Robust Duality Learning for Unsupervised Visible-Infrared Person Re-Identification

Yongxiang Li, Yuan Sun, Yang Qin, *Graduate Student Member, IEEE*, Dezhong Peng, Xi Peng, *Senior Member, IEEE*, and Peng Hu, *Member, IEEE*

Abstract—Unsupervised visible-infrared person re-identification (UVI-ReID) aims at retrieving pedestrian images of the same individual across distinct modalities, presenting challenges due to the inherent heterogeneity gap and the absence of cost-prohibitive annotations. Although existing methods employ self-training with clustering-generated pseudo-labels to bridge this gap, they always implicitly assume that these pseudo-labels are predicted correctly. In practice, however, this presumption is impossible to satisfy due to the difficulty of training a perfect model let alone without any ground truths, resulting in pseudo-labeling errors. Based on the observation, this study introduces a new learning paradigm for UVI-ReID considering Pseudo-Label Noise (PLN), which encompasses three challenges: noise overfitting, error accumulation, and noisy cluster correspondence. To conquer these challenges, we propose a novel robust duality learning framework (RoDE) for UVI-ReID to mitigate the adverse impact of noisy pseudo-labels. Specifically, for noise overfitting, we propose a novel Robust Adaptive Learning mechanism (RAL) to dynamically prioritize clean samples while deprioritizing noisy ones, thus avoiding overemphasizing noise. To circumvent error accumulation of self-training, where the model tends to confirm its mistakes, RoDE alternately trains dual distinct models using pseudo-labels predicted by their counterparts, thereby maintaining diversity and avoiding collapse into noise. However, this will lead to cross-cluster misalignment between the two distinct models, not to mention the misalignment between different modalities, resulting in dual noisy cluster correspondence and thus difficult to optimize. To address this issue, a Cluster Consistency Matching mechanism (CCM) is presented to ensure reliable alignment across distinct modalities as well as across different models by leveraging cross-cluster similarities. Extensive experiments on three benchmark datasets demonstrate the effectiveness of the proposed RoDE.

Index Terms—Unsupervised VI-ReID, pseudo-label noise, noise correspondence, cluster consistency.

I. INTRODUCTION

VISIBLE-INFRARED Person Re-Identification (VI-ReID) seeks to match pedestrians of the same identity across visible and infrared modalities, serving critical roles in various scenarios [1], [2], [3], [4] such as military surveillance, and intelligent security. This technology effectively enhances the precise identification and response capabilities in these fields, ensuring the overall security and stability of society.

The major challenge in VI-ReID is learning modality-invariant representations to bridge the significant heterogeneity gap between visible and infrared data. To this end, numerous VI-ReID methods are proposed to project different modalities into a latent common space, which could be roughly categorized into supervised VI-ReID (SVI-ReID) and unsupervised VI-ReID (UVI-ReID). Specifically, SVI-ReID methods exploit identification labels to learn the semantic consistency across distinct modalities but are impractical due to the high cost of collecting a large amount of well-labeled multimodal data. Conversely, UVI-ReID methods circumvent this limitation, making them more practical [6], but facing increased difficulty in deriving cross-modal consistency from unlabeled data.

To tackle this pivotal challenge, some UVI-ReID methods employ clustering techniques to generate pseudo-labels for each modality [7], [8], thereby establishing cross-modal correspondences and learning modality-invariant representations. However, these methods ignore the problem of various types of noise interference caused by pseudo-labels, including noise overfitting, error accumulation, and noisy cluster correspondence, collectively referred to as Pseudo-Label Noise (PLN). These training noises often occur together and misguide the optimization of multimodal models, thereby leading to serious error accumulation and overfitting. To illustrate this problem, we depict it in Fig. 1, where it can be seen that PLN is a pervasive but neglected issue. Furthermore, while some researchers [9] have explored twin noisy label problems (including noisy label and noisy correspondence) in SVI-ReID, they presume a reliable cross-modal consistency, which is absent in UVI-ReID due to the chaotic correspondence of modality-specific clusters. That is why we refer to it as noisy

(a) Noise Overfitting　　　　　(b) Error Accumulation

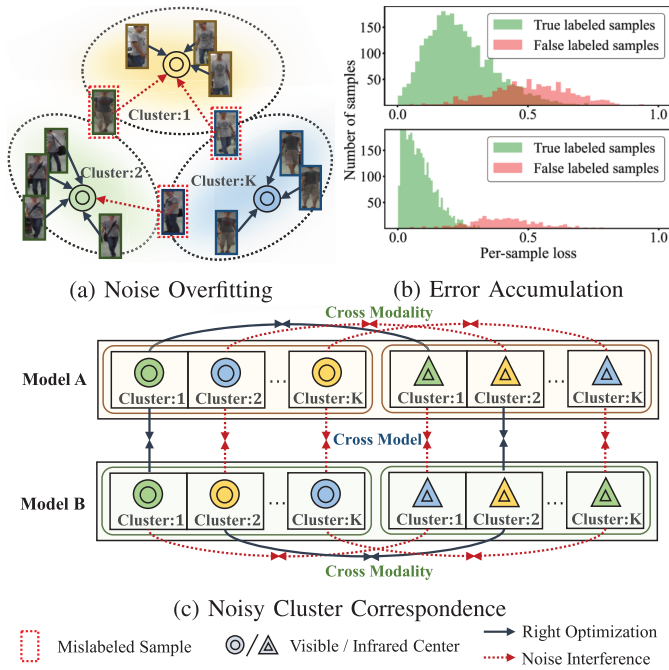(c) Noisy Cluster Correspondence

Fig. 1. Pseudo-label noise issues in UVI-ReID. (a) In intra-modality, some sample features are close to the adjacent cluster center, leading to false pseudo-label assignments and noise overfitting. (b) Error accumulation for a single model (TOP) and dual models (BOTTOM) is depicted through the per-sample loss distribution on the infrared modality of RegDB dataset using the recent IMSL method [9]. The dual models employ a cross-training strategy, using the pseudo-labels generated by one model as the ground truth for the other. During training, inevitable error annotations and cluster mismatches introduce significant noise. For a single model, noisy and clean samples intermingle due to severe error accumulation, as indicated by the overlapping color parts. In contrast, using dual models significantly mitigates this issue. (c) Semantic misalignment occurs across different clusters, including distinct models and modalities, which are regarded as noisy cluster correspondences.

cluster correspondence rather than label noise, presenting a more daunting and complex challenge.

To address the aforementioned challenges, we propose a novel unsupervised visible-infrared framework RoDE, to robustly learn from PLN using visible-infrared pedestrian images, as illustrated in Fig. 2. Specifically, RoDE incorporates the following tangible solutions: 1) To counter noise overfitting, we propose a Robust Adaptive Learning mechanism (RAL) that categorizes samples into clean and noisy subsets, and then dynamically prioritizes clean samples while deprioritizing noisy ones by using a robust loss function, thereby reducing the influence of mislabeled samples and mitigating overfitting noise. 2) To avoid error accumulation, we simultaneously train two different models using Robust Duality Learning (RDL), each using the predictions of the other model, thereby diversifying the supervision information. Thanks to this diversity, our RoDE could prevent each model from being overconfident about its own incorrect predictions, thus avoiding accumulating errors. 3) To tackle noisy cluster correspondence caused by the dual models, we present a novel Cluster Consistency Matching mechanism (CCM), which matches distinct clusters by utilizing the distances between different centers, producing more reliable correspondence.

Unlike existing UVI-ReID methods that learn from pseudo-labels [7], [9], [10], our RoDE addresses not only noise overfitting but also error accumulation. More specifically, most of these methods use a binary robust strategy [3], [10], selectively focusing on confident samples and discarding all unreliable ones, which leads to information loss and performance degradation. In contrast, our RoDE reweights all samples adaptively, thereby avoiding the rough discard of data and reducing the adverse impact of noise. However, current robust methods focus solely on robust training techniques like sample selection and robust loss functions to alleviate noise, thereby becoming overconfident in their predictions even when incorrect, i.e. error accumulation. Our RoDE counters this issue by using two distinct models that alternately guide each other, diversifying supervision and preventing overconfidence, as shown in Fig. 1 (b). In addition, the intrinsic cross-modal and cross-model gaps lead to cross-cluster misalignment, referred to as dual noisy cluster correspondence. Intuitively, this noise would be more challenging than the single cross-cluster noise across different modalities in prior works [2]. To address this challenge, our RoDE matches distinct clusters across modalities and models, ensuring more reliable correspondence and enhancing overall robustness.

Our main contributions can be summarized as follows:

- In this paper, we propose a novel UVI-ReID framework RoDE to robustly learn discriminative representations and establish cross-modal re-identification relationships in a latent common space, addressing noise overfitting, error accumulation, and noisy cluster correspondence simultaneously.

- To resist the interference of noisy overfitting, we design a novel RAL mechanism that utilizes a self-adaptive strategy and a demonstrably robust loss function to prioritize clean samples, thereby enhancing robustness.

- We present a RDL training pipeline that jointly trains two different models to prevent error accumulation in self-training. To meet UVI-ReID requirements, CCM mechanism is introduced to address noisy cluster correspondence, encompassing both cross-modal and cross-model scenarios.

- Extensive experiments on the SYSU-MM01, RegDB, and LLCM datasets highlight the superiority of our method and establish a powerful baseline for the UVI-ReID task.

