Relieving Universal Label Noise for Unsupervised Visible-Infrared Person Re-Identification by Inferring from Neighbors

Xiao Teng[1*], **Long Lan**[1*], **Dingyao Chen**[1], **Kele Xu**[1], **Nan Yin**[2†]

[1]College of Computer Science and Technology, National University of Defense Technology
[2]Hong Kong University of Science and Technology
{tengxiao14,long.lan,chendingyao}@nudt.edu.cn, {kelele.xu,yinnan8911}@gmail.com

Abstract

Unsupervised visible-infrared person re-identification (USL-VI-ReID) is of great research and practical significance yet remains challenging due to the absence of annotations. Existing approaches aim to learn modality-invariant representations in an unsupervised setting. However, these methods often encounter label noise within and across modalities due to suboptimal clustering results and considerable modality discrepancies, which impedes effective training. To address these challenges, we propose a straightforward yet effective solution for USL-VI-ReID by mitigating universal label noise using neighbor information. Specifically, we introduce the Neighbor-guided Universal Label Calibration (N-ULC) module, which replaces explicit hard pseudo labels in both homogeneous and heterogeneous spaces with soft labels derived from neighboring samples to reduce label noise. Additionally, we present the Neighbor-guided Dynamic Weighting (N-DW) module to enhance training stability by minimizing the influence of unreliable samples. Extensive experiments on the RegDB and SYSU-MM01 datasets demonstrate that our method outperforms existing USL-VI-ReID approaches, despite its simplicity.

Code — https://github.com/tengxiao14/Neighbor-guided-USL-VI-ReID

Extended version — https://arxiv.org/abs/2412.12220

Introduction

Visible-infrared person re-identification (VI-ReID) focuses on identifying the same person from the visible or infrared camera when a query image is provided from the other modality (Hao et al. 2021; Ye et al. 2021a). It has attracted widespread attention due to its potential in night vision surveillance applications, where traditional visible cameras cannot work well. Benefiting from accurate manual annotations, recent works have achieved notable improvement on the VI-ReID task. However, the reliance on annotations within and across modalities heavily limits their applications. To address the issue, the task of unsupervised visible-infrared person re-identification (USL-VI-ReID) has emerged and garnered substantial interest (Yang et al. 2022; Yang, Chen, and Ye 2024).

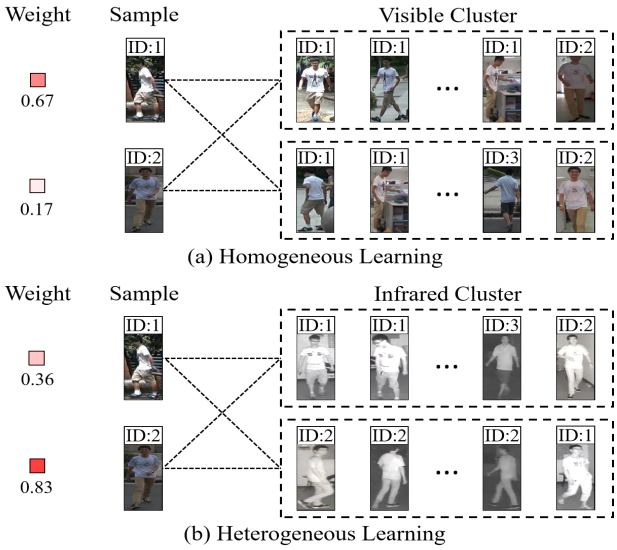


Figure 1: Illustration of the motivation of our method. Due to inferior clustering results, images of the same person can be divided into different groups in each modality. As a result, one-hot pseudo labels are limited to represent their identities in both homogeneous and heterogeneous learning processes. Additionally, the reliability of labeling can vary for diverse samples during the learning process.

The key of USL-VI-ReID is to learn modality-shareable feature representations using only an unlabeled training set (Liang et al. 2021). To achieve this, existing approaches (Yin et al. 2024a; Cheng et al. 2023b,a) typically generate pseudo labels to optimize model parameters. Specifically, these methods first apply clustering algorithms to create pseudo labels within each modality independently. Subsequently, these pseudo labels are used to establish cross-modality correspondences, which aim to create a definitive association between visible and infrared modalities. However, as illustrated in Fig. 1, suboptimal clustering results can lead to images of the same person being assigned to different groups. Consequently, relying solely on one-hot pseudo

*These authors contributed equally.
†Nan Yin is the corresponding author.

labels can introduce significant label noise both within and across modalities, thus impeding the training process.

To address the issue, we propose a simple yet effective framework for USL-VI-ReID by relieving universal label noise with neighbor information. The motivation of our method is that the real identity of a query image can be inferred from its neighbors within and across modalities. Initially, recognizing that images of the same person can be divided into groups and one-hot vectors are limited to represent their identities, we propose the neighbor-guided universal label calibration module to replace the explicit one-hot pseudo labels within and across modalities with more realistic soft pseudo labels derived from the pseudo labels of neighboring images. This approach alleviates universal label noise by providing more accurate representations of true identities. Furthermore, to enhance training stability, we introduce the neighbor-guided dynamic weighting module, which reduces the impact of unreliable samples in both homogeneous and heterogeneous learning processes by utilizing the consistency of neighbors' pseudo labels. By integrating these modules, we significantly reduce label noise at both the pseudo label and sample levels. Our contributions can be summarized as follows:

- We propose a straightforward yet efficient framework for USL-VI-ReID that utilizes neighbor information to alleviate universal label noise within and across modalities at both the pseudo label and sample levels.
- We propose the neighbor-guided label calibration module, which aims to provide more accurate identity representation by employing realistic soft labels for images within and across modalities.
- We further introduce the neighbor-guided dynamic weighting module to diminish the influence of unreliable samples by utilizing the consistency of neighboring pseudo labels in both homogeneous and heterogeneous learning contexts.
- Extensive experiments conducted on two public visible-infrared person ReID benchmarks demonstrate that our method can outperform existing approaches across various settings.

