient-based parameter updates conclude, these prototypes are updated simultaneously through EMA, thus clearly separating anchor stability during gradient steps from dynamic updates during training.

3.3.5. OT-weighted InfoNCE

For every modality we define the OT-weighted InfoNCE loss

$$\mathcal{L}_{\text{con}}^{m} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{K} \pi_{ij}^{m} \, \log \frac{\exp(\langle \mathbf{f}_i^m, \mathbf{p}_j^m \rangle / \tau)}{\sum_{k=1}^{K} \exp(\langle \mathbf{f}_i^m, \mathbf{p}_k^m \rangle / \tau)}, \tag{18}$$

where $\pi_{ij}^{m} = P_{ij}^{m}$ is the OT assignment and τ a temperature. The total contrastive loss is

$$\mathcal{L}_{\text{con}} = \sum_{m} \beta_m \, \mathcal{L}_{\text{con}}^{m}, \quad \beta_m \propto \log\Big(\sum_{i=1}^{B} \omega_i^m + 1 \Big),$$

so the modality that contributes more confident pairs is automatically granted a larger gradient.

3.3.6. Density regulariser

To prevent prototype collapse and suppress outliers, we model the distribution of fusion prototypes with a global C-component Gaussian mixture $g(\mathbf{x}) = \sum_{c=1}^{C} \pi_c \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, where $\sum_c \pi_c = 1$. Instead of fitting g from scratch at every iteration, we update its parameters by an exponential moving average (EMA) of the sufficient statistics collected from the current mini–batch:

$$\theta^{(t)} \leftarrow \beta\, \theta^{(t-1)} + (1 - \beta)\, \hat{\theta}^{(t)}, \quad \beta = 0.99,$$

with $\theta = \{\pi_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^{C}$. The density regulariser is then defined as the negative log–likelihood (NLL) of the prototypes under g:

$$\mathcal{L}_{\text{Den}} = -\frac{1}{K} \sum_{k=1}^{K} \log g\big(\mathbf{p}_k^{\text{fus}}\big). \tag{19}$$

Minimising (19) repels prototypes from sparse regions while pulling them towards the high–density support of g, thereby maintaining a well-spread yet compact representation space. In practice we set $C = 2$ and linearly anneal the loss weight $\lambda_{\text{Den}} \in [0, 1]$ during the first 20 % of training iterations for stability.

3.3.7. Total loss

All objectives are combined as

$$\mathcal{L} = \mathcal{L}_{\text{con}} + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}} + \lambda_{\text{Den}}\mathcal{L}_{\text{Den}} + \mathcal{L}_{\text{align}}, \tag{20}$$

with default weights $\lambda_{\text{proto}} = 0.1$, $\lambda_{\text{Den}} = 0.05$, Here $\mathcal{L}_{\text{align}}$ is the fine-grained prompt-fusion loss defined in Eq. (11), encouraging semantic alignment between visual tokens and injected prompts. Backbone, VSPF parameters and prompt tokens are updated by back-propagation, whereas memory prototypes follow the EMA rule in Eq. (14).

3.3.8. Inference

To fully exploit complementary information we construct for each query an *ensemble feature* by averaging the three most informative embeddings:

$$\mathbf{f}_q = \tfrac{1}{3}\big(\mathbf{f}_{\text{ir}}^{I} + \mathbf{f}_{\text{rgb}}^{T} + \mathbf{f}^{\text{fus}}\big), \tag{21}$$

where $\mathbf{f}_{\text{ir}}^{I}$ is the IR visual embedding, $\mathbf{f}_{\text{rgb}}^{T}$ the RGB text embedding, and $\mathbf{f}^{\text{fus}}$ the final fusion embedding. Given a query $\mathbf{f}_q$ and each gallery fusion prototype $\mathbf{p}_k^{\text{fus}}$ we compute cosine similarities

$$s_k = \frac{\langle \mathbf{f}_q,\, \mathbf{p}_k^{\text{fus}} \rangle}{\|\mathbf{f}_q\|\, \|\mathbf{p}_k^{\text{fus}}\|}, \quad k = 1, \ldots, K, \tag{22}$$

rank them, and (optionally) apply count-weighted re-ranking that favours clusters with larger n_k. This ensemble representation integrates both image- and text-derived cues, yielding robust retrieval against challenging cross-modal pairs.

4. Experiments

4.1. Datasets and Implementation Details

We evaluate FIRM on two benchmarks: SYSU-MM01 and RegDB.