II. RELATED WORKS

A. Supervised Visible-Infrared Person Re-Identification

SVI-ReID is a subtask of cross model retrieval [11], which aims to match visible images of individuals with their infrared counterparts. To tackle cross-modal discrepancies, several supervised VI-ReID methods have been proposed to learn modality-invariant features [12]. For example, HSME [13] uses a hypersphere manifold embedding with sphere softmax. MPANet leverages a joint modality and pattern alignment network to uncover cross-modal differences [14]. TransVI employs a Transformer-based approach with a two-stream structure to capture modality-specific features and learn shared knowledge [15]. Additionally, [16] introduces a style-agnostic
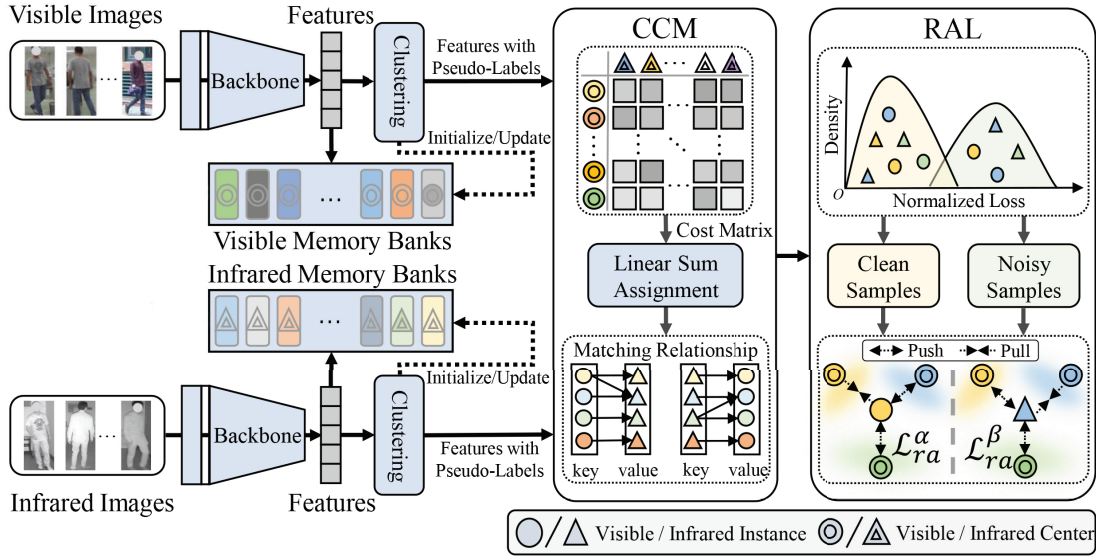
Fig. 2. The framework of the proposed RoDE. The model projects the visible and infrared images into the common space using the modality-specific networks $f^{\mathcal{P}}(\cdot; \Theta^{\mathcal{P}})$. CCM (See Section III-E) and RAL (See Section III-C) are used to alleviate noisy cluster correspondence and noisy overfitting. Specifically, cross-modal and cross-model CCM are utilized to establish the correspondence across different modalities and different models, respectively. Moreover, RAL divides the pseudo-labels into clean and noisy subsets, and adaptively adjusts the focus on them, thereby enhancing robustness against noisy overfitting.

framework that bridges modality gaps at both data and feature levels. However, these methods, while effective with ample cross-modal annotations, are limited in real-world scenarios due to their dependence on visible-infrared identity labels.

B. Unsupervised Visible-Infrared Person Re-ID

Recently, UVI-ReID has gained significant attention for its lower labeling costs and practical applications in night surveillance [2], [3], [8], [17], [18]. Unlike traditional methods, UVI-ReID cannot use cross-modal labeled pairs to learn modality-invariant features. Instead, a common approach is to generate pseudo-labels for images from each modality and use these pseudo-supervisions to learn a shared discriminative representation [19], [20], [21]. However, this strategy often requires additional RGB datasets for pre-training to acquire prior knowledge, which limits its practicality. Alternatively, some methods propose to generate pseudo-labels by exploring cluster-level relationships across different modalities through cross-modal memory aggregation, which can effectively capture the multimodal semantic consistency without any extra assistance [2], [6], [22]. Moreover, a series of methods pay attention to building cross-modal associations to embed domain information mutually, achieving remarkable performance [4], [7]. However, these methods usually ignore the serious PLN problems during the training process or only notice a certain aspect of the impact of PLN, thereby misleading model optimization direction and degrading performance.

C. Robustly Learning With Noisy Labels

The problem of noisy labels presents a significant challenge during training, potentially misdirecting the learning process [23], [24], [25]. To combat this negative influence, existing methods can be classified into three categories:

sample selection, label correction, and noise regularization. Previous researches on sample selection aim to detect noisy labels by the natural resistance of neural networks to noise, often relying on batch statistics for robustness against label noise [26]. Another research direction focuses on label correction, typically attempting to rectify sample labels using model predictions [27]. Additionally, some studies emphasize noise regularization techniques such as mixup [28], or dedicated loss terms [29]. Unsupervised regularization has also been proven to enhance the classification accuracy of neural networks when trained on noisy datasets. However, in UVI-ReID, we will face more challenging and complex problems caused by noisy labels, including noisy overfitting, error accumulation, and noisy cluster correspondence. These challenges necessitate novel strategies that can simultaneously address above issues to ensure robust and accurate model performance.

III. METHODOLOGY

A. Problem Statement and Notations

Let $\mathcal{X} = \{\mathcal{X}^{\mathcal{V}}, \mathcal{X}^{\mathcal{I}}\}$ be the label-free visible-infrared training dataset, where $\mathcal{X}^{\mathcal{V}} = \{x_1^{\mathcal{V}}, x_2^{\mathcal{V}}, \cdots, x_{N^{\mathcal{V}}}^{\mathcal{V}}\}$ denotes the $N^{\mathcal{V}}$ visible images and $\mathcal{X}^{\mathcal{I}} = \{x_1^{\mathcal{I}}, x_2^{\mathcal{I}}, \cdots, x_{N^{\mathcal{I}}}^{\mathcal{I}}\}$ denotes the $N^{\mathcal{I}}$ infrared images. For convenience, $x_i^{\mathcal{P}}$ is used to denote the i-th image in the $\mathcal{P} \in \{\mathcal{V}, \mathcal{I}\}$ modality, and $N^{\mathcal{P}}$ is the number of images in the $\mathcal{P}$ modality. UVI-ReID aims to learn common representations from the unlabeled and unaligned visible-infrared dataset $\mathcal{X}$, thereby enabling the accurate retrieval of the most semantically relevant sample in another modality using the given query.

To achieve this, we first use two modality-specific nonlinear neural network projectors, $\{f^{\mathcal{P}}(\cdot; \Theta^{\mathcal{P}})\}_{\mathcal{P} \in \{\mathcal{V}, \mathcal{I}\}}$, to map images from each modality into an L-dimensional representation space, where $\Theta^{\mathcal{P}}$ represents the trainable parameters for the

TABLE I

COMPARISON WITH THE RECENT METHODS ON SYSU-MM01 AND RegDB DATASETS. RANK-1 (%), RANK-10 (%), RANK-20 (%), MAP (%) AND MINP (%) ARE REPORTED. THE HIGHEST SCORE IS SHOWN IN BOLD, WHILE THE SECOND HIGHEST SCORE IS UNDERLINED

| | Methods | Venue | SYSU-MM01 | | | | | | RegDB | | | | | |
| | | | All Search | | | Indoor Search | | | V2T | | | T2V | | |
			Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP
SVI-ReID	AGW [31]	TPAMI'21	47.50	47.65	35.30	54.17	62.97	59.23	70.05	66.37	50.19	70.49	65.90	51.24
	CA [32]	ICCV'21	69.88	66.89	53.61	76.26	80.37	76.79	85.03	79.14	65.33	84.75	77.82	61.56
	DART [9]	CVPR'22	68.72	66.29	53.26	72.52	78.17	74.94	83.60	75.67	-	81.97	75.13	-
	SPOT [33]	TIP'22	65.34	62.25	48.86	69.42	74.63	70.48	80.35	72.46	56.19	79.37	72.26	56.06
	CMTR [34]	TMM'23	65.45	62.90	-	71.46	76.67	-	88.11	81.66	-	84.92	80.79	-
	PMT [35]	AAAI'23	67.53	64.98	51.86	71.66	76.52	72.74	84.83	76.55	-	84.16	75.13	-
	TransVI [15]	TCSVT'23	71.36	68.63	-	77.40	81.31	-	96.66	91.22	-	96.30	91.21	-
	STAR [16]	TMM'23	76.07	72.73	-	83.47	85.76	-	94.09	88.75	-	93.30	88.20	-
	DMA [1]	TIFS'24	74.57	70.41	56.50	82.85	85.10	-	93.30	88.34	-	91.50	86.80	-
UVI-ReID	SPCL [36]	NIPS'20	18.37	19.39	10.99	26.83	36.42	33.05	13.59	14.86	10.36	11.70	13.56	10.09
	MMT [37]	ICLR'20	21.47	21.53	11.50	22.79	31.50	27.66	25.68	26.51	19.56	25.59	18.66	-
	IICS [38]	CVPR'21	14.39	15.74	8.41	15.91	24.87	22.15	10.30	11.94	8.10	10.39	11.23	7.04
	CAP [39]	AAAI'21	16.82	15.71	7.02	24.57	30.74	26.15	84.83	76.55	-	84.16	75.13	-
	H2H [40]	TIP'21	23.81	18.87	-	-	-	-	13.91	12.72	-	14.11	12.29	-
	PPLR [41]	CVPR'22	11.98	12.25	4.97	12.71	20.81	17.61	10.30	11.94	8.10	10.39	11.23	7.04
	OTAL [19]	ECCV'22	29.90	27.10	-	29.80	38.80	-	32.90	29.70	-	32.10	28.60	-
	ADCA [6]	MM'22	45.51	42.73	28.29	50.60	59.11	55.17	67.20	64.05	52.67	68.48	63.81	49.62
	DOTLA [7]	MM'23	50.36	47.36	32.40	53.47	61.73	57.35	85.63	76.71	61.58	82.91	74.97	58.60
	CCLNet [42]	MM'23	54.03	50.19	-	56.68	65.12	-	69.94	65.53	-	70.17	66.66	-
	GUR [4]	ICCV'23	60.95	56.99	41.85	64.22	69.49	64.81	73.91	70.23	58.88	75.00	69.94	56.21
	PGM [2]	CVPR'23	57.27	51.78	34.96	56.23	62.74	58.13	69.48	65.41	38.72	69.85	65.17	58.47
	DFC [43]	IPM'23	40.92	36.20	-	44.12	28.36	-	38.88	38.11	-	-	-	-
	CHCR [3]	TCSVT'23	47.72	45.34	-	-	-	-	68.18	63.75	-	69.08	63.95	-
	TAA [44]	TIP'23	48.77	42.43	25.37	50.12	56.02	49.96	62.23	56.00	41.51	63.79	56.53	38.99
	IMSL [5]	TCSVT'24	57.96	53.93	-	58.30	64.31	-	70.08	66.30	-	70.67	66.35	-
	SCA-RCP [10]	TKDE'24	51.41	48.52	33.56	56.77	64.19	59.25	85.59	78.12	-	82.41	75.73	-
	RoDE	Ours	**62.88**	**57.91**	**43.04**	**64.53**	**70.42**	**66.04**	**88.77**	**78.98**	**67.99**	**85.78**	**78.43**	**62.34**