Related Work

Supervised Visible-Infrared Person ReID

Supervised visible-infrared person ReID (SVI-ReID) aims to match the same person across different modalities. The key of SVI-ReID is to obtain discriminative feature representations by leveraging accurate manual annotations (Wu et al. 2017; Kim et al. 2023). Among them, Wu *et al.* (Wu et al. 2017) introduce the task of SVI-ReID and proposes a zero-padding approach to boost the performance of one-stream model. To regularize the model from overfitting, the PartMix augmentation technique (Kim et al. 2023) is introduced by mixing local descriptions of images from different modalities for part-level data augmentation. To enhance the generalization of graph-based models, Li *et al.* (Li et al. 2022) rethink the SVI-ReID problem from the perspective of counterfactual intervention and propose a novel framework by transferring features and stressing the effect of topology structure. Regarding that convolution neural network is limited in extracting discriminative feature representations, Zhao *et al.* introduce the transformer into the task of SVI-ReID and propose a novel framework by enhancing spatial and channel information.

While these methodologies have demonstrated substantial efficacy in learning modality-shareable feature representations, they all require expensive annotations that are often challenging to acquire, thereby constraining their real-world applicability.

Unsupervised Visible-Infrared Person ReID

Unsupervised visible-infrared person re-identification (USL-VI-ReID) seeks to learn modality-shareable feature representations without requiring annotations (Yang et al. 2022; Wu and Ye 2023; Pang et al. 2024). This task is inherently more challenging than unsupervised single-modality person ReID, as the discrepancies between modalities often surpass the intra-class variance within a single modality (Teng et al. 2023; Lan et al. 2023; Zhao et al. 2022a). To effectively utilize unlabeled training data from diverse modalities, current approaches typically establish cross-modality correspondences through clustering within each modality (Ju et al. 2024; Wu and Ye 2023; Pang et al. 2023). For instance, Yang *et al.* (Yang et al. 2022) introduce a novel contrastive learning framework that aggregates cross-modality memories based on priority counts. Wu *et al.* (Wu and Ye 2023) frame cross-modality correspondence mining as a graph matching problem and propose a method featuring progressive graph matching and alternative cross-modality learning modules.

Despite these advancements, the methods still face challenges due to suboptimal clustering results and significant modality discrepancies, which can introduce label noise both within and across modalities, thereby limiting model performance.

Neighbor-Related Person ReID

Since a single image contains limited information, recent studies have sought to extract more valuable information from neighboring images in both supervised and unsupervised settings (Wang et al. 2022a; Zhong et al. 2017; Yu et al. 2021). For instance, Wang *et al.* (Wang et al. 2022a) introduce explicit relationships between input images to mitigate outlier effects and enhance the robustness of learned feature representations. Zhong *et al.* (Zhong et al. 2017) utilize neighbor information to refine retrieval results derived from direct similarity calculations of feature vectors. To address occlusion issues in person ReID, Yu *et al.* (Yu et al. 2021) propose a neighbor-guided feature reconstruction method to recover missing information using reliable neighbor data

Similar to our approach, several studies also incorporate neighbor information in USL-VI-ReID tasks as a submodule to refine cross-modality correspondences or to extract relationships among samples (Cheng et al. 2023b; Pang et al. 2023; He et al. 2024; Yin et al. 2024b; Yang et al. 2023a). However, our method is distinct in that it is exclusively based on neighbor information and employs it

in a more comprehensive and universal manner. This approach addresses pseudo label noise in both homogeneous and heterogeneous spaces through pseudo label refinement and sample weighting. As a result, our method demonstrates superior efficiency compared to existing methods, despite its inherent simplicity.

Method

The framework of our proposed method is depicted in Fig. 2. To provide a comprehensive understanding, we first revisit the Progressive Graph Matching (PGM) framework (Wu and Ye 2023), which serves as the baseline for our approach. Building upon PGM, we will then introduce two key components of our method: the Neighbor-guided Universal Label Calibration (N-ULC) module and the Neighbor-guided Dynamic Weighting (N-DW) module, which will be detailed in the subsequent sections.

Progressive Graph Matching

Given the unlabeled training sets for visible and infrared modalities, denoted as $X = \{X_v, X_i\}$, where $X_v = \{x_1^v, x_2^v, \ldots, x_N^v\}$ represents the visible training set with N images, while $X_i = \left\{x_1^i, x_2^i, \ldots, x_M^i\right\}$ denotes the infrared training set containing M images. To extract features from the unlabeled training sets, the two-stream encoders f_θ^v and f_θ^i are utilized, which share the same convolution backbone but with modality-specific classifiers. This process yields modality-specific feature sets $U_v = \{u_1^v, u_2^v, \ldots, u_N^v\}$ and $U_i = \left\{u_1^i, u_2^i, \ldots, u_M^i\right\}$ for visible and infrared modalities, respectively. To generate pseudo labels for these feature sets, we apply the DBSCAN clustering algorithm (Ester et al. 1996) independently on U_v and U_i. This results in cluster sets $\mathcal{H}_v = \{C_1^v, C_2^v, \ldots, C_K^v\}$ and $\mathcal{H}_i = \left\{C_1^i, C_2^i, \ldots, C_L^i\right\}$, where K and L denote the number of clusters in the visible and infrared modalities, respectively. Then the modalities-specific prototypes $\Phi_v \in \mathbb{R}^{K\times d}$ and $\Phi_i \in \mathbb{R}^{L\times d}$ can be obtained by applying averaging operation on feature vectors within each cluster, where d denotes the dimension of the feature vector.

