Dual-Teacher Interactive Knowledge Distillation Network for Text-to-Visible & Infrared Person Retrieval

Chenglong Li[1,2], **Zhengyu Chen**[1,2], **Yifei Deng**[3], **Aihua Zheng**[1,2,*]
[1]School of Artificial Intelligence, Anhui University, Hefei 230601, China
[2]Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University
[3]School of Computer Science and Technology, Anhui University, Hefei 230601, China
lcl1314@foxmail.com,wa24301145@stu.ahu.edu.cn,yf-ah@foxmail.com,ahzheng214@foxmail.com

Abstract

Text-to-visible & infrared person retrieval aims to retrieve the corresponding visible (RGB) and thermal infrared (TIR) images given the text descriptions. Existing methods perform semantic decoupling by aligning RGB and TIR features separately to different attributes, thereby facilitating the alignment between the fused multimodal representation and the text. However, insufficient TIR representation ability and cross-view representation capabilities of RGB and TIR modalities limit the retrieval accuracy and robustness. To address these issues, we propose a novel Dual-teacher Interactive Knowledge Distillation Network called DIKDNet, which performs the interactive knowledge distillation between two modality-specific teachers with rich cross-view representation capabilities to enhance TIR representations and the collaborative knowledge distillation from both teachers to the corresponding students to enhance the cross-modal cross-view representations, for robust text-to-visible & infrared person retrieval. Specifically, to enhance the representation ability of the TIR backbone network while preserving modality-specific characteristics, we design an Interactive Knowledge Distillation Module (IKDM), which introduces a boundary-constrained distillation strategy between RGB and TIR backbones, to transfer the semantic features of RGB backbone to TIR one. To enhance the cross-modal cross-view representation capability, we design a Collaborative Knowledge Distillation Module (CKDM) to transfer the cross-modal similarity relations and the cross-view multimodal representations from teacher networks to student ones. Experimental results demonstrate that our method consistently achieves significant performance gains on both the RGBT-PEDES and RGBNT201-PEDES datasets. The code will be released upon the acceptance.

Introduction

Person Re-identification (Wang et al. 2024; Deng et al. 2024) is a key computer vision task, but it heavily relies on the prior acquisition of target pedestrians' images for retrieval. The subsequently emerged text-to-image person retrieval (Chen, Xu, and Luo 2018), as a cross-modal retrieval task (Lei et al. 2022; Miech et al. 2021), can retrieve target pedestrians from image databases using natural language descriptions as queries. This traditional text-to-image person retrieval method (Chen et al. 2021; Chen, Xu, and Luo 2018) mainly focuses on mapping visible-light images and texts into a unified embedding space, and then improves retrieval performance through a combination of global feature matching and fine-grained alignment (Deng et al. 2025a; Zuo et al. 2024b; Deng et al. 2025b). However, these methods suffer from a sharp drop in performance under complex lighting conditions such as low light and nighttime, due to the loss of information in visible-light images. In contrast, TIR images capture long-wave infrared radiation (8-14 μm) (Ha et al. 2017; Ring and Ammer 2012) emitted by objects themselves, which can stably preserve pedestrian contours and structural features in low-light environments, providing a new solution for all-weather retrieval.

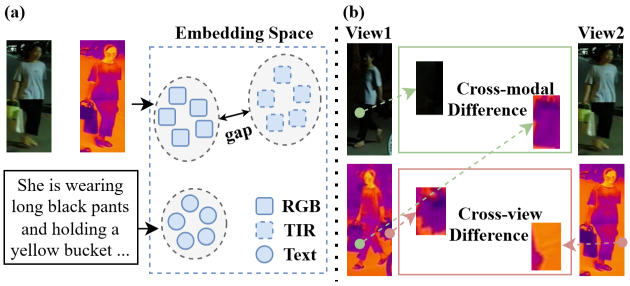

Figure 1: (a) The modal gap between RGB and TIR means pre-trained encoders cannot fully adapt to TIR characteristics, leading to insufficient TIR representation. (b) Under varying lighting and angles, cross-modal and cross-view differences exist for the same person. Failing to model these impairs retrieval accuracy.