TABLE II

COMPARISON WITH THE RECENT METHODS ON LLCM DATASET. RANK-1 (%) AND MAP (%) ARE REPORTED

| | Methods | Venue | V2T | | T2V | |
			Rank-1	mAP	Rank-1	mAP
SVI-ReID	LBA [47]	ICCV'21	50.85	55.63	43.61	51.86
	AGW [31]	TPAMI'21	51.51	55.34	43.65	51.87
	MMN [48]	MM'21	59.97	62.75	52.53	58.99
	DEEN [49]	CVPR'23	62.57	65.81	54.92	62.95
UVI-ReID	CAP [39]	AAAI'21	8.16	10.14	7.28	9.67
	P2LR [50]	AAAI'22	16.38	19.84	14.85	17.15
	OTLA [19]	ECCV'22	17.88	20.46	14.97	18.66
	ADCA [6]	MM'22	23.57	28.25	16.16	21.48
	GUR [4]	ICCV'23	31.47	34.77	29.68	33.38
	DOTLA [7]	MM'23	27.14	26.26	23.52	27.48
	IMSL [5]	TCSVT'24	22.74	19.38	17.26	24.38
	SCA-RCP [10]	TKDE'24	29.11	33.33	22.36	28.05
	RoDE	Ours	**35.13**	**37.44**	**32.73**	**36.64**

network of modality $\mathcal{P}$. In other words, each sample $\boldsymbol{x}_i^{\mathcal{P}}$ is projected as a feature vector $\boldsymbol{v}_i^{\mathcal{P}} \in \mathbb{R}^{1 \times L}$ by

$$\boldsymbol{v}_i^{\mathcal{P}} = f^{\mathcal{P}}(\boldsymbol{x}_i^{\mathcal{P}}; \Theta^{\mathcal{P}}). \tag{1}$$

To mine semantic information from unlabeled data, we follow prior works [2], [4] and employ classic clustering methods such as DBSCAN [30] to cluster the projected representations for each modality, assigning modality-specific pseudo-labels to each sample based on its nearest cluster center. Unfortunately, due to the lack of correspondence between modalities, establishing semantic associations between different modalities is challenging, which impedes the achievement of UVI-ReID. To address this issue, pseudo-labels across different modalities should be matched based on the distances between the cluster centers to mitigate the modality gap. Additionally, clustering to obtain pseudo-labels at each epoch introduces considerable variability for identical instances. To maintain the stability and reliability of pseudo-labels, we construct memory banks $\mathcal{M}^{\mathcal{P}}$ to store all cluster centers, which are iteratively updated with newly generated cluster centers. This mechanism prevents the model from frequently reassigning different pseudo-labels to the same individuals across epochs, which could introduce confusion and instability in the training process.

$$\mathcal{M}^{\mathcal{P}} = [\boldsymbol{m}_1^{\mathcal{P}}, \cdots, \boldsymbol{m}_{K^{\mathcal{P}}}^{\mathcal{P}}] \in \mathbb{R}^{K^{\mathcal{P}} \times L}, \tag{2}$$

where $\boldsymbol{m}_i^{\mathcal{P}}$ represents the center of the i-th cluster, and $K^{\mathcal{P}}$ denotes the count of clusters in modality $\mathcal{P}$. The memory banks are iteratively updated through momentum after each epoch, i.e.,

$$\boldsymbol{m}_j^{\mathcal{P}} = \eta \boldsymbol{m}_j^{\mathcal{P}} + (1 - \eta) \bar{\boldsymbol{v}}_j^{\mathcal{P}}, \tag{3}$$

where $\bar{\boldsymbol{v}}_j^{\mathcal{P}}$ is the mean feature in the j-th class, and $\eta \in [0, 1]$ is the memory updating rate. Notably, each center represents the mean feature of all samples within the same cluster, allowing

TABLE III
ABLATION STUDIES OF RODE ON SYSU-MM01, REGDB AND LLCM DATASETS

Order	Components	SYSU-MM01: All Search			RegDB:V2T			LLCM:V2T		
		Rank-1	mAP	mINP	Rank-1	mAP	mINP	Rank-1	mAP	mINP
1	RoDE w/o RAL ($\mathcal{L}_{ra}^{\alpha}$)	54.17	52.96	32.11	75.87	72.46	60.33	33.72	34.18	21.72
2	RoDE w/o RAL ($\mathcal{L}_{ra}^{\beta}$)	59.89	55.83	38.39	84.36	75.36	62.09	34.25	35.38	21.97
3	RoDE w/o CCM (cross-model)	3.56	6.39	3.70	8.03	8.08	8.69	5.37	7.28	4.49
4	RoDE w/o CCM (cross-modal)	44.31	43.90	29.83	52.88	45.98	29.25	30.18	31.94	18.44
5	RoDE	**62.88**	**57.91**	**43.04**	**88.77**	**78.98**	**67.99**	**35.13**	**37.44**	**23.04**

the memory banks to semantically distinguish the samples based on their distances from the cluster centers in the feature space.

To learn representations that are both discriminative and modality-invariant under the supervision of pseudo-labels, the standard Cross-Entropy loss (CE) could be utilized to maximize the intra-modal conditional probabilities $p(\widetilde{y}_i^{\mathcal{P}}|\boldsymbol{x}_i^{\mathcal{P}})$ and the inter-modal conditional probabilities $p(\widetilde{y}_i^{\mathcal{Q}}|\boldsymbol{x}_i^{\mathcal{P}})$ as follows:

$$\mathcal{L}_{ce} = - \sum_{\mathcal{P}\in\{\mathcal{V},\mathcal{I}\}} \sum_{i=1}^{N_B} \left(\widetilde{y}_i^{\mathcal{P}} \log\left(p(\widetilde{y}_i^{\mathcal{P}}|\boldsymbol{x}_i^{\mathcal{P}})\right) \right.$$
$$\left. + \widetilde{y}_i^{\mathcal{Q}} \log\left(p(\widetilde{y}_i^{\mathcal{Q}}|\boldsymbol{x}_i^{\mathcal{P}})\right) \right), \quad (4)$$

where $\boldsymbol{x}_i^{\mathcal{P}}$ belongs to the clustering center $\widetilde{y}_i^{\mathcal{P}}$ (i.e., pseudo-label). $\mathcal{Q} \in \{\mathcal{V},\mathcal{I}\}$ and $\mathcal{Q} \neq \mathcal{P}$, N_B is the batch size, and $p(\widetilde{y}_i^{\mathcal{P}}|\boldsymbol{x}_i^{\mathcal{P}})$ and $p(\widetilde{y}_i^{\mathcal{Q}}|\boldsymbol{x}_i^{\mathcal{P}})$ are calculated by:

$$p(\widetilde{y}_i^{\mathcal{P}}|\boldsymbol{x}_i^{\mathcal{P}}) = \frac{\exp((\boldsymbol{m}_{\widetilde{y}_i^{\mathcal{P}}}^{\mathcal{P}})^T \cdot \boldsymbol{v}_i^{\mathcal{P}}/\tau)}{\sum_{k=1}^{K^{\mathcal{P}}} \exp\left((\boldsymbol{m}_k^{\mathcal{P}})^T \cdot \boldsymbol{v}_i^{\mathcal{P}}/\tau\right)}, \quad (5)$$

$$p(\widetilde{y}_i^{\mathcal{Q}}|\boldsymbol{x}_i^{\mathcal{P}}) = \frac{\exp((\boldsymbol{m}_{\widetilde{y}_i^{\mathcal{Q}}}^{\mathcal{Q}})^T \cdot \boldsymbol{v}_i^{\mathcal{P}}/\tau)}{\sum_{k=1}^{K^{\mathcal{P}}} \exp\left((\boldsymbol{m}_k^{\mathcal{P}})^T \cdot \boldsymbol{v}_i^{\mathcal{P}}/\tau\right)}, \quad (6)$$

where τ is a temperature parameter. Although existing methods achieve promising performance by using CE, they implicitly assume that the pseudo-labels are accurately annotated. Unfortunately, it is difficult or even possible to label the unlabeled data accurately, which inevitably introduces noise into pseudo-labels, leading to noise overfitting and error accumulation. This is also demonstrated in Table III, where the noise in pseudo-labels disrupts the cross-modal association learning, seriously affecting re-identification performance.