To establish associations between different modalities, the graph matching algorithm can be formulated as follows:

$$\begin{aligned} G(M) &= \arg\min_M D^T M \\ s.t. \forall i \in [K], j \in [L] &: M_{ij} \in \{0,1\}, \\ \forall i \in [K] &: \sum_{j\in[L]} M_{ij} \le 1, \\ \forall j \in [L] &: \sum_{i\in[K]} M_{ij} \le 1, \end{aligned} \tag{1}$$

where $D \in \mathbb{R}^{K\times L}$ denotes the cost matrix representing the dissimilarity measure between the visible prototype $\Phi_v \in \mathbb{R}^{K\times d}$ and infrared prototype $\Phi_i \in \mathbb{R}^{L\times d}$. Using the progressive matching strategy detailed in (Wu and Ye 2023), it is possible to associate all clusters between the visible and infrared modalities. Due to space constraints, additional details on the matching process are omitted here. Consequently, the cross-modality label transformers $T^{V\to I}$ and $T^{I\to V}$ are derived, facilitating the transformation of pseudo labels between modalities. During training, for a sample q_v from visible modality, assumed to belong to the k'-th cluster, homogeneous learning can be expressed as follows:

$$L^{intra}(q_v) = -\log \frac{\exp\left(q_v \cdot \phi_{k'}^v/\tau\right)}{\sum_{k=1}^K \exp\left(q_v \cdot \phi_k^v/\tau\right)}, \tag{2}$$

where ϕ_k^i denotes the prototype of the k-th cluster from visible modality, while ϕ_+^v denotes the prototype of the cluster to which q_v belongs. τ is the temperature hyper-parameter. Prototypes are updated using a momentum scheme based on their corresponding query features (Wu and Ye 2023). Similarly, heterogeneous learning can be formulated as follows:

$$L^{inter}(q_v) = -\log \frac{\exp\left(q_v \cdot \phi_{T^{V\to I}[k']}^i/\tau\right)}{\sum_{l=1}^L \exp\left(q_v \cdot \phi_l^i/\tau\right)}, \tag{3}$$

where ϕ_l^i denotes the l-th cluster in the infrared modality. To mitigate the modality gap, Random Channel Augmentation (CA) (Ye et al. 2021a) is employed during training. For samples from the infrared modality, comparisons are made with both infrared and visible prototypes in a similar manner. The objective function is then formulated as follows:

$$\begin{aligned} L_{overall} = &\frac{1}{|B_V|}\sum_{q_v\in B_V}\left(L^{intra}(q_v) + \lambda L^{inter}(q_v)\right) \\ &+\frac{1}{|B_I|}\sum_{q_i\in B_I}\left(L^{intra}(q_i) + \lambda L^{inter}(q_i)\right), \end{aligned} \tag{4}$$

where λ is the hyper-parameter that balances homogeneous learning and heterogeneous learning.

Neighbor-Guided Universal Label Calibration

Due to suboptimal clustering results and significant modality discrepancies, learning with a progressive graph matching framework can be adversely affected by substantial label noise both within and across modalities. Since images of the same individual may be assigned to different clusters during the clustering process, relying solely on hard pseudo labels to establish connections between images and a single cluster, either within or across modalities, proves inadequate. To address this challenge, we propose the N-ULC module. This module aims to enhance the accuracy of sample identity representation by deriving soft labels from information provided by neighboring samples.

Specifically, given the training sets from different modalities, we obtain feature sets $\{U_v, U_i\}$ and cluster sets $\{\mathcal{H}_v, \mathcal{H}_i\}$ as described in the previous section. For each sample q_v from the visible training set, which is assumed to belong to the l'-th cluster, we identify its k-nearest neighbors from the visible training set U_v. The resulting list of neighbors is denoted as $N(q_v, U_v, k)$. Intuitively, the identity statistics of this ranking list can provide insights into the true identity of q_v. Consequently, the correlation between q_v and the visible clusters can be expressed as:

$$[\tilde{P}_{q_v}^{intra}]_l = \frac{|N(q_v, U_v, k) \cap C_l^v|}{|N(q_v, U_v, k) \cup C_l^v|}, \tag{5}$$

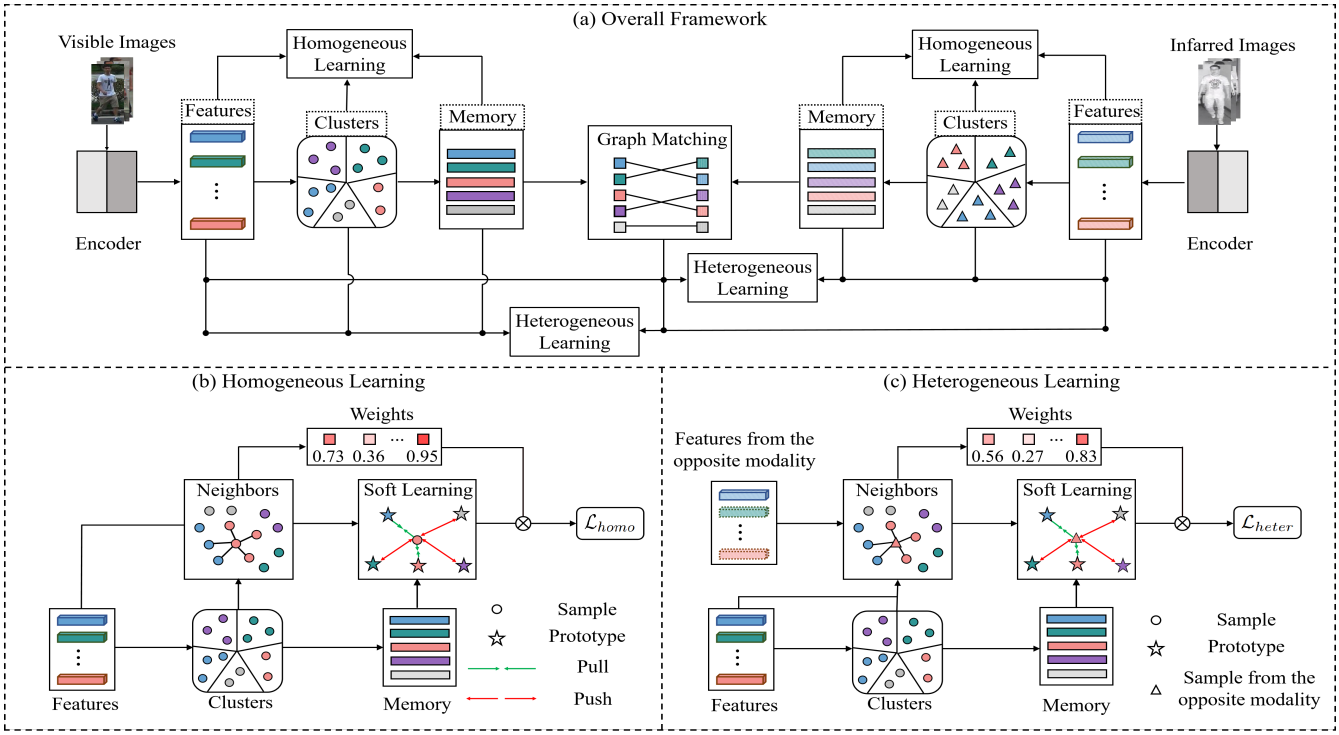


Figure 2: **Framework of the proposed method.** Based on the Progressive Graph Matching (PGM) framework, we propose the neighbor-guided universal label calibration module (Sec.) and neighbor-guided dynamic weighting module (Sec.), these module are applied on both the (b) homogeneous learning and (c) heterogeneous learning processes.

