CONTRASTIVE LABEL-AWARE SPATIAL-FREQUENCY LEARNING FOR UNSUPERVISED MULTI-MODAL RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unsupervised multi-modal re-identification (UMM-ReID) aims to learn identity-discriminative representations from multi-modalities by constructing pseudo-labels through clustering. However, it is difficult to obtain reliable and globally consistent pseudo-labels from multi-modal representations, and global pseudo-label supervision itself cannot clearly guide the representation learning and strengthen the focus on identity-discriminative regions. To address these issues, we propose Contrastive Label-aware Spatial-frequency Learning, which is a framework that alternates between global pseudo-label and minibatch representation learning. At the beginning of each epoch, we use adaptive spectral clustering and spectral gap-MDL criterion to generate global adaptive pseudo-labels for subsequent minibatch optimization. Based on pseudo-labels, cross-modal contrastive learning selectively adds reliable intra-modal negative relations to strengthen representation learning. In addition, we use pseudo-labels to guide spatial discrimination, and combine complementary frequency cues to enhance the learning of discriminative regions and suppress background noise. Experiments on real datasets fully demonstrate the effectiveness of this method.

1 INTRODUCTION

Object Re-Identification (ReID) aims to retrieve instances of the same identity across images captured by different cameras Li et al. (2020); Wu et al. (2021); Chen et al. (2021). While conventional ReID methods mainly operate on RGB images, practical sensing systems increasingly provide heterogeneous observations, such as visible, near-infrared, and thermal modalities. Multi-modal observations provide complementary information, but they also introduce substantial modality discrepancies Fan & Ling (2019); Kart et al. (2019). Moreover, acquiring identity labels consistently across multi-modalities is expensive and difficult to scale. These problems motivate the development of UMM-ReID Li et al. (2019); Huang et al. (2019), which seeks to learn cross-modal identity representations without identity labels.

Unsupervised ReID obtains pseudo-labels through clustering, which are used as supervisory signals for representation learning Pang et al. (2024). This paradigm becomes considerably more challenging in the multi-modal setting because representations are affected by modality discrepancy. As a result, two closely related difficulties arise.

First, reliable and global pseudo-labels are difficult to obtain and effectively exploit in multi-modal representations Li et al. (2019), while cross-modal contrastive learning relies on them to learn cross-modal representation during minibatch optimization. Pseudo-labels constructed from limited minibatch observations may provide inconsistent supervision, so global modeling is needed. Existing clustering-based pseudo-label methods, such as K-Means Zheng et al. (2017); Li et al. (2014) and DBSCAN Wang et al. (2023), are often sensitive to the complex non-convex structure and non-uniform density of global multi-modal representations. Spectral clustering (SC) Pang et al. (2020); Zhao et al. (2025a), by analyzing the affinity matrix, provides a more flexible method for capturing such complex global structures. However, directly applying SC to UMM-ReID still faces two limitations: affinities constructed with Gaussian kernel are sensitive to the distance scale Nie et al. (2016), while conventional SC requires the number of clusters to be specified in advance Wang et al. (2023).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This calls for adaptively learning the affinity structure while estimating the number of pseudo-labels from the data, for which the minimum description length (MDL) principle Cai et al. (2026) provides a principled criterion for cluster-number estimation. Nevertheless, global pseudo-labels inevitably contain noise. Since cross-modal contrastive learning primarily promotes cross-modal consistency Wang et al. (2024a); Zhao et al. (2025b), we further exploit reliable intra-modal negative pairs according to pseudo-label confidence to strengthen identity-discriminative representations.

Second, although global pseudo-labels provide supervisory signals, they do not explicitly guide the model to identify discriminative regions in the spatial domain Huang et al. (2019). Without explicit label guidance, the model may attend to background noise Wu et al. (2018), resulting in locally salient but irrelevant representations Yang et al. (2023). Moreover, spatial discrimination alone may be insufficient to characterize subtle identity-specific details under multi-modalities. Frequency information provides complementary cues that are not explicitly captured in spatial domain Hsu & Wu (2025); Liu et al. (2021). This motivates using pseudo-labels to guide spatial discrimination while introducing frequency cues as complementary information for representation learning.

To address these challenges, we formulate UMM-ReID as an alternating process between global adaptive pseudo-label and minibatch representation learning. At the beginning of each epoch, representations of the entire training data are extracted to construct global adaptive pseudo-labels, which are then fixed to provide consistent supervision for minibatch optimization. Specifically, we construct an adaptive affinity matrix for SC and introduce a spectral gap-MDL criterion (SG-MDL) to estimate the unknown number of identities, thereby generating global pseudo-labels. Based on pseudo-labels, during minibatch optimization, we perform cross-modal contrastive learning and estimate pseudo-label confidence, allowing reliable intra-modal negative pairs to be incorporated into contrastive optimization to strengthen identity-discriminative representations. Prototype label learning further stabilizes representation learning. Finally, we develop label-aware spatial-frequency (SF) learning in minibatch representation learning, which uses pseudo-labels to guide spatial discrimination and supplements frequency cues to enhance the learning of identity-discriminative regions and suppress background noise.

- We establish a global-to-minibatch UMM-ReID framework, where adaptive spectral clustering and SG-MDL construct global adaptive pseudo-labels to provide consistent supervision for minibatch optimization.
- We propose cross-modal contrastive learning that selectively incorporates reliable intra-modal negative pairs according to pseudo-label guide, thereby complementing cross-modal learning with stronger identity-discriminative representations.
- We develop label-aware SF learning that uses pseudo-labels to guide spatial discrimination and incorporates complementary frequency cues, strengthening the learning of identity-discriminative regions and suppressing background noise.

2 Related work

2.1 Unsupervised Re-Identification

Clustering-based pseudo-label learning is widely used in unsupervised ReID to provide identity supervision for representation learning. A line of research improves pseudo-label reliability by refining intra-cluster structures or neighborhood relations. ISE Zhang et al. (2022) progressively reduces intra-cluster discrepancies to alleviate clustering errors, while PPLR Cho et al. (2022) refines pseudo-labels through cross-consistent neighborhood relations. Multi-instance Proxy Zou et al. (2023) further models cluster structures with multiple proxies. Beyond pseudo-label construction, SpCL Ge et al. (2020) and Cluster Contrast Dai et al. (2022) exploit memory- or cluster-level representations for contrastive learning. Despite these advances, discovering reliable and globally consistent pseudo-identities remains challenging under complex feature distributions, particularly when the identity number is unknown.

2.2 Multi-modal Re-Identification

Multi-modal ReID aims to learn identity-discriminative representations across multi-modalities. Supervised methods reduce modality discrepancies through cross-modal feature aggregation, such as

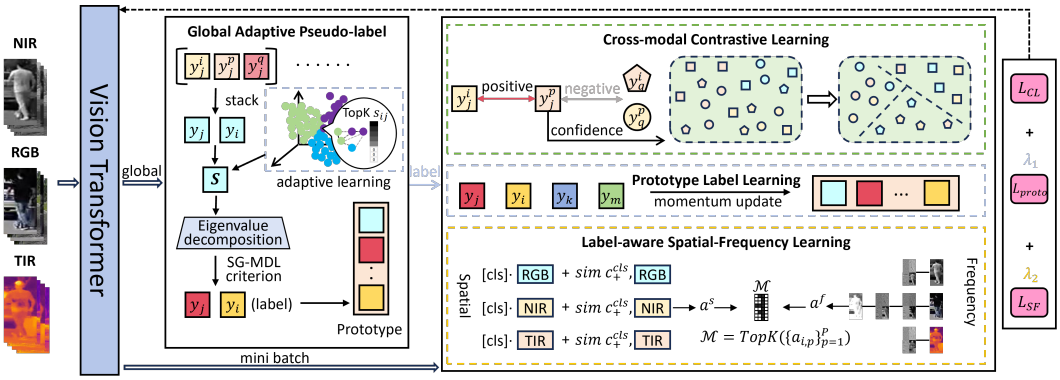


Figure 1: The architecture of our method. At the beginning of Epoch, ViT extracts representations to construct global pseudo-labels, and fixes them to guide representation learning; In minibatch processing, cross-modal contrastive learning combined with label confidence is used to strengthen identity-discriminative representations, and prototype label learning stabilizes representation learning. Then, label-aware SF learning is used to enhance the learning of discriminative regions and suppress background noise.

grayscale transformation Wu et al. (2017) and multi-spectral feature aggregation in Top-ReID Wang et al. (2024b). To reduce label dependence, H2H Liang et al. (2021) explores cross-modal contrastive learning, while OTAL Wang et al. (2022) transfers pseudo-label information across modalities. Beyond global representation learning, local region selection has also been explored: TransFG He et al. (2022) selects representative patches using self-attention, while DynamicViT Rao et al. (2021) performs dynamic token selection. However, spatial information does not fully express identity discrimination, and alone may miss complementary details available without frequency cues Hsu & Wu (2025); Liu et al. (2021). This motivates exploiting pseudo-labels to strengthen spatial discrimination while combines complementary frequency cues for representation learning.

3 Method

3.1 Framework Overview

Define $\mathbf{X} = \{\mathbf{x}_i^j | i = 1, ..., n; j = 1, ..., m\}$ as the set of original input images, where m, n are the number of modality and image batchsize respectively. We use Visual Transformer (ViT) to get $\mathbf{y}_i^j = f_\theta(\mathbf{x}_i^j)$.