B. Robust Adaptive Learning Mechanism

To tackle the noise overfitting issue in UVI-ReID, most existing UVI-ReID approaches mitigate noise overfitting by employing binary robust strategies that selectively emphasize confident samples while disregarding unreliable ones [3], [10]. Although these methods reduce the impact of noise, their unreliable partitioning inadvertently leads to the loss of information from false negative samples, resulting in suboptimal performance. In contrast, we introduce the Robust Adaptive Learning (RAL) mechanism, which adaptively ensures a selective emphasis, effectively mitigating noise without sacrificing valuable information in unreliable samples. To be specific, to alleviate excessive optimization of samples with low reliability,

we first design a robust loss function, $\mathcal{L}_{ra}$, explicitly to resist noise interference, which is defined as:

$$\mathcal{L}_{ra} = - \sum_{\mathcal{P}\in\{\mathcal{V},\mathcal{I}\}} \sum_{i=1}^{N^{\mathcal{P}}} p^{\gamma_i}\left(\widetilde{y}_i^{\mathcal{P}}|\boldsymbol{x}_i^{\mathcal{P}}\right), \quad (7)$$

where $\gamma_i \in (0,1]$ is used to adjust the strength of optimization for each sample. Moreover, we provide the property of $\mathcal{L}_{ra}$ to better understand its robustness against noisy labels:

Property 1: For any input $(\boldsymbol{x}_i^{\mathcal{P}},\widetilde{y}_i^{\mathcal{P}})$ and $\gamma_i \in (0,1]$, the loss function $\mathcal{L}_{ra}$ exhibits the following behaviors:

1) As γ_i approaches to 0, $\mathcal{L}_{ra}$ gradually behaves like the CE loss.
2) As γ_i approaches to 1, $\mathcal{L}_{ra}$ tends to optimize equally for all samples.

According to Property 1, one could infer that $\mathcal{L}_{ra}$ effectively reduces the focus on mislabeled samples, alleviating the overfitting issue caused by pseudo-labeling noise. Additionally, $\mathcal{L}_{ra}$ does not treat all samples equally, thereby mitigating the underfitting issue. Consequently, it improves performance while maintaining robustness against noise by appropriately attending to challenging samples. Meanwhile, the preference towards robustness and strong optimization is regulated by γ_i. Note that the detailed proofs for Property 1 are available in the Appendix (see Supplementary Material).

However, due to the varying requirements of different samples for optimization, it is almost impossible to tune the parameter γ_i for each sample manually, while using a single unified parameter for all samples would lead to suboptimal performance. To be ideal, the model should strongly optimize for confident samples (i.e., $\gamma_i \to 0$) while remaining robust to noisy labeled samples (i.e., $\gamma_i \to 1$). To this end, we present an adaptive method to determine the appropriate γ_i for each sample, thereby avoiding the above dilemma. Specifically, we first define the per-sample loss as $\boldsymbol{\ell}^{\mathcal{P}} = \{\ell_i^{\mathcal{P}}\}_{i=1}^{N^{\mathcal{P}}}$ to measure the difference between predictions and pseudo-labels, where $\ell_i^{\mathcal{P}}$ is calculated by $\mathcal{L}_i^{\mathcal{P}}$ (See Equation (18)). In other words, this difference could reflect the reliability of the pseudo-labels. Therefore, based on the per-sample losses, we can use a two-component Gaussian Mixture Model (GMM) to classify samples with the lower mean as the clean set and those with the higher mean as the noisy set.

$$p(\boldsymbol{\ell}^{\mathcal{P}}|\Theta^G) = \sum_{k=1}^{2} \delta_k \phi(\boldsymbol{\ell}^{\mathcal{P}}|k), \quad (8)$$
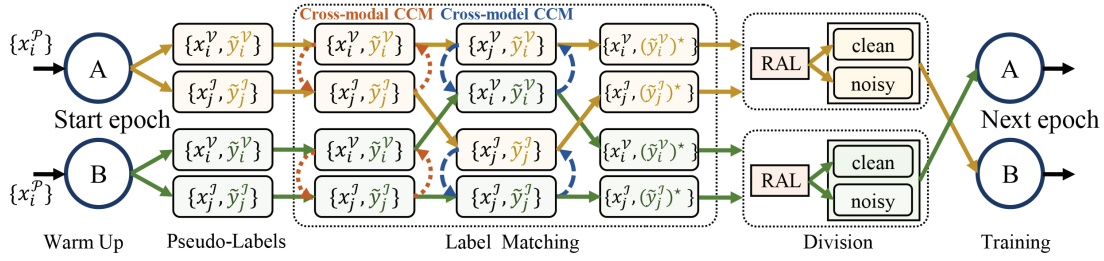
Fig. 3. The training pipeline of the proposed RoDE. RoDE consists of two individual models A and B, which are trained collaboratively by exchanging their pseudo supervisions. Before training, RoDE pre-warms up the models A and B individually by predicting pseudo-labels and self-training. After warming up, the two models are co-trained with CCM and RAL.

where δ_k and $\phi(\ell^{\mathcal{P}}|k)$ are the mixture coefficient and the probability density of the k-th component respectively, and Θ^G is the parameter of GMM. Although GMM is employed in our approach to distinguish clean and noisy samples, it is important to emphasize that GMM serves as one of several possible tools for data partitioning. Other clustering methods, such as K-Means or Beta Mixture Models (BMM), could also be utilized in place of GMM without compromising the fundamental contributions of our method. In addition, we compute the posterior probability $w_i = p(k)p(\ell_i^{\mathcal{P}}|k)/p(\ell_i^{\mathcal{P}})$ as the probability that the i-th sample belongs to the clean set. Based on the aforementioned discussion, it is optimal to assign small γ_i to reliable samples while large γ_i to noisy ones. To achieve this, we employ a sharpening strategy to calculate γ_i adaptively as follows:

$$\gamma_i = \log\left((1 - w_i)^{0.25}/\mu + 1\right), \qquad (9)$$

where μ is a scale parameter. Consequently, RAL can not only mitigate the detrimental effects of noise overfitting but also preserve useful information that might otherwise be lost.

C. Robust Duality Learning Pipeline

To mitigate the interference of the noise in pseudo-labels, current methods predominantly focus only on robust training techniques [7], [9], [10]. However, these methods often neglect the issue of overconfidence in predictions even for incorrect ones, which leads to error accumulation as shown in Fig. 1 (b). To address this problem, we propose a Robust Duality Learning (RDL) pipeline to prevent the model from being overconfident about its own incorrect predictions, which alternately trains two individual networks with the same architecture but different initializations (denoted as $A = \{f_A^{\mathcal{P}}(\cdot; \Theta_A^{\mathcal{P}}), \mathcal{M}_A^{\mathcal{P}}\}$ and $B = \{f_B^{\mathcal{P}}(\cdot; \Theta_B^{\mathcal{P}}), \mathcal{M}_B^{\mathcal{P}}\}$), as depicted in Fig. 3.

Specifically, in each epoch, models A and B first generate their own pseudo-labels $A : (\widetilde{y}_i^{\mathcal{P}}|_A, \widetilde{y}_i^{\mathcal{Q}}|_A)$ and $B : (\widetilde{y}_i^{\mathcal{P}}|_B, \widetilde{y}_i^{\mathcal{Q}}|_B)$ through a clustering method (e.g., DBSCAN [30]). Subsequently, model A leverages the pseudo-labels generated by model B for optimization, and vice versa. This mutual learning process enhances the diversity of learning and reduces the overconfidence to incorrect pseudo-labels in the specific model. However, the lack of pre-given correspondences can lead to unavoidable cluster mismatches across various modalities and models. This mismatch would seriously disrupt the optimization direction in the alternating learning

process, resulting in poor performance. To establish cluster consistency, we propose the Cross-Cluster Matching (CCM) (see Section III-D), which aligns the two sets of clusters by utilizing correlations in both centers and identities, thereby producing the pseudo-label with reliable correspondences $A : ((\widetilde{y}_i^{\mathcal{P}}|_A)^\star, (\widetilde{y}_i^{\mathcal{Q}}|_A)^\star)$ and $B : ((\widetilde{y}_i^{\mathcal{P}}|_B)^\star, (\widetilde{y}_i^{\mathcal{Q}}|_B)^\star)$. Using the aligned labels, we optimize each model with RAL (see Section III-B) by minimizing both intra-modal loss $\mathcal{L}_{ra}^\alpha$ and inter-modal loss $\mathcal{L}_{ra}^\beta$. This dual loss minimization aims to reduce the impact of incorrectly labeled samples and enhance noise tolerance. Finally, the models are guided toward the correct optimization direction under cross supervision. To be specific, the objective functions $\mathcal{L}_{ra}^\alpha$ and $\mathcal{L}_{ra}^\beta$ of model A could be rewritten as:

$$\mathcal{L}_{ra}^\alpha = -\sum_{\mathcal{P}\in\{\mathcal{V},\mathcal{I}\}} \sum_{i=1}^{N^{\mathcal{P}}} p^{\gamma_i}\left((\widetilde{y}_i^{\mathcal{P}}|_B)^\star|\boldsymbol{x}_i^{\mathcal{P}}\right), \qquad (10)$$

$$\mathcal{L}_{ra}^\beta = -\sum_{\mathcal{P}\in\{\mathcal{V},\mathcal{I}\}} \sum_{i=1}^{N^{\mathcal{P}}} p^{\gamma_i}\left((\widetilde{y}_i^{\mathcal{Q}}|_B)^\star|\boldsymbol{x}_i^{\mathcal{P}}\right), \qquad (11)$$

and the objective functions of model B could be

$$\mathcal{L}_{ra}^\alpha = -\sum_{\mathcal{P}\in\{\mathcal{V},\mathcal{I}\}} \sum_{i=1}^{N^{\mathcal{P}}} p^{\gamma_i}\left((\widetilde{y}_i^{\mathcal{P}}|_A)^\star|\boldsymbol{x}_i^{\mathcal{P}}\right), \qquad (12)$$

$$\mathcal{L}_{ra}^\beta = -\sum_{\mathcal{P}\in\{\mathcal{V},\mathcal{I}\}} \sum_{i=1}^{N^{\mathcal{P}}} p^{\gamma_i}\left((\widetilde{y}_i^{\mathcal{Q}}|_A)^\star|\boldsymbol{x}_i^{\mathcal{P}}\right). \qquad (13)$$