where $\tilde{P}_{q_v}^{intra}$ represents the correlation between q_v and the visible clusters, and $[\tilde{P}_{q_v}^{intra}]_l$ denotes the l-th entry of $\tilde{P}_{q_v}^{intra}$. Since the query sample q_v belongs to the l'-th cluster in the visible modality, its original one-hot pseudo label is denoted as $\mathcal{I}_{q_v}^{intra}$, where the l'-th entry is set to 1 and all other entries are set to 0. The calibrated soft label for q_v can then be expressed as:

$$\tilde{\mathcal{I}}_{q_v}^{intra} = \mu \mathcal{I}_{q_v}^{intra} + (1-\mu) P_{q_v}^{intra}, \tag{6}$$

where $P_{q_v}^{intra}$ denotes the ℓ_1-norm of $\tilde{P}_{q_v}^{intra}$ and μ is the hyper-parameter. Drawing inspiration from the label smoothing technique, the calibrated homogeneous learning process for the query sample q_v can be expressed as follows:

$$L_{soft}^{intra}(q_v) = -\sum_{k=1}^{K} [\tilde{\mathcal{I}}_{q_v}^{intra}]_k \log \frac{\exp\left(q_v \cdot \phi_k^v / \tau\right)}{\sum_{k'=1}^{K} \exp\left(q_v \cdot \phi_{k'}^v / \tau\right)}. \tag{7}$$

Similarly, label noise across modalities can also be mitigated using this approach. For a given query sample q_v, its cross-modality k-nearest neighbors can be identified as $N(q_v, U_i, k)$. Subsequently, the correlation between q_v and clusters from the infrared modality can be represented as:

$$[\tilde{P}_{q_v}^{inter}]_l = \frac{|N(q_v, U_i, k) \cap C_l^i|}{|N(q_v, U_i, k) \cup C_l^i|}, \tag{8}$$

where U_i denotes the feature set extracted from the infrared training data. C_l^i represents the cluster set containing samples from the l-th cluster in the infrared modality. Since q_v belongs to the l'-th cluster in the visible modality, its corresponding cluster in the infrared modality can be identified using the cross-modality label transformer $T^{V\to I}$, denoted as $T^{V\to I}[l']$. Consequently, the cross-modality one-hot pseudo label for q_v is represented as $\mathcal{I}_{q_v}^{inter}$, where the $T^{V\to I}[l']$-th entry is set to 1 and all other entries are set to 0. Thus, the calibrated cross-modality soft label for q_v is expressed as:

$$\tilde{\mathcal{I}}_{q_v}^{inter} = \mu \mathcal{I}_{q_v}^{inter} + (1-\mu) P_{q_v}^{inter}, \tag{9}$$

where $P_{q_v}^{inter}$ is the ℓ_1-norm of $\tilde{P}_{q_v}^{inter}$. The calibrated heterogeneous learning process for q_v can then be described as:

$$L_{soft}^{inter}(q_v) = -\sum_{k=1}^{L} [\tilde{\mathcal{I}}_{q_v}^{inter}]_k \log \frac{\exp\left(q_v \cdot \phi_k^i / \tau\right)}{\sum_{k'=1}^{L} \exp\left(q_v \cdot \phi_{k'}^i / \tau\right)}. \tag{10}$$

The same procedure can be applied to samples from the infrared modality for label calibration both within and across modalities. For simplicity, the details of this process are omitted.

Neighbor-Guided Dynamic Weighting

Our proposed neighbor-guided universal label calibration alleviates label noise by more accurately reflecting the true identity of images both within and across modalities (Yin et al. 2023). To enhance the stability of the learning process, we introduce the N-DW module, which reduces the impact of unreliable samples during training. Specifically, for

a query sample q_v from the l'-th cluster in visible modality, its normalized correlations with clusters within and across modalities can be represented as $P_{q_v}^{intra}$ and $P_{q_v}^{inter}$, respectively. Intuitively, samples that are more consistent with their neighbors are considered to have more reliable pseudo labels. Consequently, neighbor information is utilized to generate weights for samples, thereby minimizing the influence of unreliable samples. For homogeneous learning, the weight of q_v is computed as follows:

$$\omega_{q_v}^{intra} = \exp(-w \cdot (1 - [P_{q_v}^{intra}]_{l'})^2), \tag{11}$$

where w is the hyper-parameter controlling the degree of penalization for unreliable samples, empirically set to 10 in our experiments. The term $[P_{q_v}^{intra}]_{l'}$ is the l'-th entry of $P_{q_v}^{intra}$, with l' representing the pseudo label of q_v in the visible modality. Consequently, $\omega_{q_v}^{intra}$ increases monotonically with $[P_{q_v}^{intra}]_{l'}$. Similarly, for heterogeneous learning, the weight of q_v can be expressed in an analogous manner as:

$$\omega_{q_v}^{inter} = \exp(-w \cdot (1 - [P_{q_v}^{inter}]_{T^{V\to I}[l']})^2), \tag{12}$$

where $T^{V\to I}[l']$ denotes the matched cluster in the infrared modality for q_v. For samples originating from the infrared modality, the weights in both homogeneous and heterogeneous learning processes can be determined in a similar manner. Ultimately, the overall objective function of our method is expressed as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{homo} + \lambda \mathcal{L}_{heter} \\ &= \frac{1}{|B_V|} \sum_{q_v \in B_V} (\omega_{q_v}^{intra} L_{soft}^{intra}(q_v) + \lambda \omega_{q_v}^{inter} L_{soft}^{inter}(q_v)) \\ &+ \frac{1}{|B_I|} \sum_{q_i \in B_I} (\omega_{q_i}^{intra} L_{soft}^{intra}(q_i) + \lambda \omega_{q_i}^{inter} L_{soft}^{inter}(q_i)),\end{aligned} \tag{13}$$

where B_I and B_V represent the input batches from the infrared and visible modalities, respectively. The parameter λ is a hyper-parameter. In the objective function, the first term corresponds to the homogeneous learning process, while the second term pertains to the heterogeneous learning process.