Our goal is to jointly learn identity-discriminative representations without any supervisory signal. At epoch t, we first freeze ViT parameters θ_t and extract features for the entire training data,

$$\mathbf{Y}^{(t)} = f_{\theta_t}(\mathbf{X}). \tag{1}$$

Based on $\mathbf{Y}^{(t)}$, We estimate the number of pseudo-identities $k^{(t)}$, obtain pseudo-label assignments $\mathbf{Z}^{(t)} = \{z_i^{(t)} \in \{1, 2, ..., k^{(t)}\}\}$, and construct cluster prototypes,

$$\mathbf{C}^{(t)} = \{\mathbf{c}_i^{(t)}\}_{i=1}^{k^{(t)}}. \tag{2}$$

The pseudo-label $\mathbf{Z}^{(t)}$ is fixed while the network is optimized over minibatches within epoch t. After the epoch is completed, the updated network $f_{\theta_{t+1}}$ is used to reconstruct representations and refresh pseudo-labels.

This yields the following alternating process:

$$\theta_t \rightarrow \mathbf{Y}^{(t)} \rightarrow (k^{(t)}, \mathbf{Z}^{(t)}, \mathbf{C}^{(t)}) \rightarrow \theta_{t+1}. \tag{3}$$

All minibatches within an epoch are optimized under common global pseudo-labels, avoiding inconsistent supervision caused by constructing pseudo-labels from limited minibatch observations.

3.2 GLOBAL ADAPTIVE PSEUDO-LABEL

This part aims to discover reliable global pseudo-labels from multi-modalities. Since representation distribution can be non-convex and non-uniform, directly applying clustering with a fixed distance scale may fail to capture the underlying identity structure. We therefore perform adaptive spectral clustering on global representations.

We stack the global representations $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ from different modalities, where $\mathbf{y}_i = [\mathbf{y}_i^1, ..., \mathbf{y}_i^m]$. Meanwhile, to adapt to representation distribution with non-convex and non-uniform density, we learn the adaptive affinity matrix $\mathbf{S}$,

$$\min_{\mathbf{s}_i^T\mathbf{1}=1, \mathbf{s}_i \geq 0, s_{ii}=0} \sum_{j=1}^{n} ||\mathbf{y}_i - \mathbf{y}_j||_2^2 s_{ij} + \gamma \sum_{j=1}^{n} s_{ij}^2, \tag{4}$$

where $\mathbf{s}_i$ is a vector, its j_{th} element is s_{ij} and $\gamma > 0$ is a regularization parameter. We expect to learn from Eq (4) that smaller distance $||\mathbf{y}_i - \mathbf{y}_j||_2^2$ between $\mathbf{y}_i$ and $\mathbf{y}_j$ corresponds to a larger affinity value s_{ij}. Let us define $e_{ij} = ||\mathbf{y}_i - \mathbf{y}_j||_2^2$ and donate $\mathbf{e}_i$ as a vector with j_{th} element as e_{ij}, then Eq (4) can be simplified as

$$\min_{\mathbf{s}_i^T\mathbf{1}=1, \mathbf{s}_i \geq 0, s_{ii}=0} \frac{1}{2}||\mathbf{s}_i + \frac{\mathbf{e}_i}{2\gamma}||_2^2. \tag{5}$$

According to Lagrange theorem and KKT condition, we get the optimal affinities s_{ij}

$$s_{ij} = \begin{cases} \frac{e_{i,q+1} - e_{ij}}{q e_{i,q+1} - \sum_{h=1}^{q} e_{ih}}, & j \leq q \\ 0, & j > q \end{cases}, \tag{6}$$

where q is a sparse parameter that restricts each sample to its q most relevant neighbors, reducing noisy graph connections. See Appendix A.1.1 for the specific derivation process.

Then we compute the normalized Laplace matrix $\mathbf{L} = \mathbf{D}^{-1/2}(\mathbf{D} - \mathbf{S})\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is the degree matrix of $\mathbf{S}$.

A major difficulty in UMM-ReID is that the number of identities is unavailable during training. We therefore estimate k directly from the spectrum of $\mathbf{L}$. Let $0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$ denote the eigenvalues of $\mathbf{L}$. We define the eigengap $g(k) = \lambda_{k+1} - \lambda_k$. SG-MDL is constructed to describe the cost of pseudo-labels, and adaptively estimate k,

$$\mathcal{J}_{\text{SG-MDL}} = \frac{\sum_{i=k+1}^{n} \lambda_i}{g(k) + \epsilon} + \beta \frac{k}{n} \log n, \tag{7}$$

where $\epsilon = 10^{-8}$ guarantees numerical stability and β is regularization parameter. The first term of Eq (7) represents the cost of pseudo-labels, and the second represents the clustering complexity by parameter quantity. See Appendix A.1.2 for the specific derivation process. We determine

$$k^{(t)} = arg\min_{k} \mathcal{J}_{\text{SG-MDL}}. \tag{8}$$

The corresponding $k^{(t)}$ eigenvectors form the spectral embedding, on which clustering produces the global pseudo-label $\mathbf{Z}^{(t)}$.

For each pseudo-label k, we construct a cluster prototype from all samples assigned to the cluster,

$$\mathbf{c}_k = \frac{\sum_{i|z_i=k} \hat{\mathbf{y}}_i}{||\sum_{i|z_i=k} \hat{\mathbf{y}}_i||_2}, \tag{9}$$

where $\hat{\mathbf{y}}_i = \mathbf{y}_i / ||\mathbf{y}_i||_2$. These prototypes summarize the global label representation at the beginning of the epoch and provide guidance on label.

3.3 CROSS-MODAL CONTRASTIVE LEARNING

The global pseudo-labels provide supervisory signals for cross-modal contrastive learning. We construct cross-modal positive and negative pairs, break through the limitations of single modality and

realize modal consistency. For an anchor i, positive pairs are$\{\mathbf{y}_i^p, \mathbf{y}_i^j\}^{p\neq j}$ and negative ones are $\{\mathbf{y}_q^p, \mathbf{y}_i^j\}^{q\neq i, p\neq j}$. We use cosine similarity $sim(\cdot,\cdot)$ to express the distance between different modal representations.

The core of cross-modal contrastive learning is to aggregate cross-modal representations of the same identity and disperse cross-modal ones. In the follow-up, we use $P(\mathbf{y}_i^j, \mathbf{y}_q^p) = exp(sim(\mathbf{y}_i^j, \mathbf{y}_q^p)/\tau)$, we get

$$L_{CL} = -\frac{1}{mn}\sum_{\substack{j=1\\i=1}}^{n,m} log\frac{\sum_{j'\neq j}^{m} P(\mathbf{y}_i^j, \mathbf{y}_i^{j'})}{\sum_{\substack{i'\neq i\\j'\neq j}}^{n,m} P(\mathbf{y}_i^j, \mathbf{y}_{i'}^{j'}) + \sum_{j'\neq j}^{m} P(\mathbf{y}_i^j, \mathbf{y}_i^{j'})}, \tag{10}$$

where τ is the temperature coefficient. Although Eq (10) promotes cross-modal consistency, it does not explicitly enforce identity discrimination within each modality, which will affect the learning of identity-discriminative representations. Since pseudo-labels are inherently noisy, directly incorporating all intra-modal negatives is undesirable. Therefore, we estimate the confidence of each pseudo-label by measuring the agreement between each anchor and its corresponding cluster prototype q_i^p,

$$q_i^p = \frac{P(\mathbf{y}_i^p, \mathbf{c}_{z_{y_i}}^p)}{\sum_{j=1}^{k} P(\mathbf{y}_i^p, \mathbf{c}_j^p)}. \tag{11}$$

By obtaining $w_{i,j}^p = q_i^p q_j^p$, We integrate confidence into Eq (10),

$$L_{CL} = -\frac{1}{mn}\sum_{\substack{i=1\\j=1}}^{n,m} log\frac{\sum_{j'\neq j}^{m} P(\mathbf{y}_i^j, \mathbf{y}_i^{j'})}{\sum_{\substack{i'\neq i\\j'\neq j}}^{n,m} P(\mathbf{y}_i^j, \mathbf{y}_{i'}^{j'}) + \sum_{j'\neq j}^{m} P(\mathbf{y}_i^j, \mathbf{y}_i^{j'}) + \sum_{i':z_{i'}\neq z_i} w_{i',i}^j P(\mathbf{y}_i^j, \mathbf{y}_{i'}^j)}. \tag{12}$$

Reliable anchors contribute more to intra modal, whereas uncertain ones have a reduced influence. The resulting objective complements cross-modal consistency with confidence-aware intra-modal identity discrimination.

3.4 Prototype Label Learning

We further use the global cluster prototypes to provide cluster-level loss constrain to stabilize representation learning during minibatch optimization,

$$L_{proto} = -\frac{1}{n}\sum_{i=1}^{n} log\frac{P(\mathbf{y}_i, \mathbf{c}_{z_{y_i}})}{\sum_{j=1}^{k} P(\mathbf{y}_i, \mathbf{c}_j)}. \tag{13}$$

Although pseudo-label assignments remain fixed within each epoch, the prototypes are updated online to track the evolving representations,

$$\mathbf{c}_{z_{y_i}} \leftarrow \mu\mathbf{c}_{z_{y_i}} + (1-\mu)\mathbf{y}_h, \tag{14}$$

where $\mathbf{y}_h$ stands for the hardest sample in a batch and μ is an updating rate Dai et al. (2022).

3.5 Label-aware SF Learning

The preceding parts exploit global pseudo-labels to optimize multi-modal representations, but they do not explicitly determine which local regions provide identity-discriminative information. We therefore further exploit pseudo-labels through label-aware spatial-frequency learning. Specifically, pseudo-labels guide spatial discrimination to enhance the learning of identity-discriminative regions and suppress background noise, while frequency information provides complementary cues.