D. Cluster Consistency Matching

In the RDL pipeline, inherent discrepancies between cross-modal and cross-model elements lead to cross-cluster misalignment, referred to as dual noisy cluster correspondence. This form of noise presents a more significant challenge compared to the single cross-cluster noise encountered in previous studies on different modalities [2]. Dual noisy cluster correspondences exacerbate mismatching issues, leading to unstable training and even failure to converge. This significantly hinders the learning of association information between cross-modal and cross-model elements. To overcome this issue, we introduce the CCM mechanism, which aims to correlate the intrinsic characteristics of each cluster, thus aligning cluster centers across different modalities or models. Additionally, CCM accounts for the complex interactions between clusters across multiple modalities and models, providing a more comprehensive solution, as illustrated in Fig. 4.
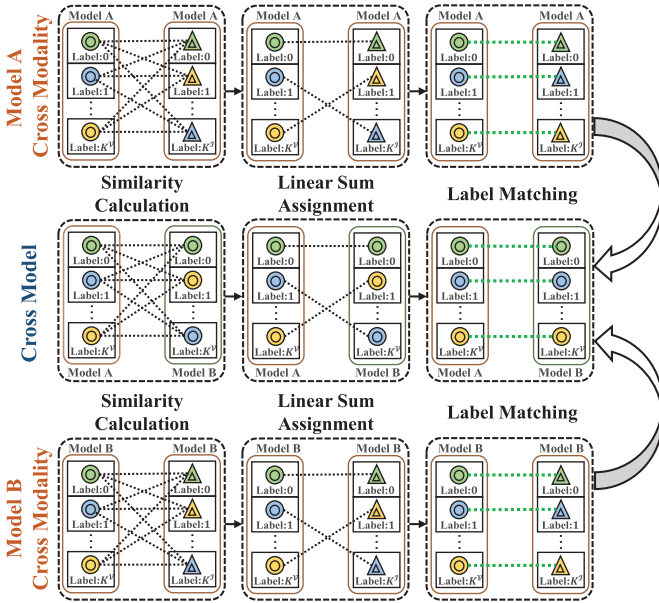
Fig. 4. The solution of cluster inconsistency issue. ◎ and △ represent visible and infrared modality centers respectively. The *green dotted lines* denote correct matches after CCM.

modalities are harmonized. Then, we align clusters across distinct models (i.e. Equation (17)) to ensure cross-model consistency and coherence. By maintaining consistency between the clusters produced by different models, we reduce the risk of error accumulation and enhance the overall robustness of the system. These two alignment strategies collectively improve the ability of model to accurately identify and match individuals across modalities.

$$\text{Cross-modal Correspondence:} \begin{cases} A : \mathcal{I} \leftrightarrow A : \mathcal{V} \\ B : \mathcal{I} \leftrightarrow B : \mathcal{V} \end{cases}, \quad (16)$$

$$\text{Cross-model Correspondence:} \begin{cases} A : \mathcal{V} \leftrightarrow B : \mathcal{V} \\ A : \mathcal{I} \leftrightarrow B : \mathcal{I} \end{cases}. \quad (17)$$

Although the cross-modal/model gap challenges cluster alignment, two key strategies address it. First, common prior knowledge plays a crucial role. The backbones of both modalities/models are initialized with the same pre-trained weights, ensuring a shared feature representation that helps narrow the cross-modal gap from the start. Second, the robust loss in Equation (7) further strengthens alignment by guiding the model toward reliable correlations, prioritizing consistent clusters, and mitigating the impact of noisy correspondences.

E. Training and Inference Strategy

In a specific model A or B, given an input image $x_i^{\mathcal{P}}$, its training loss $\mathcal{L}_i^{\mathcal{P}}$ is actually a combination of intra-modal loss and inter-modal loss with a trade-off parameter λ. $\mathcal{L}_i^{\mathcal{P}}$ is defined as follow:

$$\mathcal{L}_i^{\mathcal{P}} = -\lambda p^{\gamma_i}\left((\widetilde{y}_i^{\mathcal{P}}|_Z)^\star | x_i^{\mathcal{P}}\right) - (1-\lambda) p^{\gamma_i}\left((\widetilde{y}_i^{\mathcal{Q}}|_Z)^\star | x_i^{\mathcal{P}}\right), \quad (18)$$

where Z refers to the model A or B, i.e., $Z \in \{A, B\}$. The overall loss $\mathcal{L}_{all}$ can be formulated as:

$$\begin{aligned} \mathcal{L}_{all} &= \sum_{\mathcal{P}\in\{\mathcal{V},\mathcal{I}\}} \sum_{i=1}^{N^{\mathcal{P}}} \mathcal{L}_i^{\mathcal{P}} \\ &= \lambda \underbrace{\sum_{\mathcal{P}\in\{\mathcal{V},\mathcal{I}\}} \sum_{i=1}^{N^{\mathcal{P}}} -p^{\gamma_i}\left((\widetilde{y}_i^{\mathcal{P}}|_Z)^\star | x_i^{\mathcal{P}}\right)}_{\mathcal{L}_{ra}^\alpha} \\ &\quad + (1-\lambda) \underbrace{\sum_{\mathcal{P}\in\{\mathcal{V},\mathcal{I}\}} \sum_{i=1}^{N^{\mathcal{P}}} -p^{\gamma_i}\left((\widetilde{y}_i^{\mathcal{Q}}|_Z)^\star | x_i^{\mathcal{P}}\right)}_{\mathcal{L}_{ra}^\beta}. \end{aligned} \quad (19)$$

Overall, the training process alternates between two models initialized with distinct parameters by minimizing the objective $\mathcal{L}_{all}$. Notably, each model uses pseudo-labels generated by the other model for optimization. The detailed optimization procedure of RoDE is given in Algorithm 1.

In the inference stage, we integrate the features of models A and B to obtain more comprehensive and robust representations, thus enhancing the representation capability and improving the performance. More specifically, given a query image $x_i^{\mathcal{P}}$, its corresponding joint feature is:

$$v_i^{\mathcal{P}} = \frac{1}{2}\left(f_A^{\mathcal{P}}(x_i^{\mathcal{P}}; \Theta_A^{\mathcal{P}}) + f_B^{\mathcal{P}}(x_i^{\mathcal{P}}; \Theta_B^{\mathcal{P}})\right). \quad (20)$$

In brief, the target of CCM could be formulated as a binary linear programming problem, which aims to match clusters with similar features. Specifically, we assume two sets of clustering center groups denoted as $C^P = \{c_i^P\}_{1 \le i \le N^P}$ and $C^Q = \{c_j^Q\}_{1 \le j \le N^Q}$, where c_i^P and c_j^Q represent the i-th and j-th cluster centers for the modalities or models P and Q respectively. N^P and N^Q are the numbers of clusters for P and Q, respectively. Based on this, we design a cost matrix $S = \{S_{ij}\}_{1 \le i \le N^P, 1 \le j \le N^Q}$, where S_{ij} satisfies:

$$S_{ij} = \exp\left(1 - \frac{c_i^P}{\|c_i^P\|}\left(\frac{c_j^Q}{\|c_j^Q\|}\right)^T\right). \quad (14)$$

Therefore, the objective function can be formulated as:

$$\begin{aligned} &\arg\min_{M} S^T M, \\ &s.t. \quad M\mathbf{1} = \mathbf{1}, \ M^T\mathbf{1} = \mathbf{1} \quad \text{and} \ \forall M_{ij} \in \{0, 1\}, \end{aligned} \quad (15)$$

where $\mathbf{1}$ is a column vector consisting entirely of ones, and the M serves as an indicator factor matrix, whose (i, j)-th element determines whether c_i^P and c_j^Q belong to the same class with M_{ij} equaling 1, and 0 otherwise. Intuitively, many existing binary linear matching methods could be utilized to solve the problem of Equation (15), such as maximum weight matching, bipartite matching, and linear sum assignment. However, these methods would fail if N^P and N^Q are not equal, since they may leave many clusters unmatched in this case. To address this issue, we advocate for aligning unmatched clusters through multiple dynamic matches until all clusters have been progressively matched.

To mitigate inter-cluster misalignment, we perform consistency matching across different modalities and models, respectively. Specifically, we first align cross-modal clusters (i.e. Equation (16)), ensuring that clusters from different

---

Algorithm 1 Optimization Procedure of RoDE

Input: Training dataset $\mathcal{X}$, models A and B, λ, τ, η, epochs N_e, batch size N_b, and learning rate lr.