Discussion. In this section, we propose a method for generating weights for samples to mitigate the influence of unreliable samples during both the homogeneous and heterogeneous learning stages. Initially, we assign higher weights to easily labeled samples to aid the model in learning general patterns. As training advances, more challenging samples are assigned greater weights if they demonstrate increased consistency with their neighbors. This approach can be considered a form of curriculum learning.

Theoretical Analysis

In this study, we derive soft labels from neighbor information to optimize model parameters both in the homogeneous learning and heterogeneous learning processes. In this part, we aim to investigate the Rademacher generalization bound of our method. For simplicity, we merely consider the binary classification setting, i.e., each sample x is paired with a label $y \in \{0, 1\}$. Our method relies on soft labels derived from a probabilistic distribution, leading to the following loss function formulation: $\ell^{\mathrm{SOFT}}(h(x), y) = (1-\beta) \cdot \ell(h(x), y) + \beta \cdot \ell(h(x), 1-y)$, where ℓ represents a commonly used loss function such as cross-entropy loss and β denotes the probability of the irrelevant class. Our method aims to obtain the classifier h by minimizing the following expected risk:

$$\mathcal{R}(h) := \mathbb{E}_{(x,y)\sim\mathbb{D}} \left[\ell^{\mathrm{SOFT}}\left(h\left(x\right), y\right)\right], \tag{14}$$

where $\mathbb{D}$ is the distribution of the dataset and y is the corresponding generated pseudo label. ℓ is the loss function. Due to the limited scale of our dataset, our actual learning objective becomes the empirical risk as follows:

$$\hat{\mathcal{R}}(h) := \frac{1}{N} \sum\nolimits_{i=1}^{N} \left[\ell^{\mathrm{SOFT}}\left(h\left(x_i\right), y_i\right)\right], \tag{15}$$

where N is the size of the training set.

Theorem 1(Wei et al. 2022). With probability at least $1-\delta$, for all $h \in \mathcal{H}$, we have:

$$\mathcal{R}(h) \leq \hat{\mathcal{R}}(h) + 2 \cdot L \cdot \Re(\mathcal{H}) + (1-2\beta)(\overline{\ell} - \underline{\ell}) \cdot \sqrt{\frac{log(1/\delta)}{2N}}, \tag{16}$$

where $\overline{\ell}$ and $\underline{\ell}$ are the upper bound and lower bound of ℓ and $\Re$ is the Rademacher complexity.

Remark 1. Theorem 1 offers an upper bound for the gap between expected risk $\mathcal{R}(h)$ and empirical risk $\hat{\mathcal{R}}(h)$. Since β is usually set to a small value to ensure the pivot can be classified, it can be regarded as a constant here. When the size of the dataset N is large enough and the Rademacher complexity of the hypothesis space $\Re(\mathcal{H})$ is limited, $\mathcal{R}(h)$ can approach $\hat{\mathcal{R}}(h)$ well with soft labels derived from probabilistic distributions.

Experiment

Datasets and Evaluation Protocols

Our experiments are conducted on two public datasets: RegDB (Nguyen et al. 2017) and SYSU-MM01 (Wu et al. 2017). To ensure fair comparison, we use mean average precision (mAP), Cumulative Matching Characteristics (CMC), and mean Inverse Negative Penalty (mINP) (Ye et al. 2021b) as evaluation metrics, which are commonly employed in existing research (Yang et al. 2022; Wu and Ye 2023).

SYSU-MM01 is a large-scale VI-ReID dataset collected from 4 visible and 2 infrared cameras. The training set includes 395 identities with 22,258 visible images and 11,909 infrared images, while the testing set includes 96 identities.

RegDB is another dataset acquired from one visible and one infrared camera. It contains 412 identities, each identity includes 10 visible and 10 infrared images. In line with previous studies (Yang et al. 2022; Wu and Ye 2023), we perform 10 experiments on this dataset and report the average performance as the final results.

Implementation Details

Our method is implemented in the PyTorch platform. Following existing works (Yang et al. 2022; Cheng et al. 2023a;