Spatial Domain For each anchor, taking RGB modality as an example, the last ViT layer produces a global classification and P patches, $\mathbf{H}_i = [\mathbf{h}_i^{cls}, \mathbf{h}_{i,1}, ..., \mathbf{h}_{i,P}]$. We estimate the score of patch $\mathbf{h}_{i,P}$ by combining its consistency with $\mathbf{h}_i^{cls}$ and its distance from the corresponding cluster prototype,

$$a_{i,P}^{RGB} = \alpha Softmax(\frac{(\mathbf{h}_i^{cls})^T\mathbf{h}_{i,P}}{\sqrt{D}}) + (1-\alpha)sim(\mathbf{c}_{z_{h_i}}^{cls}, \mathbf{h}_{i,P}), \tag{15}$$

where D is the size of patches and $\mathbf{c}_i^{cls}$ represent the cluster prototype of the classification. The first term captures instance-specific spatial information, while the second introduces pseudo-label signals. Finally, we average the scores of different modalities to get $a_{i,P}^s$.

Frequency Domain Spatial discrimination alone may not characterize subtle identity-specific details. We adopt the Haar wavelet transform as a lightweight frequency extraction. Its high-frequency subbands provide complementary texture cues for spatial discrimination without introducing additional learnable parameters,

$$\mathbf{I}_i^{ll}, \mathbf{I}_i^{lh}, \mathbf{I}_i^{hl}, \mathbf{I}_i^{hh} = HWT(\mathbf{x}_i^{RGB}). \tag{16}$$

We sum up the separation results obtained by different modalities first. Then, the sum result is transformed by Inverse HWT, and the obtained image highlights the significant features. Finally, we calculate the pixel values in each patch as score $\{a_{i,p}^f\}_{p=1}^P$.

Aggregation By adding the normalized spatial score $a_{i,p}^s$ and frequency score$a_{i,p}^f$, we get $a_{i,P}$ and retain the most informative patches,

$$\mathcal{M}_i = TopK(\{a_{i,p}\}_{p=1}^P).$$

We aggregate the selected patches into a local representation $\mathbf{r}_i = Pool(\{\mathbf{h}_{i,p}|p \in \mathcal{M}_i\})$, and align it with the corresponding classification prototype,

$$L_{SF} = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{P(\mathbf{r}_i, \mathbf{c}_{z_{r_i}}^{cls})}{\sum_{j=1}^{k} P(\mathbf{r}_i, \mathbf{c}_j^{cls})}. \tag{17}$$

Eq (17) encourages the selected patches to enhance identity-discriminative representation and suppress other background noise.

During each epoch, pseudo-labels are fixed, and the ViT is optimized using

$$Loss = L_{CL} + \lambda_1 L_{proto} + \lambda_2 L_{SF}, \tag{18}$$

where λ_1 and λ_2 represent the weight of each part. See Appendix A.3.3 for the sensitivity analysis.

4 Experiment

4.1 Experiment Setup

Dataset and Evaluation We evaluate our method on three multi-modal ReID datasets: RGBNT201, RegDB, and MSVR310, covering person and vehicle re-identification under RGB, near-infrared, and thermal modalities. We follow the standardized processing rules of each dataset, and report Cumulative Matching Characteristic (CMC) accuracy at Rank-1 and Rank-10, together with mean Average Precision (mAP). Detailed dataset statistics are provided in Appendix.

Implementation Datails We use a ViT backbone. Unless otherwise specified, images are resized to 128×256, the batchsize is 128, and the model is optimized using SGD with an initial learning rate of 10^{-3}. In the SF learning, the number of the selected patches is set to 10. For datasets with naturally aligned multi-modal observations , the aligned observations are treated as a single clustering instance. For datasets without explicit cross-modal shot correspondence, each image is treated as an individual clustering instance, and cross-modal positive and negative pairs are determined by shared pseudo-labels.

4.2 Comparison with State-of-the-Art Methods

We compare our method with state-of-the-art (SOTA) methods, and the results are shown in Table 1. It shows that our method has better performance than most UMM-ReID methods. It is worth noting that even compared with supervised methods, our method shows strong competitiveness. In Table 1, our method is compared with unsupervised single-modal ReID, UMM-ReID and supervised multi-modal ReID methods. The result shows:

(1) As a multi-modal method, our method exhibits significantly superior performance compared to SOTA single-modal methods. These methods don't take into account the differences between modalities, which proves that our methods can fuse multi-modal information and effectively make up for the defects of single modality.

Table 1: Performance comparisons with state-of-the-art methods on real datasets. The figures in boldface are best results

Method	Reference	RGBNT201			RegDB			MSVR310		
		Rank-1	Rank-10	mAP	Rank-1	Rank-10	mAP	Rank-1	Rank-10	mAP
Unsupervised Single-modal										
ISE	Zhang et al.	21.34	37.52	19.66	9.91	18.60	10.31	4.23	7.63	5.22
Cluster Contrast	Dai et al.	29.14	55.64	30.12	22.17	46.93	21.36	26.59	38.16	26.74
DCMIP	Zou et al.	28.88	51.20	29.00	27.61	43.01	28.69	14.23	24.27	15.31
PAT	Chen et al.	34.02	55.69	35.61	33.41	52.87	35.81	27.83	40.13	29.64
Unsupervised Multi-modal										
H2H	Liang et al.	17.75	39.04	15.61	13.91	30.39	12.72	7.28	15.26	8.81
H2H+AGW	Liang et al.	33.56	60.28	32.91	23.81	45.31	18.87	15.61	25.97	14.37
OTLA	Wang et al.	42.19	59.28	43.66	32.90	50.19	29.70	22.59	38.29	23.30
CHCR	Pang et al.	60.02	79.84	59.31	**69.31**	82.60	64.74	29.82	54.11	28.96
SCL	Zhang et al.	55.24	77.78	51.78	67.20	82.02	64.05	25.68	42.23	26.51
Ours	-	**69.15**	**85.24**	**68.33**	68.63	**86.72**	**65.35**	**35.68**	**61.57**	**34.37**
Supervised Multi-modal										
DEnet	Zheng et al.	42.22	64.49	42.37	45.26	70.39	44.64	21.32	36.41	25.26
EDITOR	Zhang et al.	68.32	88.21	66.48	85.59	92.14	82.85	43.12	61.53	35.79
TOP-ReID	Wang et al.	76.61	89.38	72.30	72.48	85.47	70.25	33.72	58.91	30.15

Table 2: Ablation experimental results of module discarding or module replacement.

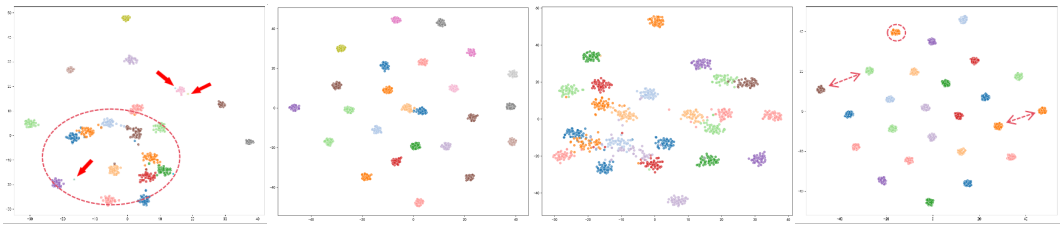
Method	①	②	③	RGBNT201			RegDB			MSVR310		
				Rank-1	Rank-10	mAP	Rank-1	Rank-10	mAP	Rank-1	Rank-10	mAP
A1	×	✓	✓	36.94	60.28	38.21	42.53	60.01	40.78	18.45	39.21	17.39
A2	✓	×	✓	43.48	65.27	44.59	48.19	68.66	46.41	21.81	43.27	20.77
A3	✓	non-w	✓	61.33	81.24	59.51	63.59	76.82	61.27	28.88	50.16	29.17
A4	✓	✓	×	58.37	75.71	60.85	59.22	75.38	55.63	27.30	45.23	29.61
A5	✓	✓	a^s	62.26	78.71	61.43	60.10	74.37	58.51	29.20	50.84	30.68
Ours	✓	✓	✓	69.15	85.24	68.33	68.63	86.72	65.35	35.68	61.57	34.37

(2) Compared with UMM-ReID methods, although H2H and H2H+AGW deviate from purely unsupervised methods, our method still achieves substantially better performance than them in the absence of labels. For the second-ranked method (CHCR), our method demonstrates a deficit in Rank-1 but outperforms it in both Rank-10 and mAP. This performance pattern underscores two key advantages of our framework: First, compared to simple bimodal scenarios (RegDB), our method leverages cross-modal contrastive learning to more effectively address cross-modal consistency in complex multi-modalities (RGBNT201, MSVR310). Second, label-aware SF learning enhance its capability to capture identity-discriminative representations (RGBNT201, MSVR310).

(3) Compared with supervised multi-modal methods, ours has approached and even outperformed some of them, reaching a level close to SOTA. It narrows the performance gap brought by real labels.

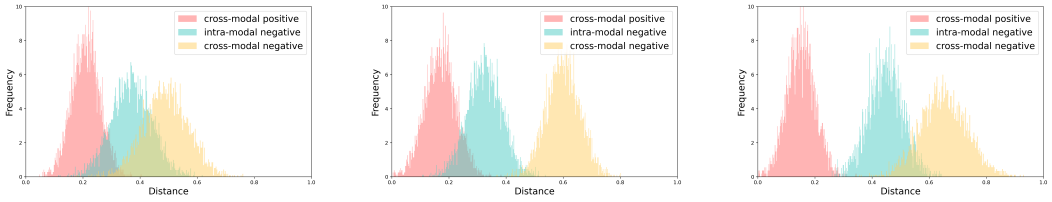
4.3 Ablation Study

For UMM-ReID, our method optimization mainly consists of three modules, ①adaptive global pseudo-label, ②cross-modal contrastive learning and ③label-aware SF learning. We have conducted ablation experiments, as shown in Table 2, to verify the effectiveness of each modules. Note that compared with our method, A1 uses DBSCAN to construct local pseudo-labels in minibatch, A2 does not use any contrastive learning, A3 does not introduce confidence in cross-modal contrastive learning, A4 does not use any discrimination guidance, and A5 only uses spatial discrimination. We introduce these methods in more detail in Appendix A.3.1 and analyze the time complexity in Appendix A.3.2.