1: Warm up A and B with Equation (4), respectively;
2: **for** $n_e \to \{1, 2, \cdots, N_e\}$ **do**
3:     Calculate the features $v_i^{\mathcal{P}} = f^{\mathcal{P}}(x_i^{\mathcal{P}}; \Theta^{\mathcal{P}})$ for each sample $x_i^{\mathcal{P}}$ using models A and B;
4:     Perform modality-specific clustering on the features $v_i^{\mathcal{P}}$ to acquire the centers for models A and B;
5:     **if** $n_e == 1$ **then**
6:         Initialize memory banks $\mathcal{M}_A^{\mathcal{P}}$ and $\mathcal{M}_B^{\mathcal{P}}$ using the clustering centers;
7:     **else**
8:         Update memory banks $\mathcal{M}_A^{\mathcal{P}}$ and $\mathcal{M}_B^{\mathcal{P}}$ with Equation (3);
9:         Align cluster centers across different modalities or models through Equation (16) and Equation (17).
10:         Generate model-specific pseudo-labels $(\widetilde{y}_i^{\mathcal{P}}|_A, \widetilde{y}_i^{\mathcal{Q}}|_A)$ and $(\widetilde{y}_i^{\mathcal{P}}|_B, \widetilde{y}_i^{\mathcal{Q}}|_B)$;
11:         Obtain pseudo-labels with cross cluster consistency, i.e. $((\widetilde{y}_i^{\mathcal{P}}|_A)^\star, (\widetilde{y}_i^{\mathcal{Q}}|_A)^\star)$ and $((\widetilde{y}_i^{\mathcal{P}}|_B)^\star, (\widetilde{y}_i^{\mathcal{Q}}|_B)^\star)$;
12:         **repeat**
13:         Randomly select N_b samples;
14:         Update parameters $\Theta_A^{\mathcal{P}}$ with Equation (19) using the pseudo-labels $((\widetilde{y}_i^{\mathcal{P}}|_B)^\star, (\widetilde{y}_i^{\mathcal{Q}}|_B)^\star)$;
15:         Update parameters $\Theta_B^{\mathcal{P}}$ with Equation (19) using the pseudo-labels $((\widetilde{y}_i^{\mathcal{P}}|_A)^\star, (\widetilde{y}_i^{\mathcal{Q}}|_A)^\star)$;
16:     **until** *All samples are selected;*
17:     **end if**
18: **end for**
Output: Optimized parameters $\Theta_A^{\mathcal{P}}$ and $\Theta_B^{\mathcal{P}}$.

Subsequently, we use the joint features $v_i^{\mathcal{P}}$ to identify the pedestrian image with the highest cross-modal similarity, thereby obtaining the re-identification results for $x_i^{\mathcal{P}}$.

IV. EXPERIMENTS

A. Datasets

We evaluate our proposed RoDE using three publicly available datasets:

SYSU-MM01 [45] is a large-scale visible-infrared person re-identification dataset with four visible and two near-infrared cameras, covering both indoor and outdoor settings. The training set includes 22,257 visible images and 11,909 infrared images of 395 identities. For single-shot evaluation, the query and gallery sets comprise 3,803 infrared images and 301 randomly selected visible images from 96 identities.

RegDB [46] is a smaller dataset with two aligned cameras (one visible and one thermal). The similarity in body pose and capture distance between modalities reduces the challenge of visible-infrared re-identification. Each identity is represented by 10 visible and 10 infrared images.

LLCM [46] is a low-light multimodal dataset with 9 cameras in dim environments, including 46,767 bounding boxes across 1,064 identities. Both modalities suffer from issues like blurring and pose variation.

B. Evaluation Metric

To ensure fair comparisons, we use established protocols to evaluate retrieval performance [9]. These include Cumulative Matching Characteristic (CMC), mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP). Following [14], we assess SYSU-MM01 dataset performance using both testing modes across 10 randomly selected gallery sets. For RegDB and LLCM, we consider two scenarios: Visible to Thermal (V2T) and Thermal to Visible (T2V). The training is performed entirely in an unsupervised manner, with identity labels used only during testing.

C. Experimental Settings

The experiments and evaluations of RoDE are conducted on four NVIDIA Tesla V100 GPUs with Ubuntu 18.04.6 OS using PyTorch. We utilize AGW [31] as the feature extractor for both visible and infrared modalities. All the input images are resized to 288×144 and then executed data augmentation, including random flipping, random erasing, and random cropping. In RoDE, the initial learning rate is set to 3.5×10^{-4} and decays by a factor of 10 every 25 epochs. We train the model for a total of 50 epochs. The batch size is 32, with a memory updating rate η of 0.15 and a temperature factor τ of 0.05. The trade-off parameter λ is analyzed in Section IV-E.

D. Comparison With the State-of-the-Art Methods

To evaluate the effectiveness of our RoDE, we compare it with 26 state-of-the-art methods across three benchmark datasets. These methods are grouped into two categories: 9 supervised VI-ReID (SVI-ReID) methods and 17 unsupervised VI-ReID (UVI-ReID) methods. From the results in Tables I and II, one can be drawn the following observations:

- *Comparison with UVI-ReID Methods:* Our RoDE achieves state-of-the-art performance on the three benchmark datasets in the unsupervised setting. To be specific, on the SYSU-MM01 dataset, RoDE achieves 62.88% Rank-1, 57.91% mAP, and 43.04% mINP in the All Search mode, and 64.53% Rank-1, 70.42% mAP, and 66.04% mINP in the Indoor-Search mode. On the RegDB dataset, RoDE shows significant advancements over the latest SCA-RCP [10], with a notable Rank-1 improvement of 3.18% in V2T and 3.37% in T2V. Moreover, on the more challenging LLCM dataset, RoDE demonstrates both outstanding and promising performance. Compared to the second-best methods, GUR [4], it achieves a 3.66% Rank-1 increase in the V2T setting and a 3.05% Rank-1 increase in the T2V setting, respectively. Overall, these observations highlight the immense potential of RoDE, particularly in real-world scenarios that require high accuracy and involve challenging conditions.

- *Comparison with Noisy-labels based UVI-ReID Methods:* Compared to existing methods, our RoDE systematically considered handling various noise issues raised by pseudo-labels. More importantly, RoDE reveals error accumulation in unsupervised cross-modal learning, which has been overlooked by previous noisy label
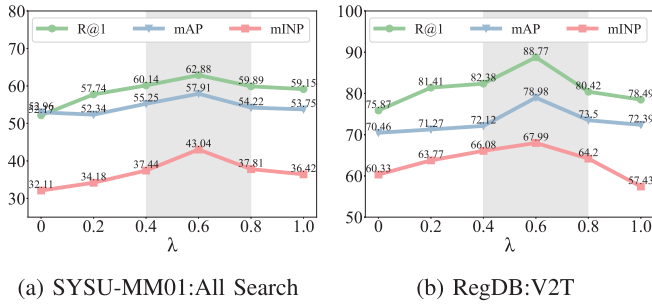
(a) SYSU-MM01:All Search   (b) RegDB:V2T

Fig. 5. The impact of parameter λ. The **gray shaded area** represents the recommended parameter range for further fine-tuning, as suggested by the authors.

learning based UVI-ReID methods such as CHCR [3], DOTLA [7], IMSL [5], and SCA-RCP [10]. From the results, one could find that our RoDE outperforms these methods in overall performance, demonstrating the effectiveness of RDL in tackling error accumulation.

- *Comparison with SVI-ReID Methods:* Our RoDE achieves comparable performance to or even surpasses some supervised methods, especially on the RegDB dataset, which demonstrates the superiority of RoDE in effectively extracting discrimination from unlabeled and unaligned VI-ReID data. However, UVI-ReID still faces challenges in achieving a more precise cross-modal semantic understanding, indicating potential space for improvement.

E. Parameter Analysis

As shown in Fig. 5, we analyze the performance variation of the model for different values of λ within the range of [0, 1]. The sensitivity of λ can vary depending on the characteristics of different datasets, but our results demonstrate that RoDE generally achieves the best performance when λ is around 0.6. Notably, when $\lambda = 0$ (i.e., without $\mathcal{L}_{ra}^{\alpha}$) or $\lambda = 1$ (i.e., without $\mathcal{L}_{ra}^{\beta}$), the model's performance significantly degrades. This highlights the critical role of balancing both discriminative learning and modality-invariant feature representation. The poor performance at the extremes of λ underscores the necessity of jointly optimizing these two objectives to effectively address the challenges posed by noisy and incomplete data.

F. Ablation Studies

To address the three challenging issues arising from noise in pseudo-labels (i.e, noise overfitting, error accumulation, and noisy cluster correspondence), we designed three components, RAL, RDL, and CCM, respectively. The ablation studies are conducted to validate the effectiveness of them.

1) Effectiveness of RAL: In this experiment, two variants are presented to study the effectiveness of our RAL: RoDE without $\mathcal{L}_{ra}^{\alpha}$ and RoDE without $\mathcal{L}_{ra}^{\beta}$, which is shown in Table III. The experimental results demonstrate that each component (i.e., $\mathcal{L}_{ra}^{\alpha}$ or $\mathcal{L}_{ra}^{\beta}$) contributes to the person re-identification performance. More specifically, RAL reduces over-optimization on low-reliability samples by using the robust loss function $\mathcal{L}_{ra}$, directly addressing noise interference. Furthermore, RAL adaptively ensures selective emphasis,

TABLE IV
ABLATION STUDIES ON RDL

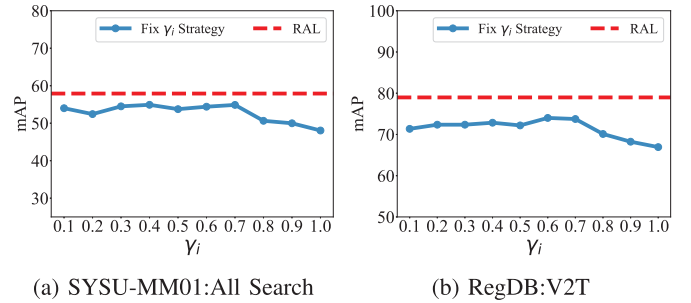|	SYSU-MM01: All Search			RegDB: V2T		
	Rank-1	mAP	mINP	Rank-1	mAP	mINP
A	56.81	52.77	35.92	70.41	65.77	57.92
B	56.66	53.02	36.88	71.12	66.02	58.88
A+B (A)	60.23	54.86	40.18	84.32	73.86	63.18
A+B (B)	59.97	54.27	37.44	83.97	73.27	63.44
A+B	**62.88**	**57.91**	**43.04**	**88.77**	**78.98**	**67.99**

(a) SYSU-MM01:All Search   (b) RegDB:V2T

Fig. 6. The impact of the parameter γ_i and the advantages of the adaptive strategy in RAL. The blue points indicate results with a fixed value of γ_i, while the red line represents the results of the RAL, which serves as the upper bound for the fixed γ_i strategy.

effectively minimizing noise while preserving valuable information in mislabeled samples. Notably, compared to the absence of $\mathcal{L}_{ra}^{\beta}$, the lack of $\mathcal{L}_{ra}^{\alpha}$ often results in worse results. This phenomenon occurs because CCM establishes preliminary associations between cross-modal clusters with the same identity, enabling learning potential modality-invariant representations.