			SYSU-MM01 dataset						RegDB dataset					
			Indoor Search			All search			Infrared to Visible			Visible to Infrared		
	Method	Reference	r1	mAP	mINP	r1	mAP	mINP	r1	mAP	mINP	r1	mAP	mINP
SVI-ReID	AGW	TPAMI-21	54.17	62.97	59.23	47.50	47.65	35.30	70.49	65.90	51.24	70.05	66.37	50.19
	CA	ICCV-21	76.26	80.37	76.79	69.88	66.89	53.61	84.75	77.82	61.56	85.03	79.14	65.33
	MAUM	CVPR-22	76.97	81.94	-	71.68	68.79	-	86.95	84.34	-	87.87	85.09	-
	DFLN-ViT	TMM-22	62.13	69.03	-	59.84	57.70	-	91.21	81.62	-	92.10	82.11	-
	PartMix	CVPR-23	81.52	84.38	-	77.78	74.62	-	84.93	82.52	-	85.66	82.27	-
	MUN	ICCV-23	79.42	82.06	-	76.24	73.81	-	91.86	85.01	-	95.19	87.15	-
	SAAI	ICCV-23	83.20	88.01	-	75.90	77.03	-	92.09	92.01	-	91.07	91.45	-
USL-ReID	SPCL	NeurIPS-20	26.83	36.42	33.05	18.37	19.39	10.99	11.70	13.56	10.09	13.59	14.86	10.36
	MMT	ICLR-20	24.42	25.59	18.66	25.68	26.51	19.56	22.79	31.50	27.66	21.47	21.53	11.50
	ICE	ICCV-21	12.18	14.82	10.60	12.98	15.64	11.91	29.81	38.35	34.32	20.54	20.39	10.24
	CCL	ACCV-22	11.14	12.99	8.99	11.76	13.88	9.94	23.33	34.01	30.88	20.16	22.00	12.97
	PPLR	CVPR-22	8.11	9.07	5.65	8.93	11.14	7.89	12.71	20.81	17.61	11.98	12.25	4.97
	ISE	CVPR-22	10.83	13.66	10.71	16.12	16.99	13.24	14.22	24.62	21.74	20.01	18.93	8.54
USL-VI-ReID	H2H*	TIP-21	-	-	-	30.15	29.40	-	-	-	-	23.81	18.87	-
	OTLA*	ECCV-22	47.4	56.8	-	48.2	43.9	-	49.6	42.8	-	49.9	41.8	-
	ADCA	MM-22	50.60	59.11	55.17	45.51	42.73	28.29	68.48	63.81	49.62	67.20	64.05	52.67
	PGM	CVPR-23	56.23	62.74	58.13	57.27	51.78	34.96	69.85	65.17	-	69.48	65.41	-
	DOTLA*	MM-23	53.47	61.73	57.35	50.36	47.36	32.40	82.91	74.97	58.60	85.63	76.71	61.58
	MBCCM	MM-23	55.21	61.98	57.13	53.14	48.16	32.41	82.82	76.74	61.73	83.79	77.87	65.04
	CCLNet	MM-23	56.68	65.12	-	54.03	50.19	-	70.17	66.66	-	69.94	65.53	-
	GUR†	ICCV-23	64.22	69.49	64.81	60.95	56.99	41.85	75.00	69.94	56.21	73.91	70.23	58.88
	SCA-RCP	TKDE-24	56.77	64.19	59.25	51.41	48.52	33.56	82.41	75.73	-	85.59	79.12	-
	Ours	-	**67.04**	**73.08**	**69.42**	**61.81**	**58.92**	**45.01**	**88.17**	**81.11**	**66.05**	**88.75**	**82.14**	**68.75**

Table 1: Experimental results (%) of our method and SOTA methods on the SYSU-MM01 and RegDB datasets under different settings. $*$ means the model is pre-trained on an extra labeled visible dataset. GUR† deontes results without camera information.

Yang, Chen, and Ye 2023; Li et al. 2024a), we utilize the ImageNet-pretrained ResNet50 (He et al. 2016) as the backbone. In each mini-batch, we select 16 identities per modality, with each identity comprising 16 instances. Input images are resized to 288 $\times$ 144 pixels. Standard data augmentation techniques, including random flipping, random cropping, and random erasing, are applied. Pseudo labels are generated using the DBSCAN clustering algorithm, with a maximum distance of 0.2 for the RegDB dataset and 0.6 for the SYSU-MM01 dataset. Hyper-parameters μ and λ are set to 0.7 and 3, respectively, while the number of neighbors k is set to 20 and 30 for the RegDB and SYSU-MM01 datasets, respectively. All other experimental settings follow those of previous works (Yang et al. 2022; Wu and Ye 2023).

Comparison with SOTA Methods

To assess the effectiveness of our method, we compare it against state-of-the-art approaches across three ReID settings: supervised visible-infrared person ReID (SVI-ReID), unsupervised single-modality person ReID (USL-ReID), and unsupervised visible-infrared person ReID (USL-VI-ReID). The results are presented in Table 1.

Comparison with SVI-ReID methods. We compare our method with several recent SVI-ReID approaches, including AGW (Ye et al. 2021b), CA (Ye et al. 2021a), MAUM (Liu et al. 2022), DFLN-ViT (Zhao et al. 2022b), PartMix (Kim et al. 2023), MUN (Yu et al. 2023), and SAAI (Fang, Yang, and Fu 2023). Despite these methods benefiting from precise manual annotations, our method achieves comparable performance to some of them, such as AGW and DFLN-ViT.

Comparison with USL-ReID methods. We compare our method with recent state-of-the-art USL-ReID techniques, including SPCL (Ge et al. 2020), MMT (Ge, Chen, and Li 2020), CCL (Dai et al. 2022), ICE (Chen, Lagadec, and Bremond 2021), PPLR (Cho et al. 2022), and ISE (Zhang et al. 2022). These approaches are generally constrained by their focus on USL-ReID, which limits their ability to address severe modality discrepancies.

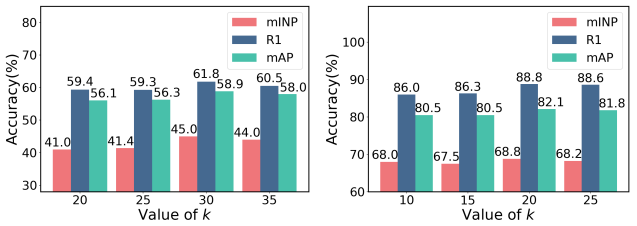
Comparison with USL-VI-ReID methods. We also evaluate our method against advanced USL-VI-ReID approaches, such as H2H (Liang et al. 2021), OTLA (Wang et al. 2022b), ADCA (Yang et al. 2022), TAA (Yang et al. 2023b), PGM (Wu and Ye 2023), CCLNet (Chen et al. 2023), MBCCM (Cheng et al. 2023a), DOTLA (Cheng et al. 2023b), GUR (Yang, Chen, and Ye 2023), and SCA-RCP (Li et al. 2024b). Our method outperforms all of these approaches on the tested datasets, likely because it effectively addresses universal label noise in homogeneous and heterogeneous learning contexts at both the pseudo label and sample levels, thus demonstrating superior efficiency.

Index	Components			SYSU-MM01 Settings						RegDB Settings					
				Indoor Search			All search			Infrared-to-Visible			Visible-to-Infrared		
	Baseline	N-ULC	N-DW	r1	mAP	mINP	r1	mAP	mINP	r1	mAP	mINP	r1	mAP	mINP
1	✓			58.80	65.35	60.58	54.14	50.37	33.56	77.44	66.47	45.40	77.92	67.51	48.43
2	✓	✓		61.14	67.15	62.69	55.87	51.79	35.71	79.29	70.45	52.19	80.60	71.72	54.64
3	✓		✓	62.04	69.06	65.16	58.54	55.92	41.79	87.82	**81.80**	**67.66**	87.88	**82.24**	**70.13**
4	✓	✓	✓	**67.04**	**73.08**	**69.42**	**61.81**	**58.92**	**45.01**	**88.17**	81.11	66.05	**88.75**	82.14	68.75

Table 2: Ablation study on the SYSU-MM01 and RegDB datasets (%).