(a) A1 (b) A1-b (c) A1-c (d) Ours

Figure 2: Comparison of t-SNE and representation distribution on RGBNT201. Different colors represent different labels. A1-b and A1-c are two variants of A1, using SC with Gaussian kernel of 1 and 100 respectively to construct global pseudo-label.



(a) A2 (b) A3 (c) Ours

Figure 3: Distance distribution of cross-modal positive, intra-modal negative and cross-modal negative sample pairs on RGBNT201.

4.3.1 Analysis of Adaptive Global Pseudo-Label

The purpose of this module has two points: one is to solve the sensitivity of traditional clustering to data distribution by using adaptive SC, and the other is to judge the number of clusters by SG-MDL. We will analyze the effectiveness of SG-MDL in Appendix A.3.1. As shown in Table 2, local pseudo-labels in minibatch can not provide a stable supervision signal, our method gains the increase of evaluation metrics on datasets compared with A1.

To further verify the gain effect of adaptive SC on global pseudo-labels, Fig 2 shows the clustering distribution of different methods. It can be observed that:

(1) A1 can not effectively distinguish different labels in dense areas. The performance of two variants A1-b and A1-c with different Gaussian kernels is significantly different, which indicates that traditional SC needs manual parameter adjustment to find the adaptive distance of data.

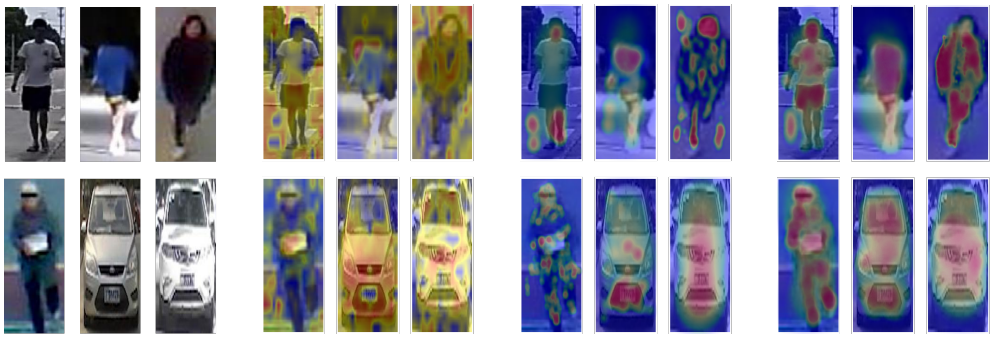
(2) Compared with the other three methods, ours can effectively gather representations inside cluster and increase the distance between different clusters, that is, adaptive SC can adapt to complex data distribution and dynamically adjust the distance, thus improving the accuracy of pseudo-labels.

4.3.2 Analysis of Cross-Modal Contrastive Learning

The purpose of this module is to enhance identity discrimination and cross-modal consistency. As shown in Table 2, our method is obviously superior to A2 and A3. It is worth noting that, compared with A3 using traditional contrastive learning, by introducing confidence of intra-modal negative pairs, our method achieves considerable evaluation metrics on datasets.

To further elaborate on the effectiveness of cross-modal contrastive learning, we visualize the distance distribution in Fig 3. Two phenomena can be observed:

(1) Compared with A2, both A3 and Ours exhibit better alignment between cross-modal positive-pair distributions, showing that cross-modal contrastive learning effectively reduces modality discrepancies for the same identity.



(a) Original (b) A4 (c) A5 (d) Ours

Figure 4: Comparison of smoothed attention visualizations between our method and ablation methods, showing progressively enhanced identity-discriminative responses through spatial and frequency guidance.

(2) Compared with A3, Ours produces a clearer separation between intra-modal positive and negative pairs. This supports our pseudo-label-aware confidence strategy, which selectively introduces reliable intra-modal negatives to enhance identity discrimination.

4.3.3 Analysis of Label-aware SF learning

The purpose of Label-aware SF learning aims to enhance the learning of identity-discriminative regions and suppress background noise. As shown in Table 2, the progressive improvements from A4 to A5 and finally to our full model demonstrate the complementary roles of label-aware SF learning.

To further investigate how this module affects discriminative regions, we present the original images and attention heatmaps of different methods on three datasets in Fig 4. The experiment observed the following three key phenomena:

(1) A4 does not use any discrimination guidance, yet it still possesses a basic ability to focus on key features. The core reason lies in that cross-modal learning indirectly guides the model to focus on identity-related representations. However, its attention tends to be diffuse and is frequently distracted by background regions.

(2) A5 introduces label-aware spatial discrimination and achieves a consistent improvement over A4, which show label-aware spatial discrimination provide effective guidance for suppressing irrelevant regions. Nevertheless, the lack of frequency cues limits its ability to characterize subtle identity-specific details.

(3) Our full model further incorporates frequency cues on top of the label-aware spatial guidance. The complementary frequency cues enhance fine-grained patterns that are difficult to distinguish using spatial discrimination alone. Moreover, this module is performed adaptively for each instance, allowing the model to preserve different discriminative evidence according to the input.

5 Conclusion

We present a unified framework for contrastive label-aware SF learning for UMM-ReID. Adaptive global pseudo-labels capture relable identity structures and provide supervisory guidance for subsequent representation learning. On this basis, our cross-modal contrastive learning selectively incorporates reliable intra-modal negative pairs, complementing cross-modal consistency with stronger identity discrimination under noisy pseudo-label. Label-aware SF learning further strengthens the learning of identity-discriminative regions and suppress background noise. Extensive experiments and analyses demonstrate the effectiveness of these designs, highlighting the benefit of jointly modeling global identity information, contrastive relations, and SF discrimination for robust multi-modal representations.

REFERENCES

Xing Cai, Tong Zhang, Yide Qiu, Xudong Wang, Baotong Su, and Zhen Cui. Minimum description length-driven fragment mining for pretraining molecule property prediction model. *IEEE Transaction on Computational Biology and Bioinformatics*, 23(4):1747–1760, 2026.

Shuoyi Chen, Mang Ye, Xingping Dong, and Bo Du. Perception assisted transformer for unsupervised object re-identification. *IEEE Transactions on Image Processing*, 34:2112–2123, 2025.

Yehansen Chen, Lin Wan, Zhihang Li, Qianyan Jing, and Zongyuan Sun. Neural feature search for rgb-infrared person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 587–597, 2021.

Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7298–7308, 2022.

Zuozhuo Dai, Guangyuan Wang, Weihao Yuan, Siyu Zhu, and Ping Tan. Cluster contrast for unsupervised person re-identification. In *Computer Vision - ACCV 2022 - 16th Asian Conference on Computer Vision*, volume 13846, pp. 319–337, 2022.

Heng Fan and Haibin Ling. Siamese cascaded region proposal networks for real-time visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7952–7961, 2019.

Yixiao Ge, Feng Zhu, Dapeng Chen, Rui Zhao, and Hongsheng Li. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*, 2020.

Ju He, Jieneng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*, pp. 852–860, 2022.

Wei-Yen Hsu and Chun-Hsiang Wu. Wavelet structure-texture-aware super-resolution for pedestrian detection. *Information Sciences*, 691:121612, 2025.

Yan Huang, Qiang Wu, Jingsong Xu, and Yi Zhong. SBSGAN: suppression of inter-domain background shift for person re-identification. In *2019 IEEE/CVF International Conference on Computer Vision*, pp. 9526–9535, 2019.

Ugur Kart, Alan Lukezic, Matej Kristan, Joni-Kristian Kamarainen, and Jiri Matas. Object tracking by reconstruction with view-specific discriminative correlation filters. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1339–1348, 2019.

Hongchao Li, Chenglong Li, Xianpeng Zhu, Aihua Zheng, and Bin Luo. Multi-spectral vehicle re-identification: A challenge. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pp. 11345–11353, 2020.

Peixia Li, Boyu Chen, Wanli Ouyang, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Gradnet: Gradient-guided network for visual object tracking. In *2019 IEEE/CVF International Conference on Computer Vision*, pp. 6161–6170, 2019.

Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152–159, 2014.

Wenqi Liang, Guangcong Wang, Jianhuang Lai, and Xiaohua Xie. Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. *IEEE Transactions on Image Processing*, 30:6392–6407, 2021.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20973–20982, 2022.

Shaolei Liu, Manning Wang, and Zhijian Song. Wavefuse: A unified unsupervised framework for image fusion with discrete wavelet transform. In *Neural Information Processing - 28th International Conference*, volume 13111, pp. 162–174, 2021.

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.

Feiping Nie, Xiaoqian Wang, Michael I. Jordan, and Heng Huang. The constrained laplacian rank algorithm for graph-based clustering. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 1969–1976, 2016.

Yanwei Pang, Jin Xie, Feiping Nie, and Xuelong Li. Spectral clustering by joint spectral embedding and spectral rotation. *IEEE Transactions on Cybernetics*, 50(1):247–258, 2020.

Zhiqi Pang, Chunyu Wang, Lingling Zhao, Yang Liu, and Gaurav Sharma. Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification. *IEEE Transaction on Circults and Systems for Video Technology*, 34(4):2706–2718, 2024.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pp. 13937–13949, 2021.