2) Effectiveness of CCM: As shown in Table III, removing CCM leads to suboptimal model performance, especially in the absence of cross-model alignment. The main reason may be that the lack of consistent correspondence leads to discontinuous and inconsistent supervision. Without cross-model consistency, the model suffers from interference caused by semantically inconsistent pseudo-labels generated by other models, leading to poor performance (i.e., Order #3 in Table III). In contrast, inter-modalities can maintain a certain degree of relevance during training, even without cross-modal alignment. This internal relevance helps integrate information and mitigate the impact of mismatches, thus avoiding significant performance degradation (i.e., Order #4 in Table III). Therefore, cross-model mismatch is a crucial issue affecting model performance and even making the model invalid.

3) Effectiveness of RDL: We evaluate the effectiveness of RDL in capturing diverse information, as shown in Table IV. Our method, which jointly trains and tests both models $(A + B)$, outperforms independently trained models (A and B) and collaborative models where only one model is used for testing. Specifically, $A + B(A)$ refers to collaborative training with model A used for testing, and $A + B(B)$ refers to collaborative training with model B used for testing. In contrast, our approach allows both models to benefit from each
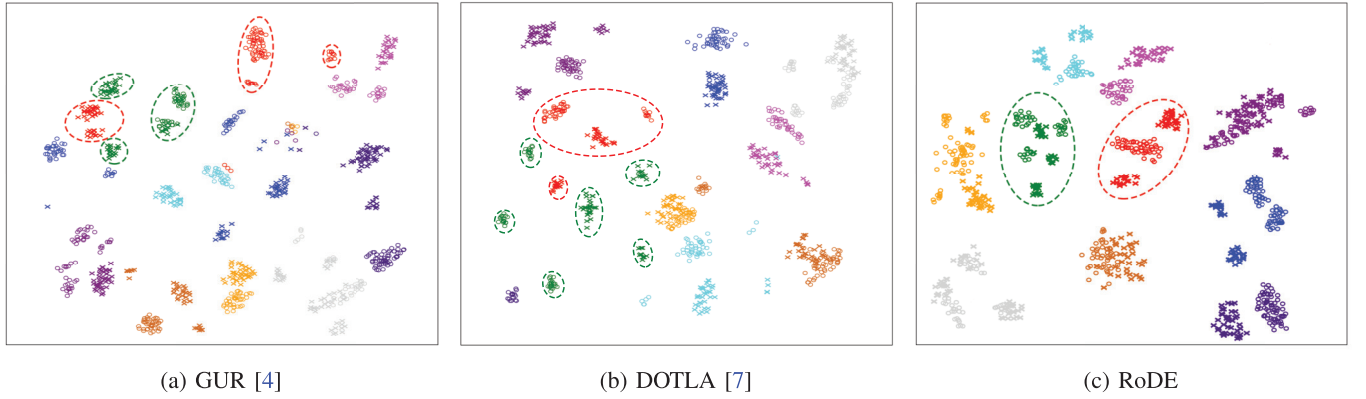
(a) GUR [4]   (b) DOTLA [7]   (c) RoDE

Fig. 7. The t-SNE plot for 10 randomly selected identities from SYSU-MM01 is presented, with ∘ representing visible modality and × representing infrared modality.

other's predictions during training and testing, reducing error accumulation and leveraging complementary information.

4) The Beneficial Effects of Self-Adaption Selecting γ_i: We conducted a series of experiments to demonstrate the advantages of the adaptive optimization strategy in RAL. Specifically, we compared RAL with strategies using fixed parameter γ_i, where γ_i ranges in $(0, 1]$, which is shown in Fig. 6. From the figure, one can see that the result of RAL serves as an upper bound (the red line) compared to the results achieved with various fixed parameters γ_i, demonstrating that the fixed parameter γ_i limits the ability of adaptive optimization for the clean and noisy samples. In other words, RAL effectively mitigates performance degradation by adaptively reducing the emphasis on mislabeled samples.

G. Visualization Analysis

We conduct a detailed visualization comparing RoDE with the most competitive baselines GUR [4] and DOTLA [7].

1) T-SNE Visualization: We plot the t-SNE map feature distribution of 10 randomly selected identities from SYSU-MM01 in Fig. 7. We observe that GUR [4] and DOTLA [7] fail to come together pedestrian images with the same identities. For example, in Fig. 7 (a) and (b), the samples marked in red and green do not flock together within the same dashed ellipse. This issue likely arises from their inability to robustly handle noise interference in pseudo-labels, which results in the distribution of some samples in the feature space being shifted by label noise. In contrast, RoDE (i.e., Fig. 7 (c)) demonstrates robustness against pseudo-label noise, producing a consistent understanding of images with the same identity in the feature space even under noisy label conditions.

2) Visualization on the Qualitative Results: We further evaluated the qualitative results of RoDE with the benchmark methods. The qualitative results are presented by retrieving the Top-5 gallery images with the highest similarity scores for each query image, as shown in Fig. 8. RoDE provides more stable matching results compared to other competitive methods. Notably, RoDE selects accurate results even for the challenging query image with severe occlusion (i.e., the first-row case), demonstrating its ability to handle complex scenarios. However, RoDE faces inevitable failure when the

(a) GUR [4]   (b) DOTLA [7]   (c) RoDE

Fig. 8. Some person re-identification results of (a) GUR [4], (b) DOTLA [7], and (c) RoDE. Each row presents a query image of a person (marked with an **orange** bounding box) on the left, with the retrieved images highlighted in **green** bounding boxes denoting correct matches and those in **red** indicating incorrect matches. The results are arranged in descending order. The first four cases show successful outcomes, while the last two represent failures.

query image is severely blurred and unclear (i.e., the third-row and fourth-row case). This is because, in cases of significant image degradation, the blurred visual information makes it difficult for the model to extract effective features, thereby affecting the matching accuracy.

H. Robustness Analysis

To verify the robustness of the proposed RoDE against pseudo-label noise (PLN), we conduct detailed experiments and analyses focusing on three challenges: noisy overfitting, error accumulation, and noisy cluster correspondence. While these three types of noise are interrelated and can all significantly impact model performance, no single type can be considered more critical than the others.

1) Robustness Analysis on Noisy Overfitting: To illustrate the issue of noisy overfitting, we analyze the sample loss

(a) without RAL   (b) with RAL

(c) Single Model   (d) Dual Models

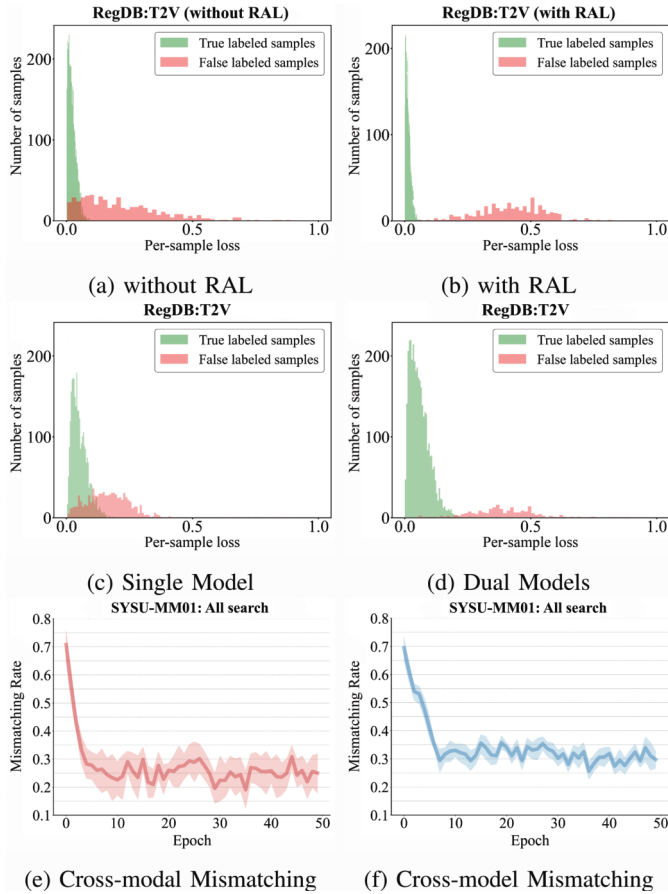(e) Cross-modal Mismatching   (f) Cross-model Mismatching

Fig. 9. Robustness analysis of RoDE.

distribution, as illustrated in Fig. 9 (a) and (b). When RAL is absent, many clean samples and noisy labeled samples appear simultaneously near the low loss value, indicating severe overfitting to the noisy samples. In contrast, with the introduction of RAL, the loss distribution exhibits a clear separation between clean and noisy samples. This improvement is because RAL can reduce the attention to noisy samples through adaptive optimization, thus preventing the training process from being dominated by pseudo-label noise.