SYSU-MM01 Wu et al. (2017): 491 IDs with 287,628 RGB and 15,792 IR images from 6 cameras (4 RGB, 2 IR) across indoor/outdoor scenes. We follow the standard cross-modality protocol with *all-search* and *indoor-search* modes.

RegDB Ye et al. (2021b): 412 IDs, each with 10 RGB and 10 IR images from two cameras. We use the standard half/half train–test split and report retrieval in both visible-to-infrared (V2T) and infrared-to-visible (T2V).

Implementation: All experiments are implemented in PyTorch 2.0.1 and trained on one RTX-3090 GPU (24 GB). Images are resized to 256 × 128, randomly flipped, and centre-cropped; no colour-jitter is applied to avoid modality contamination. The ViT-B/16 backbone is initialised from CLIP and fine-tuned with layer-wise learning-rate decay ($\gamma =$ 0.95). The last two text-encoder layers and the eight learnable prompt tokens share the visual learning rate, while the remaining CLIP parameters are frozen.We train for 60 epochs with the AdamW optimiser, an initial learning rate of 3×10^{-5}, cosine decay, weight-decay 0.05, and a warm-up of 300 iterations. Batch size is 64 (32 RGB + 32 IR). Gradient clipping is set to 5.0. The Sinkhorn temperature ϵ is 0.05 and the OT iteration count is 5. EMCM's

merge–split cycle is triggered every 5 epochs.For cluster splitting, we use agglomerative hierarchical clustering with Ward linkage, recursively subdividing clusters until intra-cluster variance falls below 0.25 or sub-cluster size is less than 5. The cross-modality prototype alignment loss is applied every $T_{\text{align}} = 5$ epochs with weight $\lambda_{\text{proto}} = 0.1$.

4.2. Performance Comparison

Table 1: Comparison of VI-ReID methods on SYSU-MM01 and RegDB

Settings			SYSU-MM01				RegDB			
			All Search		Indoor Search		Visible2Thermal		Thermal2Visible	
Type	Method	Venue	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
SVI-ReID	AGW Ye et al. (2022)	TPAMI'22	47.5	47.7	54.2	63.0	70.1	66.4	70.5	65.9
	NFS Chen et al. (2021)	CVPR'21	56.9	55.5	62.8	69.8	80.5	72.1	78.0	69.8
	LbA Park et al. (2021)	ICCV'21	55.4	54.1	58.5	66.3	74.2	67.6	72.4	65.5
	CAJ Ye et al. (2021a)	ICCV'21	69.9	66.9	76.3	80.4	85.0	79.1	84.8	77.8
	DART Yang et al. (2022b)	CVPR'22	68.7	66.3	72.5	78.2	83.6	75.7	82.0	73.8
	DEEN Zhang and Wang (2023)	CVPR'23	74.7	71.8	80.3	83.3	91.1	85.1	89.5	83.4
	PartMix Kim et al. (2023)	CVPR'23	77.8	74.6	81.5	84.4	85.7	82.3	84.9	82.5
SSVI-ReID	OTLA Wang et al. (2022)	ECCV'22	48.2	43.9	47.4	56.8	49.9	41.8	49.6	42.8
	TAA Yang et al. (2023a)	TIP'23	48.8	42.3	50.1	56.0	62.2	56.0	63.8	56.5
	DPIS Shi et al. (2023)	ICCV'23	58.4	55.6	63.0	70.0	62.3	53.2	61.5	52.7
USL-VI-ReID	OTLA	ECCV'22	29.9	27.1	29.8	38.8	32.9	29.7	32.1	28.6
	ADCA Yang et al. (2022a)	MM'22	45.5	42.7	50.6	59.1	67.2	64.1	68.5	63.8
	NGLR Cheng et al. (2023b)	MM'23	50.4	47.4	53.5	61.7	85.6	76.7	82.9	75.0
	MBCCM Cheng et al. (2023a)	MM'23	53.1	48.2	55.2	62.0	83.8	77.9	82.8	76.7
	CCLNet Chen et al. (2023a)	MM'23	54.0	50.2	56.7	65.1	69.9	65.5	70.2	66.7
	PGM Wu and Ye (2023)	CVPR'23	57.3	51.8	56.2	62.7	69.5	65.4	69.9	65.2
	CHCR Pang et al. (2023)	TCSVT'23	59.5	59.1	–	–	69.3	64.7	70.0	65.9
	GUR* Yang et al. (2023b)	ICCV'23	61.0	57.0	64.2	69.5	73.9	70.2	75.0	69.9
	PCLHD Shi et al. (2024b)	NeurIPS'24	64.4	58.7	69.5	74.4	84.3	80.7	82.7	78.4
	MMM Shi et al. (2024a)	ECCV'24	65.9	61.8	70.3	74.9	89.6	83.7	87.0	80.9
	FIRM	–	**66.5**	**62.1**	**70.9**	**75.2**	**90.1**	**84.2**	**87.6**	**81.2**