Ablation Study

In Table 2, we present experiments conducted on the SYSU-MM01 and RegDB datasets to evaluate the components of our method, specifically the Neighbor-Guided Label Universal Calibration (N-ULC) and Neighbor-Guided Dynamic Weighting (N-DW) modules. The results indicate that applying the N-ULC and N-DW modules individually yields substantial improvements on both datasets, highlighting the effectiveness of these components. When both modules are combined, our method achieves optimal performance on the SYSU-MM01 dataset. However, improvements on the RegDB dataset are more limited. This limitation is likely due to the already high performance of the N-DW module on RegDB and the partial overlap in functionality between the modules, which restricts further gains in performance.



(a) Impact of k on SYSU-MM01 under all search setting. (b) Impact of k on RegDB under visible-to-infrared setting.

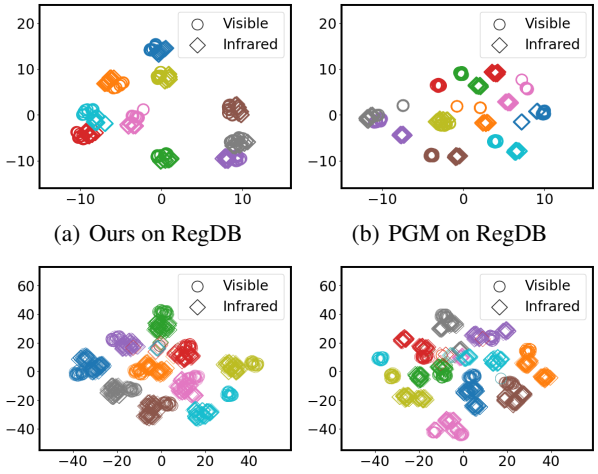
Figure 3: Impact of hyper-parameter k on different datasets.

Hyper-parameter Analysis

Our method relies on information from nearest neighbors, making the nearest neighbor number k in Eq. (5) a crucial factor in its performance. Figure 3 presents experiments conducted on various datasets to assess the impact of this hyper-parameter. The results indicate that the performance of our method remains relatively stable across different values of k. For optimal results, k is set to 30 for the SYSU-MM01 dataset and 20 for the RegDB dataset.

Visualization

To further evaluate the effectiveness of our method in learning modality-shareable feature representations, we employ t-SNE (Van der Maaten and Hinton 2008) to visualize the features learned by both our method and PGM on the SYSU-MM01 and RegDB datasets, as shown in Fig. 4. The results



(a) Ours on RegDB (b) PGM on RegDB

(c) Ours on SYSU-MM01 (d) PGM on SYSU-MM01

Figure 4: T-SNE visualization of features learned by PGM and our method on a subset of RegDB and SYSU-MM01 datasets. Different colors represent different identities.

reveal that PGM often separates images of the same person across different modalities into distinct clusters, whereas our method generates more compact feature representations for the same individual. In contrast, PGM tends to confuse more identities and blend features from different individuals compared with our method.

Conclusion

In this paper, we present a simple yet effective framework for USL-VI-ReID that addresses universal label noise through the use of neighbor information. To mitigate label noise, we introduce the neighbor-guided universal label calibration module, which refines explicit hard pseudo labels by leveraging information from neighbors in both homogeneous and heterogeneous spaces. Additionally, to enhance training stability, we propose the neighbor-guided dynamic weighting module, which reduces the impact of unreliable samples within and across modalities. Extensive experiments conducted on the SYSU-MM01 and RegDB datasets demonstrate the effectiveness of our proposed method, despite its inherent simplicity.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62376282)

References

Chen, H.; Lagadec, B.; and Bremond, F. 2021. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14960–14969.

Chen, Z.; Zhang, Z.; Tan, X.; Qu, Y.; and Xie, Y. 2023. Unveiling the power of clip in unsupervised visible-infrared person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3667–3675.

Cheng, D.; He, L.; Wang, N.; Zhang, S.; Wang, Z.; and Gao, X. 2023a. Efficient Bilateral Cross-Modality Cluster Matching for Unsupervised Visible-Infrared Person ReID. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1325–1333.

Cheng, D.; Huang, X.; Wang, N.; He, L.; Li, Z.; and Gao, X. 2023b. Unsupervised visible-infrared person reid by collaborative learning with neighbor-guided label refinement. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7085–7093.

Cho, Y.; Kim, W. J.; Hong, S.; and Yoon, S.-E. 2022. Part-based pseudo label refinement for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7308–7318.

Dai, Z.; Wang, G.; Yuan, W.; Zhu, S.; and Tan, P. 2022. Cluster contrast for unsupervised person re-identification. In *Proceedings of the Asian Conference on Computer Vision*, 1142–1160.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 226–231.

Fang, X.; Yang, Y.; and Fu, Y. 2023. Visible-infrared person re-identification via semantic alignment and affinity inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11270–11279.

Ge, Y.; Chen, D.; and Li, H. 2020. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*.

Ge, Y.; Zhu, F.; Chen, D.; Zhao, R.; et al. 2020. Self-paced contrastive learning with hybrid memory for domain adaptive object re-id. *Advances in Neural Information Processing Systems*, 33: 11309–11321.

Hao, X.; Zhao, S.; Ye, M.; and Shen, J. 2021. Cross-modality person re-identification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International conference on computer vision*, 16403–16412.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, L.; Cheng, D.; Wang, N.; and Gao, X. 2024. Exploring Homogeneous and Heterogeneous Consistent Label Associations for Unsupervised Visible-Infrared Person ReID. *arXiv preprint arXiv:2402.00672*.

Ju, W.; Yi, S.; Wang, Y.; Xiao, Z.; Mao, Z.; Li, H.; Gu, Y.; Qin, Y.; Yin, N.; Wang, S.; et al. 2024. A survey of graph neural networks in real world: Imbalance, noise, privacy and ood challenges. *arXiv preprint arXiv:2403.04468*.