2) Robustness Analysis on Error Accumulation: The mutual learning mechanism in the RDL strategy reduces error accumulation by alternating training between two models and sharing pseudo-labels. It also prevents both models from becoming overly biased towards incorrect pseudo-labels, mitigating error accumulation. This alternating process ensures that both models benefit from each other's predictions, improving generalization and reducing susceptibility to errors from relying on a single model. Moreover, the RDL framework uses a dynamic weighting mechanism RAL to adjust the influence of pseudo-labels based on their reliability, ensuring that errors do not dominate the learning process. To display error accumulation, we have counted the loss distributions for each sample in the infrared modality. These distributions are obtained through training a single model (i.e., Fig. 9 (c)) and dual models (i.e., Fig. 9 (d)), respectively. In the case of single models, noisy and clean samples intermingle due to

significant error accumulation, as evidenced by the overlapping colored areas. This overlap indicates that the model struggles to differentiate between noisy and clean samples, leading to imprecise predictions. In contrast, dual models with RoDE effectively alleviate this issue, producing a clear separation between noisy and clean samples.

3) Robustness Analysis on Noisy Cluster Correspondence: We observe that both cross-model and cross-modal cluster mismatches decrease progressively during training and eventually stabilize within a certain range, as illustrated in Fig. 9 (e) and (f). The shaded area indicates the standard deviation. This demonstrates that CCM can mitigate biases introduced by noise or inconsistent correspondences, helping the model maintain a more accurate optimization path. Unfortunately, while CCM reduces most of the matching errors, a mismatching rate of 15% to 35% still persists, which underscores the need for further refinement in the matching process.

V. CONCLUSION AND FUTURE WORKS

This paper proposes a novel learning paradigm, RoDE, for UVI-ReID that simultaneously addresses three key challenges: noisy overfitting, error accumulation, and noisy cluster correspondence. To mitigate noisy overfitting, RoDE employs a pivotal RAL to dynamically and adaptively reduce the emphasis on noisy samples. It also alternates training between two individual models, thereby maintaining diversity and avoiding error accumulation. Additionally, RoDE incorporates the CCM to establish reliable alignment across distinct modalities and different models by leveraging cross-cluster similarities. Numerous experiments demonstrate the excellent performance of our proposed method. In the future, we plan to extend RoDE to tackle additional challenges in VI-ReID, particularly noise filtering techniques and domain adaptation, to better handle the variability of real-world scenarios.

REFERENCES

[1] Z. Cui, J. Zhou, and Y. Peng, "DMA: Dual modality-aware alignment for visible-infrared person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 2696–2708, 2024.

[2] Z. Wu and M. Ye, "Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 9548–9558.

[3] Z. Pang, C. Wang, L. Zhao, Y. Liu, and G. Sharma, "Cross-modality hierarchical clustering and refinement for unsupervised visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2706–2718, Apr. 2024.

[4] B. Yang, J. Chen, and M. Ye, "Towards grand unified representation learning for unsupervised visible-infrared person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 11035–11045.

[5] B. Yang, M. Ye, J. Chen, and Z. Wu, "Augmented dual-contrastive aggregation learning for unsupervised visible-infrared person re-identification," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 2843–2851.

[6] D. Cheng, X. Huang, N. Wang, L. He, Z. Li, and X. Gao, "Unsupervised visible-infrared person ReID by collaborative learning with neighbor-guided label refinement," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 7085–7093.

[7] J. Shi et al., "Dual pseudo-labels interactive self-training for semi-supervised visible-infrared person re-identification," in *Proc. IEEE Int. Conf. Compute. Vis. (ICCV)*, Oct. 2023, pp. 11218–11228.

[8] M. Yang, Z. Huang, P. Hu, T. Li, J. Lv, and X. Peng, "Learning with twin noisy labels for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Oct. 2022, pp. 14308–14317.

[9] Z. Pang, L. Zhao, Y. Liu, G. Sharma, and C. Wang, "Inter-modality similarity learning for unsupervised multi-modality person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 10, pp. 10411–10423, Oct. 2024.

[10] Z. Li, H. Liu, X. Peng, and W. Jiang, "Inter-intra modality knowledge learning and clustering noise alleviation for unsupervised visible-infrared person re-identification," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 8, pp. 3934–3947, Aug. 2024.

[11] Y. Sun, J. Dai, Z. Ren, Y. Chen, D. Peng, and P. Hu, "Dual self-paced cross-modal hashing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, Mar. 2024, pp. 15184–15192.

[12] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 407–419, 2019.

[13] Y. Hao, N. Wang, J. Li, and X. Gao, "HSME: Hypersphere manifold embedding for visible thermal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 8385–8392.

[14] Q. Wu et al., "Discover cross-modality nuances for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4330–4339.

[15] Z. Chai, Y. Ling, Z. Luo, D. Lin, M. Jiang, and S. Li, "Dual-stream transformer with distribution alignment for visible-infrared person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6764–6776, Nov. 2023.

[16] J. Wu, H. Liu, W. Shi, M. Liu, and W. Li, "Style-agnostic representation learning for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 26, pp. 2263–2275, 2023.

[17] J. Shi et al., "Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification," in *Proc. 38th Annu. Conf. Neural Inf. Process. Syst.*, 2024, pp. 1–14.

[18] J. Shi et al., "Multi-memory matching for unsupervised visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Sep. 2024, pp. 456–474.

[19] J. Wang et al., "Optimal transport for label-efficient visible-infrared person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV).* Cham, Switzerland: Springer, 2022, pp. 93–109.

[20] Y. Bai et al., "RaSa: Relation and sensitivity aware representation learning for text-based person search," 2023, *arXiv:2305.13653*.

[21] M. Cao, M. Y. Bai, Z. Zeng, M. Ye, and M. Zhang, "An empirical study of clip for text-based person search," in *Proc. AAAI Con. Artif. Intell.*, vol. 38, 2024, pp. 465–473.

[22] M. Cao, S. Li, J. Li, L. Nie, and M. Zhang, "Image-text retrieval: A survey on recent research and development," 2022, *arXiv:2203.14713*.

[23] Y. Li, Y. Qin, Y. Sun, D. Peng, X. Peng, and P. Hu, "RoMo: Robust unsupervised multimodal learning with noisy pseudo labels," *IEEE Trans. Image Process.*, vol. 33, pp. 5086–5097, 2024.

[24] Y. Feng, H. Zhu, D. Peng, X. Peng, and P. Hu, "RONO: Robust discriminative learning with noisy labels for 2D-3D cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 11610–11619.

[25] Y. Sun, Y. Qin, Y. Li, D. Peng, X. Peng, and P. Hu, "Robust multi-view clustering with noisy correspondence," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 12, pp. 9150–9162, Dec. 2024.

[26] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, and X. Peng, "Learning with noisy correspondence for cross-modal matching," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 29406–29419.

[27] Y. Yao et al., "Jo-SRC: A contrastive approach for combating noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5188–5197.

[28] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[29] P. Hu, X. Peng, H. Zhu, L. Zhen, and J. Lin, "Learning cross-modal retrieval with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5403–5413.

[30] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96, Aug. 1996, pp. 226–231.

[31] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[32] M. Ye, W. Ruan, B. Du, and M. Z. Shou, "Channel augmented joint learning for visible-infrared recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 13567–13576.

[33] C. Chen, M. Ye, M. Qi, J. Wu, J. Jiang, and C.-W. Lin, "Structure-aware positional transformer for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 31, pp. 2352–2364, 2022.

[34] T. Liang, Y. Jin, W. Liu, and Y. Li, "Cross-modality transformer with modality mining for visible-infrared person re-identification," *IEEE Trans. Multimedia*, vol. 25, pp. 8432–8444, 2023.

[35] H. Lu, X. Zou, and P. Zhang, "Learning progressive modality-shared transformers for effective visible-infrared person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 1835–1843.

[36] Y. Ge, F. Zhu, D. Chen, and R. Zhao, "Self-paced contrastive learning with hybrid memory for domain adaptive object Re-ID," in *Proc. NIPS*, vol. 33, 2020, pp. 11309–11321.

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

[38] S. Xuan and S. Zhang, "Intra-inter camera similarity for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11926–11935.

[39] M. Wang, B. Lai, J. Huang, X. Gong, and X.-S. Hua, "Camera-aware proxies for unsupervised person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 4, pp. 2764–2772.

[40] W. Liang, G. Wang, J. Lai, and X. Xie, "Homogeneous-to-heterogeneous: Unsupervised learning for RGB-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 6392–6407, 2021.

[41] Y. Cho, W. J. Kim, S. Hong, and S.-E. Yoon, "Part-based pseudo label refinement for unsupervised person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7308–7318.

[42] Z. Chen, Z. Zhang, X. Tan, Y. Qu, and Y. Xie, "Unveiling the power of CLIP in unsupervised visible-infrared person re-identification," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 3667–3675.

[43] T. Si, F. He, P. Li, Y. Song, and L. Fan, "Diversity feature constraint based on heterogeneous data for unsupervised person re-identification," *Inf. Process. Manage.*, vol. 60, no. 3, May 2023, Art. no. 103304.

[44] B. Yang, J. Chen, X. Ma, and M. Ye, "Translation, association and augmentation: Learning cross-modality re-identification from single-modality annotation," *IEEE Trans. Image Process.*, vol. 32, pp. 5099–5113, 2023.

[45] A. Wu, W.-S. Zheng, H.-X. Yu, S. Gong, and J. Lai, "RGB-infrared cross-modality person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5380–5389.

[46] D. Nguyen, H. Hong, K. Kim, and K. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, Mar. 2017.

[47] H. Park, S. Lee, J. Lee, and B. Ham, "Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12046–12055.

[48] Y. Zhang, Y. Yan, Y. Lu, and H. Wang, "Towards a unified middle modality learning for visible-infrared person re-identification," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 788–796.

[49] Y. Zhang and H. Wang, "Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2153–2162.

[50] J. Han, Y.-L. Li, and S. Wang, "Delving into probabilistic uncertainty for unsupervised domain adaptive person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 790–798.