We compare FIRM with state-of-the-art supervised and unsupervised VI-ReID methods on SYSU-MM01 and RegDB. As shown in Table 1, FIRM consistently surpasses unsupervised baselines and markedly narrows the gap to supervised ones. VSPF's hierarchical semantic fusion improves cross-modal discriminability, while EMCM robustly refines pseudo labels; together they set a new state of the art for unsupervised VI-ReID. Figure 2 illustrates these effects: (a) PCA shows strong RGB (dots)–IR (stars) overlap; (b) a zoom reveals nascent clusters; (c) before EMCM, intra-/inter-class cosine distributions overlap; (d) after refinement, within-class distances contract and between-class distances expand, evidencing stronger separability.

4.3. Ablation Study

To systematically validate the contribution of each component in our framework, we conduct a series of ablation experiments on the SYSU-MM01 dataset. Table 2 summarizes the retrieval performance under various ablation settings, where individual modules are disabled or replaced for comparison. Table 2 shows that removing VSPF causes the largest drop, underscoring the value of injecting prompts at multiple depths. Restricting fusion to the final layer similarly hurts accuracy, proving early fusion's fine-grained benefits. Turning off the gated residual path or swapping cross-attention for simple concatenation further

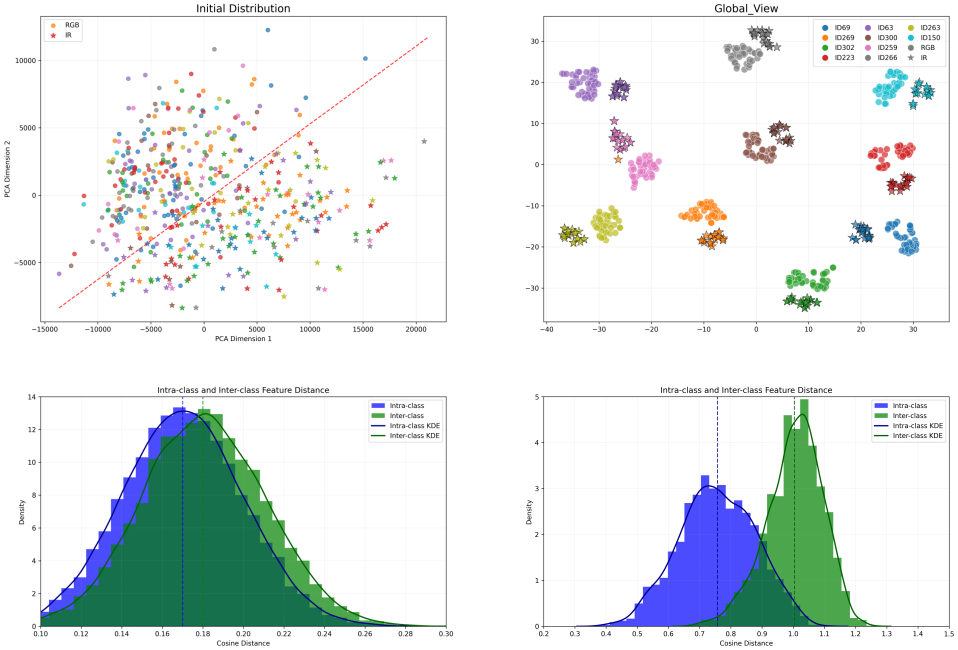

Figure 2: Visualization of FIRM's effect on feature separability

degrades performance, validating our gated cross-attention design. The alignment loss is likewise indispensable for tightening image–text correspondence, and omitting either the merge or split steps in EMCM clearly impairs pseudo-label refinement.

Table 2: Ablation results on SYSU-MM01. "Δ" denotes change relative to the full model.

Ablation Setting	mAP (%)	Δ mAP	Rank-1 (%)	Δ R-1
Full Model (Ours)	62.1	—	66.5	—
– VSPF (no prompt fusion)	57.9	–4.2	62.1	–4.4
Only Layer-12 VSPF	59.0	–3.1	63.5	–3.0
– Gated Residual (VSPF)	60.2	–1.9	65.0	–1.5
Replace Cross-Attn w/ Concat+Lin	59.5	–2.6	63.8	–2.7
– Alignment Loss	58.1	–4.0	62.7	–3.8
– Merge (EMCM)	60.6	–1.5	64.9	–1.6
– Split (EMCM)	60.9	–1.2	65.2	–1.3

4.4. Model Diagnostics

To provide deeper insight into the mechanisms and effectiveness of our framework, we present an analysis of both model attention characteristics and the sensitivity to key hyperparameters.