Tongzhen Si, Fazhi He, Penglei Li, Yupeng Song, and Linkun Fan. Diversity feature constraint based on heterogeneous data for unsupervised person re-identification. *Information Processing and Management*, 60(3):103304, 2023.

Jiangming Wang, Zhizhong Zhang, Mingang Chen, Yi Zhang, Cong Wang, Bin Sheng, Yanyun Qu, and Yuan Xie. Optimal transport for label-efficient visible-infrared person re-identification. In *Computer Vision - ECCV 2022: 17th European Conference*, volume 13684, pp. 93–109, 2022.

Pengfei Wang, Minglian Wang, and Dongzhi He. Multi-scale feature pyramid and multi-branch neural network for person re-identification. *Visual Computer*, 39(10):5185–5197, 2023.

Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *Computer Vision - ECCV 2024 - 18th European Conference*, volume 15087, pp. 286–303, 2024a.

Yuhao Wang, Xuehu Liu, Pingping Zhang, Hu Lu, Zhengzheng Tu, and Huchuan Lu. Top-reid: Multi-spectral object re-identification with token permutation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*, pp. 5758–5766, 2024b.

Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *IEEE International Conference on Computer Vision*, pp. 5390–5399, 2017.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4330–4339, 2021.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

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

Bin Yang, Jun Chen, and Mang Ye. Top-k visual tokens transformer: Selecting tokens for visible-infrared person re-identification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17117–17126, 2024.

Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit sample extension for unsupervised person re-identification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7359–7368, 2022.

Yuxin Zhang, Zhu Teng, and Baopeng Zhang. Semantic consistency learning for unsupervised multi-modal person re-identification. *Image and Vision Computing*, 155:105434, 2025.

Yang Zhao, Zixuan Bi, Peican Zhu, Aihong Yuan, and Xuelong Li. Deep spectral clustering with projected adaptive feature selection. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1–12, 2025a.

Yang Zhao, Daidai Zhu, Aihong Yuan, and Xuelong Li. Consensus multi-view spectral clustering network with unified similarity. *Neural Networks*, 190:107647, 2025b.

Aihua Zheng, Zi Wang, Zi-Han Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pp. 3529–3537, 2021.

Aihua Zheng, Xianpeng Zhu, Chenglong Li, Jin Tang, and Jixin Ma. Multi-spectral vehicle re-identification with cross-directional consistency network and a high-quality benchmark. *CoRR*, abs/2208.00632, 2022.

Aihua Zheng, Ziling He, Zi Wang, Chenglong Li, and Jin Tang. Dynamic enhancement network for partial multi-modality person re-identification. *CoRR*, abs/2305.15762, 2023.

Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro. In *IEEE International Conference on Computer Vision*, pp. 3774–3782, 2017.

Chang Zou, Zeqi Chen, Zhichao Cui, Yuehu Liu, and Chi Zhang. Discrepant and multi-instance proxies for unsupervised person re-identification. In *IEEE/CVF International Conference on Computer Vision*, pp. 11024–11034, 2023.

A APPENDIX

A.1 MATHEMATICAL THEORY DERIVATION

A.1.1 AFFINITY MATRIX DERIVATION

In this part, we will deduce the adaptive affinity matrix $\mathbf{S}$ in detail.

The loss function of spectral clustering is

$$\begin{aligned} &\min \sum_{j=1}^{m} (||\mathbf{y}_i - \mathbf{y}_j||_F^2 s_{ij} + \gamma s_{ij}^2). \\ &s.t. \forall i,\ \mathbf{s}_i \mathbf{1} = 1,\ \mathbf{s}_i \geq 0. \end{aligned} \tag{19}$$

Eq (19) is a convex function because all parts involving $\mathbf{S}$ are linear functions and quadratic functions. Given $d_{ij} = ||\mathbf{y}_i - \mathbf{y}_j||_F^2$, we denote $\mathbf{d}_i$ with its j_{th} element as d_{ij}, thus Eq (19) can be optimized to

$$\min_{\mathbf{s}_i \mathbf{1}=1,\, \mathbf{s}_i \geq 0} ||\mathbf{s}_i + \frac{1}{2\gamma} \mathbf{d}_i||_F^2. \tag{20}$$

To ensure the sparsity of $\mathbf{S}$, we learn the affinities with the maximal γ such that the optimal solution $\mathbf{s}_i$ to Eq (20) has exactly q nonzero values. i.e., the l_0 norm of $\mathbf{s}_i$ is constrained to be q. To this end, we solve the following problem

$$\max_{\gamma, ||\mathbf{s}_i||_0 = q} \gamma. \tag{21}$$

The Lagrangian function of Eq (20) is

$$L(\mathbf{s}_i, \psi, \beta_i) = \frac{1}{2} ||\mathbf{s}_i + \frac{1}{2\gamma} \mathbf{d}_i||_F^2 - \psi(\mathbf{s}_i^T \mathbf{1} - 1) - \beta_i^T \mathbf{s}_i, \tag{22}$$

where ψ and $\beta_i \geq 0$ are the Lagrange multipliers.

The optimal solution $\mathbf{s}_i$ should satisfy that the derivative of Eq (22) w.r.t. $\mathbf{s}_i$ is equal to zero, so we have

$$\mathbf{s}_i + \frac{1}{2\gamma} \mathbf{d}_i - \psi \mathbf{1} - \beta_i = 0. \tag{23}$$

Then for the j-th element of $\mathbf{s}_i$, we have

$$s_{ij} + \frac{1}{2\gamma} \mathbf{d}_{ij} - \psi - \beta_{ij} = 0. \tag{24}$$

Theorem 1. *At the optimal solution $\mathbf{x}^*$ of the following constrained optimization problem*

$$\min_{g_i(\mathbf{x} \leq 0)} f(\mathbf{x}), \tag{25}$$

if Lagrange multiplier $\lambda_i \geq 0$ correspondence and inequality constraint $g_i(\mathbf{x}) \leq 0$, then the following conditions hold

$$\lambda_i g_i(\mathbf{x}^*) = 0, \quad \forall i = 1, ..., m. \tag{26}$$

Proof. Define the Lagrange function

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}), \tag{27}$$

where $\lambda_i \geq 0$ is a Lagrange multiplier and $g_i(\mathbf{x})$ is the inequality constraint. At the optimal solution $\mathbf{x}^*$, the gradient condition requires

$$\nabla_x L(\mathbf{x}^*, \lambda) = \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) = 0. \tag{28}$$

The complementary relaxation condition requires

$$\lambda_i g_i(\mathbf{x}^*) = 0, \quad \forall i = 1, ..., m. \tag{29}$$

When $g_i(\mathbf{x}) < 0$, the inequality constraint is inactive, $\mathbf{L}$ does not depend on it, and $\lambda_i = 0$. When $g_i(\mathbf{x}) = 0$, the inequality constraint takes effect, $\lambda \geq 0$, which ensures the gradient condition of $\mathbf{L}$. □

Noting that $s_{ij}\beta_{ij} = 0$ according to Theorem 1, from Eq (24) we have

$$s_{ij} = (-\frac{1}{2\gamma}\mathbf{d}_{ij} + \psi)_+. \tag{30}$$

In order not to lose generality, suppose $\mathbf{d}_{i1}, \mathbf{d}_{i2}, ..., \mathbf{d}_{im}$ are ordered from small to large. To impose $s_{ii} = 0$, we always let $\mathbf{d}_{ii}$ place this value last despite having $\mathbf{d}_{ii} = 0$. According to the constrain $||\mathbf{s}_i||_0 = q$ in Eq (21), we know $s_{iq} > 0$ and $s_{i,q+1} = 0$. Therefore, we have

$$-\frac{1}{2\gamma}\mathbf{d}_{iq} + \psi > 0, \ and \ -\frac{1}{2q}\mathbf{d}_{i,q+1} + \psi \leq 0. \tag{31}$$

According to Eq (30) and the constraint $\mathbf{s}_i^T\mathbf{1} = 1$ in Eq (19), we have

$$\sum_{j=1}^{q}(-\frac{1}{2\gamma}\mathbf{d}_{ij} + \psi) = 1 \Rightarrow \psi = \frac{1}{q} + \frac{1}{2q\gamma}\sum_{j=1}^{q} d_{ij}. \tag{32}$$

So we have the following inequality for γ according to Eq (31) and Eq (32):

$$\frac{q}{2}d_{iq} - \frac{1}{2}\sum_{j=1}^{q} d_{ij} < \gamma \leq \frac{q}{2}d_{i,q+1} - \frac{1}{2}\sum_{j=1}^{q} d_{ij}. \tag{33}$$

Therefore, to obtain the optimal solution $\mathbf{s}_i$ to Eq (21) that has exactly α nonzero values, the maximal γ is

$$\gamma = \frac{q}{2}d_{i,q+1} - \frac{1}{2}\sum_{j=1}^{q} d_{ij}. \tag{34}$$

According to Eq (30), (32) and (34), we get the optimal affinities s_{ij} as follows

$$s_{ij} = \begin{cases} \frac{d_{i,q+1} - d_{ij}}{q d_{i,q+1} - \sum_{h=1}^{q} d_{ih}} & j \leq q \\ 0 & j > q \end{cases}. \tag{35}$$

In most cases, $q < 10$ is likely to yield reasonable results

A.1.2 Mathematical Motivation of SG-MDL

We formulate adaptive cluster-number estimation as a spectral model selection problem inspired by the minimum description length (MDL) principle, which balances representation fidelity against model complexity.