Kim, M.; Kim, S.; Park, J.; Park, S.; and Sohn, K. 2023. PartMix: Regularization Strategy To Learn Part Discovery for Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18621–18632.

Lan, L.; Teng, X.; Zhang, J.; Zhang, X.; and Tao, D. 2023. Learning to Purification for Unsupervised Person Re-identification. *IEEE Transactions on Image Processing*.

Li, C.; Teng, X.; Ding, Y.; and Lan, L. 2024a. Instance-Level Scaling and Dynamic Margin-Alignment Knowledge Distillation for Remote Sensing Image Scene Classification. *Remote Sensing*, 16(20): 3853.

Li, X.; Lu, Y.; Liu, B.; Liu, Y.; Yin, G.; Chu, Q.; Huang, J.; Zhu, F.; Zhao, R.; and Yu, N. 2022. Counterfactual Intervention Feature Transfer for Visible-Infrared Person Re-identification. In *European Conference on Computer Vision*, 381–398. Springer.

Li, Z.; Liu, H.; Peng, X.; and Jiang, W. 2024b. Inter-Intra Modality Knowledge Learning and Clustering Noise Alleviation for Unsupervised Visible-Infrared Person Re-Identification. *IEEE Transactions on Knowledge and Data Engineering*.

Liang, W.; Wang, G.; Lai, J.; and Xie, X. 2021. Homogeneous-to-heterogeneous: Unsupervised learning for rgb-infrared person re-identification. *IEEE Transactions on Image Processing*, 30: 6392–6407.

Liu, J.; Sun, Y.; Zhu, F.; Pei, H.; Yang, Y.; and Li, W. 2022. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19366–19375.

Nguyen, D. T.; Hong, H. G.; Kim, K. W.; and Park, K. R. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 17(3): 605.

Pang, Z.; Wang, C.; Pan, H.; Zhao, L.; Wang, J.; and Guo, M. 2024. MIMR: Modality-Invariance Modeling and Refinement for unsupervised visible-infrared person re-identification. *Knowledge-Based Systems*, 285: 111350.

Pang, Z.; Wang, C.; Zhao, L.; Liu, Y.; and Sharma, G. 2023. Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

Teng, X.; Lan, L.; Zhao, J.; Li, X.; and Tang, Y. 2023. Highly Efficient Active Learning With Tracklet-Aware Co-Cooperative Annotators for Person Re-Identification. *IEEE Transactions on Neural Networks and Learning Systems*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, H.; Shen, J.; Liu, Y.; Gao, Y.; and Gavves, E. 2022a. Nformer: Robust person re-identification with neighbor transformer. In *Proceedings of the IEEE/CVF confer-*

ence on computer vision and pattern recognition, 7297–7307.

Wang, J.; Zhang, Z.; Chen, M.; Zhang, Y.; Wang, C.; Sheng, B.; Qu, Y.; and Xie, Y. 2022b. Optimal transport for label-efficient visible-infrared person re-identification. In *European Conference on Computer Vision*, 93–109. Springer.

Wei, J.; Liu, H.; Liu, T.; Niu, G.; Sugiyama, M.; and Liu, Y. 2022. To Smooth or Not? When Label Smoothing Meets Noisy Labels. In *International Conference on Machine Learning*, 23589–23614. PMLR.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. RGB-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, 5380–5389.

Wu, Z.; and Ye, M. 2023. 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*, 9548–9558.

Yang, B.; Chen, J.; Chen, C.; and Ye, M. 2023a. Dual Consistency-Constrained Learning for Unsupervised Visible-Infrared Person Re-Identification. *IEEE Transactions on Information Forensics and Security*.

Yang, B.; Chen, J.; Ma, X.; and Ye, M. 2023b. Translation, association and augmentation: Learning cross-modality re-identification from single-modality annotation. *IEEE Transactions on Image Processing*.

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

Yang, B.; Chen, J.; and Ye, M. 2024. Shallow-Deep Collaborative Learning for Unsupervised Visible-Infrared Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16870–16879.

Yang, B.; Ye, M.; Chen, J.; and Wu, Z. 2022. Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2843–2851.

Ye, M.; Ruan, W.; Du, B.; and Shou, M. Z. 2021a. Channel augmented joint learning for visible-infrared recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13567–13576.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021b. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 2872–2893.

Yin, N.; Shen, L.; Chen, C.; Hua, X.-S.; and Luo, X. 2024a. Sport: A subgraph perspective on graph classification with label noise. *ACM Transactions on Knowledge Discovery from Data*, 18(9): 1–20.

Yin, N.; Shen, L.; Wang, M.; Luo, X.; Luo, Z.; and Tao, D. 2023. Omg: Towards effective graph classification against label noise. *IEEE Transactions on Knowledge and Data Engineering*, 35(12): 12873–12886.

Yin, X.; Shi, J.; Zhang, Y.; Lu, Y.; Zhang, Z.; Xie, Y.; and Qu, Y. 2024b. Robust Pseudo-label Learning with Neighbor Relation for Unsupervised Visible-Infrared Person Re-Identification. *arXiv preprint arXiv:2405.05613*.

Yu, H.; Cheng, X.; Peng, W.; Liu, W.; and Zhao, G. 2023. Modality unifying network for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11185–11195.

Yu, S.; Chen, D.; Zhao, R.; Chen, H.; and Qiao, Y. 2021. Neighbourhood-guided feature reconstruction for occluded person re-identification. *arXiv preprint arXiv:2105.07345*.

Zhang, X.; Li, D.; Wang, Z.; Wang, J.; Ding, E.; Shi, J. Q.; Zhang, Z.; and Wang, J. 2022. Implicit sample extension for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7369–7378.

Zhao, J.; Lan, L.; Huang, D.; Ren, J.; and Yang, W. 2022a. Heterogeneous pseudo-supervised learning for few-shot person re-identification. *Neural Networks*, 154: 521–537.

Zhao, J.; Wang, H.; Zhou, Y.; Yao, R.; Chen, S.; and El Saddik, A. 2022b. Spatial-channel enhanced transformer for visible-infrared person re-identification. *IEEE Transactions on Multimedia*.

Zhong, Z.; Zheng, L.; Cao, D.; and Li, S. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1318–1327.