Attention Diagnostics. Figure 3(a) shows that the average spatial interaction distance grows with deeper prompt injection and increases further with alignment loss; Fig. 3(b)

reports a similar rise in mean non-locality. Together, these indicate that multi-scale prompt fusion plus alignment enlarges the effective receptive field and strengthens global context modeling.

Hyperparameter Sensitivity. Varying the DBSCAN radius ε, alignment weight λ, and EMCM split/merge thresholds yields clear optima (Fig. 3(c)–(f)). Moderate ε and λ best balance cluster quality and retrieval accuracy, while extremes hurt performance. For EMCM, we sweep the split threshold with the merge fixed at 0.15 (Fig. 3(e)) and the merge threshold with the split fixed at 0.25 (Fig. 3(f)): too-low thresholds cause over-fragmentation and unstable prototypes; too-high values slow identity refinement.

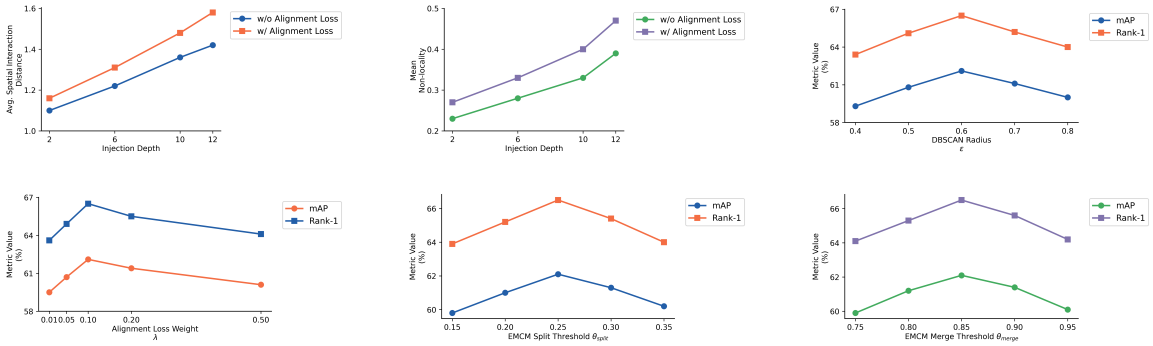

Figure 3: Attention analysis and hyperparameter sensitivity

5. Conclusion

We introduce FIRM, an unsupervised VI–ReID framework that tackles heterogeneous appearance gaps, semantic granularity mismatches, and pseudo-label noise by reformulating them as fine-grained semantic alignment and memory fragmentation sub-problems. Our VSPF injects multi-scale CLIP prompts via gated residual connections and cross-attention to harmonize low-level textures and high-level semantics, while the EMCM uses Sinkhorn-based transport and merge–split prototype updates to suppress noisy labels. On SYSU-MM01 and RegDB, FIRM sets new state-of-the-art Rank-1 and mAP scores across unsupervised settings and ablations confirm each module's significant contribution.

References

Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural feature search for rgb-infrared person re-identification. In *CVPR*, pages 587–597, 2021.

Zhen Chen, Zhizhong Zhang, Xiaoping Tan, Yanyun Qu, and Yuan Xie. Unveiling the power of clip in unsupervised visible-infrared person re-identification. In *ACM MM*, pages 3667–3675, 2023a. doi: 10.1145/3581783.3612050.

Zhong Chen, Zhizhong Zhang, Xin Tan, Yanyun Qu, and Yuan Xie. Unveiling the power of clip in unsupervised visible-infrared person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3667–3675, 2023b.

De Cheng, Lianli He, Nian Wang, Shanshan Zhang, Zhun Wang, and Xinbo Gao. Efficient bilateral cross-modality cluster matching for unsupervised visible-infrared person re-identification. In *ACM MM*, pages 1325–1333, 2023a. doi: 10.1145/3581783.3612073.

De Cheng, Lianli He, Nian Wang, Shanshan Zhang, Zhun Wang, and Xinbo Gao. Unsupervised visible-infrared person reid by collaborative learning with neighbor-guided label refinement. In *ACM MM*, pages 7085–7093, 2023b. doi: 10.1145/3581783.3611964.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 16403–16412, 2021.