Let $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ be the eigendecomposition of the graph Laplacian, with eigenvalues $0 \leq \lambda_1 \leq \cdots \leq \lambda_n$. For a candidate cluster number k, the spectral representation is constructed from the first k eigenvectors. The residual spectral energy outside the selected subspace is given by

$$R(k) = \mathrm{tr}\left((\mathbf{I} - \mathbf{U}_k\mathbf{U}_k^\top)\mathbf{L}\right) = \sum_{i=k+1}^{n} \lambda_i. \tag{36}$$

Although increasing k reduces the residual energy, it does not necessarily yield a well-separated spectral representation. We therefore incorporate the eigengap $g(k) = \lambda_{k+1} - \lambda_k$, which characterizes the separation between the selected subspace and the remaining spectrum. Motivated by spectral perturbation theory, we define the gap-normalized residual energy as $R(k)/(g(k)+\epsilon)$ where ϵ is a perturbation term, favoring spectral representations with low residual energy and pronounced spectral separation.

Following the MDL principle, we further introduce a complexity penalty proportional to the number of clusters. Assuming that the effective model dimension grows linearly with k, the asymptotic model description length scales as $O(k \log n)$. Combining the spectral criterion with a normalized complexity penalty yields,

$$\mathcal{J}_{\text{SG-MDL}} = \frac{\sum_{i=k+1}^{n} \lambda_i}{g(k) + \epsilon} + \beta\frac{k}{n}\log n, \tag{37}$$

where β is is the balance coefficient, which is obtained by network training. This criterion jointly considers residual spectral energy, spectral separation, and model complexity, enabling adaptive selection of the spectral representation dimension without requiring a predefined cluster count.

A.2 ALGORITHM

Algorithm 1 Training Procedure of Contrastive Label-aware Spatial-Frequency Learning

Input: Unlabeled multi-modal training set $\mathbf{X} = \{\mathbf{x}_i^m\}$, ViT f_θ, number of epochs T
Output: Trained encoder f_θ
for $t = 1$ to T **do**
  Global adaptive pseudo-label
  Extract global representations for all training samples:

$$\mathbf{y}_i^m = f_\theta(\mathbf{x}_i^m)$$

  Construct clustering representations from the extracted representations:

$$\mathbf{y}_i = \mathrm{Stack}[\mathbf{y}_i^1, \mathbf{y}_i^2, ..., \mathbf{y}_i^m]$$

  Construct the adaptive affinity matrix $\mathbf{S}$ from $\{\mathbf{y}_i\}$
  Compute the normalized graph Laplacian $\mathbf{L}$ and its eigenvalues
  Estimate the cluster number K_t using the SG-MDL criterion:

$$k^{(t)} = \arg\min_k \mathcal{J}_{\text{SG-MDL}}(k)$$

  Perform spectral clustering with $k^{(t)}$ clusters to obtain pseudo-labels $\mathbf{Z}^{(t)}$
  Compute the cluster prototype dictionary

$$\mathbf{C}^{(t)} = \{\mathbf{c}_i^{(t)}\}_{i=1}^{k^{(t)}}$$

  Fix $\mathbf{Z}^{(t)}$ and $\mathbf{C}^{(t)}$ within epoch t
  Minibatch representation learning
  for each minibatch $\mathcal{B} \subset \mathbf{X}$ **do**
    Forward the minibatch through f_θ and obtain the global token and patch tokens:

$$\mathbf{y}_i = \mathbf{H}_i = [\mathbf{h}_i^{cls}, \mathbf{h}_{i,1}, \ldots, \mathbf{h}_{i,P}]$$

    Retrieve pseudo-labels and prototypes for samples in $\mathcal{B}$
    Construct positive pairs and negative pairs
    Compute the cross-modal contrastive loss $\mathcal{L}_{CL}$
    Compute the prototype label learning loss $\mathcal{L}_{proto}$
    Compute label-aware spatial importance score $a_{i,P}^s$ for each patch
    Combine it with complementary frequency cues to get $a_{i,P}$ and retain the most informative patches

$$\mathcal{M}_i = TopK(\{a_{i,p}\}_{p=1}^P)$$

    Compute the label-aware SF learning loss $\mathcal{L}_{\text{SF}}$
    Optimize the total loss: $\mathcal{L} = \mathcal{L}_{CL} + \lambda_1 \mathcal{L}_{proto} + \lambda_2 \mathcal{L}_{SF}$
    Update the network parameters θ
  end for
end for
return f_θ

Table 3: Performance comparisons with more methods on double-modal datasets.

Method	Reference	SYSU-MM01			RegDB			HITSZ-VCM		
		Rank-1	Rank-10	mAP	Rank-1	Rank-10	mAP	Rank-1	Rank-10	mAP
Unsupervised Single-modal										
ISE	Zhang et al.	20.54	57.50	20.39	9.91	18.60	10.31	6.38	15.21	7.33
Cluster Contrast	Dai et al.	20.16	59.27	22.00	22.17	46.93	21.36	17.77	30.15	19.40
DCMIP	Zou et al.	18.37	54.00	19.27	27.61	43.01	28.69	17.51	32.47	15.71
PAT	Chen et al.	21.41	59.53	21.11	33.41	52.87	35.81	31.79	57.38	27.14
Unsupervised Multi-modal										
H2H	Liang et al.	25.49	63.85	25.16	13.91	30.39	12.72	21.54	57.20	20.39
H2H+AGW	Liang et al.	30.15	65.92	29.40	23.81	45.31	18.87	37.13	51.40	38.11
CMRR	Liang et al.	42.13	72.35	43.38	25.47	52.78	27.89	40.10	61.78	42.17
ADCA	Yang et al.	45.51	72.78	47.99	25.20	55.29	22.70	42.51	65.31	40.72
DFC	Si et al.	40.92	79.07	36.20	20.92	41.73	17.13	40.33	61.57	39.21
OTLA	Wang et al.	29.98	71.79	27.13	32.90	50.19	29.70	37.24	60.11	35.35
CHCR	Pang et al.	47.72	**87.27**	45.34	**69.31**	82.60	64.74	51.00	70.21	45.79
SCL	Zhang et al.	45.34	85.11	40.92	67.20	82.02	64.05	50.38	70.21	45.79
Ours	-	**60.13**	83.24	**59.47**	68.63	**86.72**	**65.35**	**57.31**	**79.38**	**55.45**
Supervised Multi-modal										
DEnet	Zheng et al.	41.30	81.21	42.50	45.26	70.39	44.64	41.33	68.23	40.69
EDITOR	Zhang et al.	61.58	85.40	60.31	85.59	92.14	82.85	56.10	76.28	51.17
TOP-ReID	Wang et al.	64.23	90.15	66.21	72.48	85.47	70.25	63.14	80.73	55.31

Table 4: Performance comparisons with more methods on triple-modal datasets.

Method	Reference	RGBNT100			RGBNT201			MSVR310		
		Rank-1	Rank-10	mAP	Rank-1	Rank-10	mAP	Rank-1	Rank-10	mAP
Unsupervised Single-modal										
ISE	Zhang et al.	32.51	42.87	33.10	21.34	37.52	19.66	4.23	7.63	5.22
Cluster Contrast	Dai et al.	40.87	67.42	40.11	29.14	55.64	30.12	26.59	38.16	26.74
DCMIP	Zou et al.	42.45	64.51	42.89	28.88	51.20	29.00	14.23	24.27	15.31
PAT	Chen et al.	65.00	82.57	64.41	34.02	55.69	35.61	27.83	40.13	29.64
Unsupervised Multi-modal										
H2H	Liang et al.	45.82	72.31	41.50	17.75	39.04	15.61	7.28	15.26	8.81
H2H+AGW	Liang et al.	52.33	76.72	51.24	33.56	60.28	32.91	15.61	25.97	14.37
CMRR	Liang et al.	57.01	82.43	55.90	40.17	70.19	40.23	17.24	30.51	16.51
ADCA	Yang et al.	55.73	80.71	52.41	40.51	72.23	39.40	20.20	33.17	23.81
DFC	Si et al.	57.23	83.51	55.70	43.37	75.21	42.22	21.37	34.15	23.85
OTLA	Wang et al.	58.54	82.00	58.17	42.19	59.28	43.66	22.59	38.29	23.30
CHCR	Pang et al.	59.41	81.77	58.35	60.02	79.84	59.31	29.82	54.11	28.96
SCL	Zhang et al.	67.14	91.88	65.20	55.24	77.78	51.78	25.68	42.23	26.51
Ours	-	**81.20**	**96.35**	**82.15**	**69.15**	**85.24**	**68.33**	**35.68**	**61.57**	**34.37**
Supervised Multi-modal										
DEnet	Zheng et al.	75.00	94.36	74.43	42.22	64.49	42.37	21.32	36.41	25.26
EDITOR	Zhang et al.	96.42	99.87	82.13	68.32	88.21	66.48	43.12	61.53	35.79
TOP-ReID	Wang et al.	96.57	99.50	81.29	76.61	89.38	72.30	33.72	58.91	30.15

A.3 SUPPLEMENTARY EXPERIMENTS

To evaluate the performance of our method, we use six multi-modal benchmark datasets.

(1) SYSU-MM01 Pang et al. (2024) is a dataset containing RGB and IR modalities, with 491 pedestrian identities and 303,420 images captured by six cameras, which has significant modality discrepancies and complex indoor and outdoor scenes.

(2) RegDB Nguyen et al. (2017) is a small-scale multi-modal dataset with 412 pedestrian identities, each of which contains 10 RGB images and 10 TIR images, with a total of 4,120 image binary groups, with relatively little background interference.

(3) HITSZ-VCM Lin et al. (2022) is a dataset with 927 pedestrian identities, containing 11,785 visible and 10,078 infrared tracklets captured by 12 cameras, with diverse scenes and considerable variations in viewpoints and pedestrian appearance.

(4) RGBNT100 Zheng et al. (2021) is a dataset containing RGB, NIR and TIR modalities, with 100 vehicle identities and 17,250 image triplets, which has considerable cross-modal heterogeneity and diverse vehicle viewpoints.

(5) RGBNT201 Zheng et al. (2021) is a dataset containing RGB, NIR and TIR modalities, with 201 pedestrian identities and 4,787 image triplets, which has great modal heterogeneity.

(6) MSVR310 Zheng et al. (2022) is a large-scale multi-modal vehicle ReID dataset, which contains three modes, a total of 310 vehicle identities and 2,087 image triples, and has complex visual scenes and subtle distinguishing features.

We compare our method with more methods on more datasets, and the results are shown in Table 3 and Table 4. The results show:

(1) To examine whether multi-modal data can be effectively exploited without explicit cross-modal learning, we additionally adapt single-modal Re-ID methods by normalizing and concatenating inputs from different modalities into a unified representation. As shown in Table 3 4, these methods generally struggle to achieve competitive performance. This comparison shows that simply increasing the available modal information is insufficient, learning the consistency and discrepancy cross-modalities is important for effectively exploiting multi-modal data.

(2) Compared with existing UMM-ReID methods, it can be clearly seen that our advantage is more pronounced on triple-modal datasets than on double-modal ones. Our framework appears to gain more from scenarios involving richer modalities. This performance is consistent with the design of our method: cross-modal contrastive learning explicitly models cross-modal relations, while label-aware spatial-frequency learning further guides. The results therefore suggest that the proposed framework is particularly effective when representation learning requires integrating information from multiple modalities.

(3) We further compare our method with supervised multi-modal approaches. Despite using no supervised signals during training, our method achieves performance competitive with many supervised methods. These results provide additional evidence that adaptive global pseudo-label supervision can recover meaningful identity structure and serve as effective guidance for subsequent representation learning.

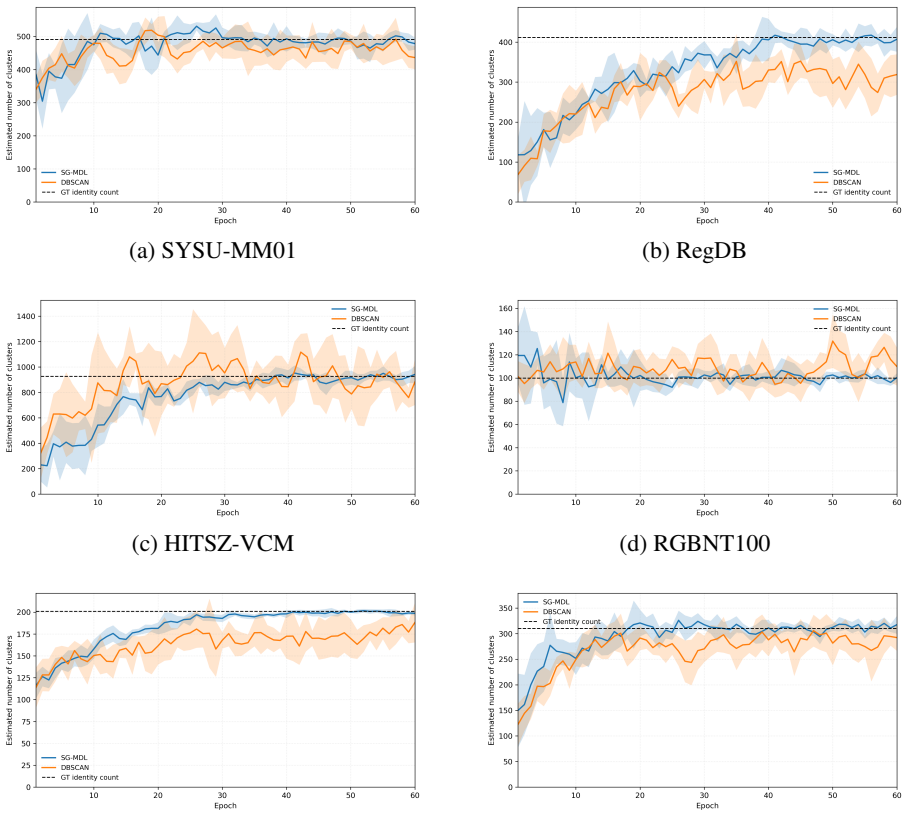
A.3.1 ABLATION STUDY

In this section, we will analyze the effectiveness of three modules through ablation experiments. There are 5 contrast methods to compare ablation experiments:

(1) A1 does not include the global adaptive pseudo-label module. The network is optimized by minibatch samples, and local pseudo-labels are constructed by DBSCAN. Its variants A1-b and A1-c introduce global pseudo-labels, but use Gaussian kernel to calculate affinity matrix.

(2) A2 does not include cross-modal contrastive learning, and the overall loss of the network only includes L_{proto} and L_{SF}, so the network cannot learn cross-modal consistency. And A3 only uses basic contrastive learning, and does not include the confidence introduced by pseudo-labels.

(3) A4 does not include label-aware SF learning module, and only cross-modal contrastive learning indirectly guides the model focusing. A5 only introduces label-aware spatial information learning, and does not contain more information domains.



(a) SYSU-MM01 (b) RegDB

(c) HITSZ-VCM (d) RGBNT100

(e) RGBNT201 (f) MSVR310

Figure 5: The standard deviation of the evolution of the estimated cluster number across training epochs in two methods (DBSCAN, SG-MDL).

Effectiveness of SG-MDL To further investigate the effectiveness of the proposed SG-MDL criterion, we conduct additional experiments on six datasets. Specifically, we analyze the evolution of the estimated cluster number throughout training and compare the downstream retrieval performance of different clustering strategies, including fixed-K, DBSCAN, and our SG-MDL criterion.

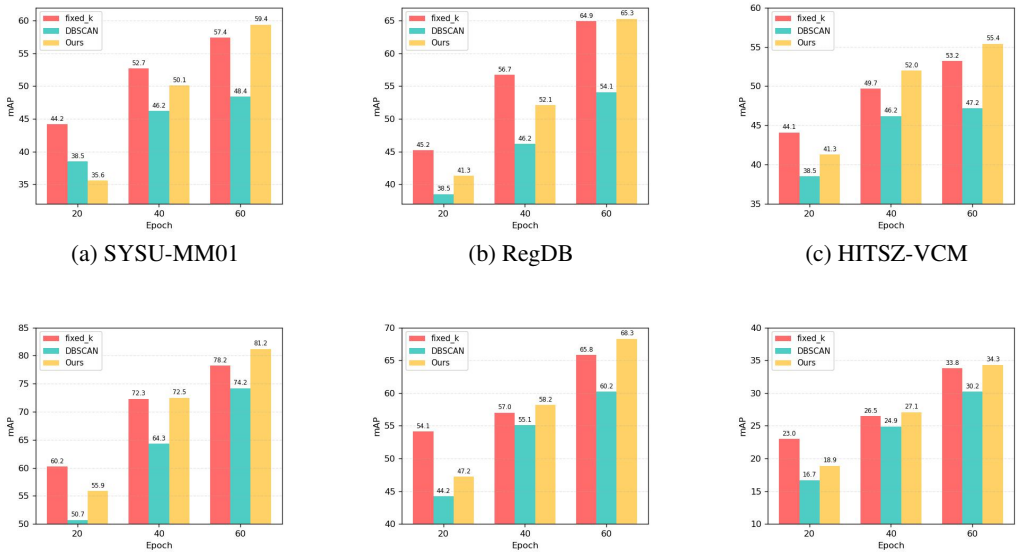
In Fig 5, we use two methods to estimate the number of global clusters, namely DBSCAN and our SG-MDL strategy.

(1) Fig 5 shows that SG-MDL shows gradual adjustment of GT identity count on most datasets, and shows relatively stable estimates in later training stages. In contrast, DBSCAN shows larger fluctuations and more pronounced deviations from GT identity count. These observations suggest that SG-MDL can provide more effective cluster number estimation with the development of representation learning.

(2) It is worth noting that the estimated number of clusters does not necessarily converge to GT identity count accurately. This is because pseudo-labels of unsupervised ReID usually have the phenomenon of over-segmentation, but the samples in the cluster are more similar and the noise of pseudo-labels is less, which is helpful for contrastive learning to learn more accurate representation and will not affect the final downstream retrieval performance.

In Fig 6, we further study whether SG-MDL is beneficial to representation learning by comparing the effects of different estimation strategies on retrieval performance (fixed_k, DBSCAN, ours).

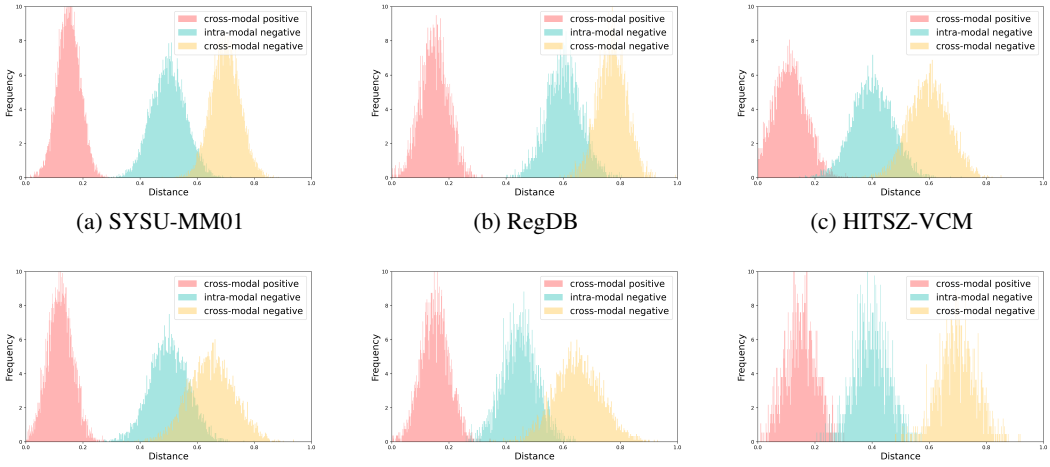
(1) As shown in Fig 6, SG-MDL finally achieved higher mAP than DBSCAN with the iteration of epoch, indicating that the choice of clustering strategy can significantly affect the subsequent representation learning.