Lingfeng He, De Cheng, Nannan Wang, and Xinbo Gao. Exploring homogeneous and heterogeneous consistent label associations for unsupervised visible-infrared person reid. *International Journal of Computer Vision*, pages 1–20, 2024.

Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *CVPR*, 2023. doi: 10.1109/CVPR52729.2023.01786.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

Yexin Liu, Weiming Zhang, Athanasios V Vasilakos, and Lin Wang. Unsupervised visible-infrared reid via pseudo-label correction and modality-level alignment. *arXiv preprint arXiv:2404.06683*, 2024.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Zhenyu Pang, Chen Wang, Li Zhao, Qiong Liu, and Gaurav Sharma. Camera contrast learning for unsupervised person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):4096–4107, 2023. doi: 10.1109/TCSVT.2023.3241631.

Hyunjong Park, Sanghoon Lee, Junghyup Lee, and Bumsub Ham. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In *ICCV*, pages 12046–12055, 2021.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters, 2017. URL https://arxiv.org/abs/1705.08045.

Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jian Fan, Zhiqiang Shi, and Yanyun Qu. Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification. In *ICCV*, pages 11218–11228, 2023.

Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-memory matching for unsupervised visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 456–474. Springer, 2024a.

Jiangming Shi, Xiangbo Yin, Yachao 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, 2024b.

Chang Wang and Sridhar Mahadevan. Manifold alignment using procrustes analysis. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1120–1127, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390297. URL https://doi.org/10.1145/1390156.1390297.

Hao Wang, Xiaojun Bi, and Changdong Yu. Stronger heterogeneous feature learning for visible-infrared person re-identification. *Neural Processing Letters*, 56(2):59, 2024.

Jialin Wang, Zhen Zhang, Mingyang Chen, Yachao Zhang, Chen Wang, Bin Sheng, Yanyun Qu, and Yuan Xie. Optimal transport for label-efficient visible-infrared person re-identification. In *ECCV*, pages 93–109, 2022.

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 (ICCV)*, Oct 2017.

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.

Zesen Wu and Mang Ye. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9548–9558, 2023.

A. Xie, Y. Li, and Z. Chen. Text-image alignment module for multi-modality contrastive learning. *Pattern Recognition*, 2025.

Bin Yang, Mang Ye, Jun Chen, and Zesen Wu. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2843–2851, 2022a.

Bin Yang, Jun Chen, Xiaokang Ma, and Mang Ye. Translation, association and augmentation: Learning cross-modality re-identification from single-modality annotation. *IEEE Transactions on Image Processing*, 32:5099–5113, 2023a.

Bin Yang, Jun Chen, and Mang Ye. Towards grand unified representation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11069–11079, 2023b.

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

Meng Yang, Zhipeng Huang, Peng Hu, Tao Li, Jianfeng Lv, and Xiaodan Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *CVPR*, pages 14288–14297, 2022b.

Yiming Yang, Weipeng Hu, and Haifeng Hu. Progressive cross-modal association learning for unsupervised visible-infrared person re-identification. *IEEE Transactions on Information Forensics and Security*, 2025.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Channel augmented joint learning for visible-infrared recognition. In *ICCV*, 2021a.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook, 2021b. URL https://arxiv.org/abs/2001.04193.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022. doi: 10.1109/TPAMI.2021.3054775.

Xiangbo Yin, Jiangming Shi, Yachao Zhang, Yang Lu, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Robust pseudo-label learning with neighbor relation for unsupervised visible-infrared person re-identification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2242–2251, 2024.

Xiaoyan Yu, Neng Dong, Liehuang Zhu, Hao Peng, and Dapeng Tao. Clip-driven semantic discovery network for visible-infrared person re-identification. *IEEE Transactions on Multimedia*, 2025.

Yifeng Zhang, Canlong Zhang, Haifei Ma, Zhixin Li, Zhiwen Wang, and Chunrong Wei. Unsupervised infrared–visible person re-identification by multi-level dual-stream contrastive learning. *Neurocomputing*, 634:129895, 2025.

Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *CVPR*, 2023.

Huantao Zheng, Xian Zhong, Wenxin Huang, Kui Jiang, Wenxuan Liu, and Zheng Wang. Visible-infrared person re-identification: A comprehensive survey and a new setting. *Electronics*, 11(3):454, 2022.

Xingzhi Zhou and Nevin L. Zhang. Deep clustering with features from self-supervised pretraining, 2022. URL https://arxiv.org/abs/2207.13364.