(a) SYSU-MM01 (b) RegDB (c) HITSZ-VCM

(d) RGBNT100 (e) RGBNT201 (f) MSVR310

Figure 6: Comparison of retrieval performance under different clustering strategies (fixed_k, DBSCAN, Ours).



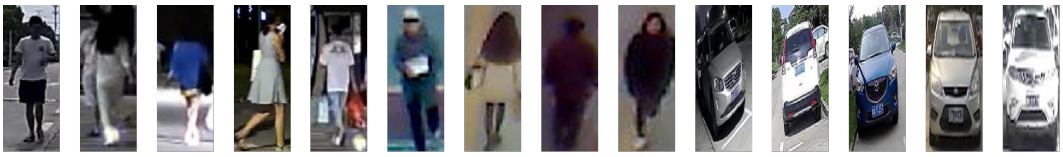
(a) SYSU-MM01 (b) RegDB (c) HITSZ-VCM

(d) RGBNT100 (e) RGBNT201 (f) MSVR310

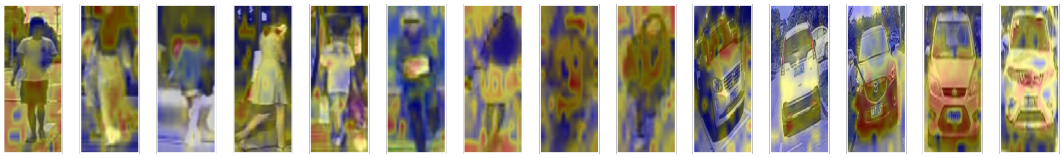
Figure 7: Distance distribution of cross-modal positive, intra-modal negative and cross-modal negative sample pairs in our method on 6 datasets.

(2) Compared with fixed_k, although SG-MDL did not show better results in the early stage of training, the final mAP showed comparable performance. This shows that SG-MDL can reduce the dependence on the predefined number of clusters, while maintaining the retrieval performance across multi-modal.

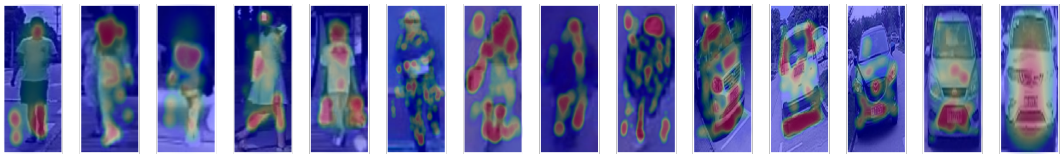
Cross-modal Contrastive Learning To complement the analysis on our method in the main paper, we further visualize the pairwise distance distributions learned by our method on five additional datasets. As shown in Fig 7, a consistent distance structure can be observed across these datasets despite their differences in modality composition, data distribution, and object category.



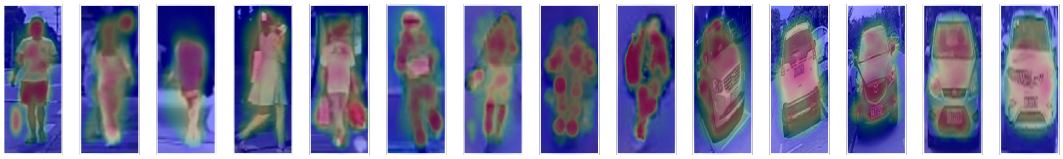
(a) Original



(b) A4



(c) A5



(d) Ours

Figure 8: Additional smoothed attention visualizations for label-aware spatial-frequency learning.

(1) Specifically, cross-modal positive pairs are consistently concentrated in the low-distance region, indicating that representations of the same identity from multi-modalities are effectively aligned. In contrast, intra-modal negative pairs are shifted toward larger distances, showing that the model preserves identity-discrimination within each modality. Cross-modal negative pairs generally occupy an even larger-distance region, reflecting separation of different identities across heterogeneous modalities.

(2) Notably, the degree of overlap between intra- and cross-modal negative pairs distributions varies across datasets, which is expected given their different modality gaps and identity representations. Nevertheless, the relative relationship of positive and negative pairs remains broadly consistent.

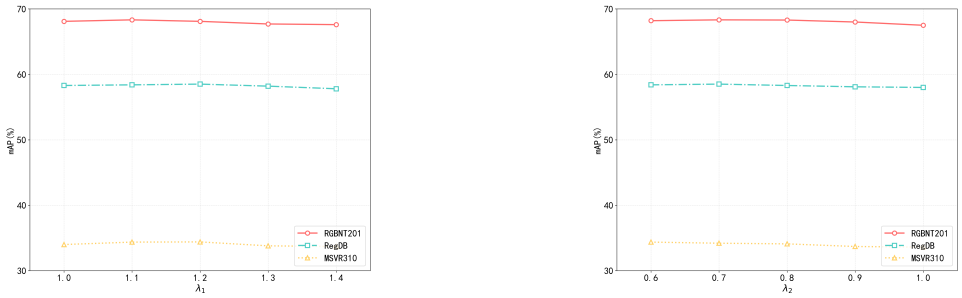
Label-aware SF Learning To complement the visual analysis in the main paper, we provide additional attention visualizations of A4, A5, and ours in Fig 8. The additional examples with diverse appearances and imaging conditions. Overall, they exhibit a consistent tendency with the conclusion in the main paper.

(1) A4 can roughly respond to foreground objects under the guidance of cross-modal learning, it frequently extend to background or identity-irrelevant regions. This phenomenon is consistently observed across different modalities and object categories, suggesting that cross-modal learning provides insufficient spatial selectivity.

(2) introducing label-aware spatial learning in A5 substantially concentrates the attention toward the objects. Nevertheless, its attention is often fragmented into several local regions or dominated by a limited set of salient parts. This performance further illustrates the limitation of relying solely on spatial information: identifying identity-relevant locations does not necessarily guarantee sufficient representation of identity within these regions.

Table 5: Time cost of single-round epoch with four methods on six datasets (DBSCAN, spectral clustering (SC), adaptive SC with fixed_k, ours).

Method	SYSU-MM01	RegDB	HITSZ-VCM	RGBNT100	RGBNT201	MSVR310
DBSCAN	277s	35s	344s	210s	74s	24s
SC	531s	50s	622s	322s	103s	28s
fixed_k adaptive SC	297s	37s	371s	225s	78s	25s
ours	312s	42s	418s	245s	81s	27s



(a) (b)

Figure 9: The impact of the weights of L_{proto} and L_{SF} (corresponding to λ_1 and λ_2) is evaluated.

(3) ours produces more coherent and label-oriented performance across the additional examples. For pedestrian images, the attention covers informative body regions rather than surrounding backgrounds, while for vehicle images they concentrate on characteristic structural regions of the vehicle. Compared with A5, complementary frequency cues enriches the discriminative representation beyond spatial localization alone. Notably, similar behavior is observed across different input, which proves ours is robust.

A.3.2 Time Complexity Analysis

In Table 5, we evaluate the time cost of single-round epoch of four methods on six datasets. The results shows that the time cost of ours is slightly higher than that of DBSCAN and fixed_k adaptive SC, but far less than that of SC. The main additional cost of SG-MDL arises from estimating the partial spectrum of the sparse graph Laplacian. Given N samples with d-dimensional representations, pairwise affinity construction requires $O(N^2d)$, which is shared with the fixed_K adaptive SC variant. Since the learned affinity graph contains only $O(qN)$ nonzero entries, sparse eigensolvers can be employed without performing dense $O(N^3)$ eigendecomposition. Once the required eigenvalues are obtained, evaluating SG-MDL over the candidate set requires only $O(k)$ operations. Consequently, SG-MDL introduces only the cost of estimating additional eigenpairs compared with fixed_k adaptive SC.

A.3.3 Sensitivity Analysis

In Fig 9, we evaluate the impact of the weights of L_{proto} and L_{SF} respectively. Considering the time cost, we only select three representative datasets (RegDB, RGBNT201, MSVR310). Specifically, we enlarge the weight λ_1 of L_{proto} and narrow the weight λ_2 of L_{SF}. Fig 9(a) shows the performance of our method at different values of λ_1, where the fixed value of λ_2 is set to 0.6. The value of λ_1 in Fig 9(b) is fixed at 1. The Experimental result shows that our method achieves better performance while maintaining a relatively low weight. According to our experimental observation, we suggest that λ_1 should be in the range of 1.0-1.2, because it often performs better on different datasets. Similarly, λ_2 is chosen between 0.6 and 0.8. Although the weights will vary according to different datasets, these findings can be used as guidelines for future experiments.

A.4 THE USE OF LLM

We used ChatGPT (OpenAI) to assist with language editing, improving academic clarity, and refining the organization and presentation of the manuscript. All scientific contributions, methodological designs, experimental results, and conclusions were developed and verified by the authors. The authors have reviewed and revised the AI-assisted content and take full responsibility for the final manuscript.

A.5 ANONYMOUS GITHUB CODE

anonymous.4open.science/r/Anonymous-5C9B