Building on the above advantages, (Deng et al. 2026) are the first to propose the Text-to-visible & Infrared Person Retrieval task. This task uses textual descriptions as queries to retrieve paired RGB and TIR person images. By fusing the complementary information of the two modalities, it achieves robust retrieval under complex lighting conditions. Through a modality-specific text masking mechanism, they categorize textual attributes into color-related, color-unrelated, and other categories, which are aligned with RGB and TIR features respectively. Then, the retrieval is completed through global-local matching between fused features

*Corresponding author
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

and text. However, it still has significant limitations. Currently, encoders are mainly pre-trained on visible-light data, facing an inherent modal gap between RGB and TIR (Fig. 1 (a)) that makes them incompatible with TIR characteristics. Meanwhile, cross-modal and cross-view differences (Fig. 1 (b)) affect the task's ability to model feature associations of the same identity. These two challenges jointly impair retrieval performance and robustness.

To address the above issues, this paper proposes a Dual-Teacher Interactive Knowledge Distillation Network (DIKDNet). It improves the quality of TIR features and the capability of cross-modal and cross-view representation through interactive distillation between teacher networks and collaborative distillation between teachers and students. Specifically, in the dual-teacher network architecture, we design an Interactive Knowledge Distillation Module (IKDM) to enhance the representation ability of the TIR modality. Considering the advantages of the RGB modality in semantic information such as texture and color, as well as the structural preservation characteristics of the TIR modality in low-light scenarios, this module realizes knowledge interaction between the RGB and TIR teacher networks through a boundary-constrained distillation strategy. On the one hand, it transfers the ability of semantic feature extraction from the RGB teacher to the TIR teacher, making up for the deficiency of TIR in detailed representation. On the other hand, by constraining the distance boundary between RGB and TIR features, it prevents the TIR modality from losing its unique core attributes such as contours and structures. This controllable knowledge transfer mechanism not only improves the semantic richness of TIR features but also preserves their modality specificity.

To further enhance the model's cross-modal and cross-view representation capabilities, we design a Collaborative Knowledge Distillation Module (CKDM) to enable knowledge transfer from the dual-teacher network to the student network. Notably, during training, the teacher network can build associative knowledge using multi-view RGBT images and corresponding texts of the same identity (Yan et al. 2023). However, in inference, only single-view query texts and retrieval images are available. Thus, it is necessary to distill the teacher's associative modeling ability into the student, ensuring robust performance under single-view input. Specifically, the dual-teacher network learns cross-modal similarity relationships. By constraining the student network to have a similarity distribution consistent with that of the teacher, the accuracy of its cross-modal semantic alignment is improved. Meanwhile, the cross-view multi-modal representations constructed by the dual-teacher guide the student network to learn robust feature associations across views, enabling it to maintain retrieval consistency when facing view changes. Through the collaborative transfer of these two aspects, the student network inherits the cross-modal modeling ability and cross-view robustness of the teacher.

The contributions of this work are summarized as follows:

- We propose DIKDNet, which adopts an online distillation paradigm without requiring pre-trained teacher models, enabling knowledge interaction between modality-specific teachers and collaborative transfer to students.
- We design IKDM to enhance TIR feature representation while preserving its modality-specific attributes through a boundary-constrained distillation strategy.
- We construct CKDM to enhance the student model's cross-modal and cross-view collaborative representation capabilities via cross-view multimodal representation transfer and cross-modal similarity relationship transfer.
- We verify the superiority of the proposed method through experiments on the RGBT-PEDES and RGBNT201-PEDES datasets, with performance significantly outperforming existing state-of-the-art technologies.

Related Work

Text-Image Person Retrieval

The text-to-image person retrieval task, proposed in (Li et al. 2017b), aims to retrieve the target person's image from a dataset using textual descriptions by aligning features between text and images. Its core lies in mapping image and text features into a unified space and strengthening the associations between corresponding pairs. In feature extraction, the backbone has evolved from early independent encoders (VGG (Li et al. 2017a; Simonyan and Zisserman 2014), LSTM (Hochreiter and Schmidhuber 1997)) to ResNet (He et al. 2016) and BERT (Devlin et al. 2019), and now commonly uses CLIP's dual-stream encoder (Liu et al. 2024; Radford et al. 2021). CLIP excels at pre-trained cross-modal matching but lacks TIR adaptation, leading to suboptimal TIR feature extraction. Meanwhile, in cross-modal alignment, early methods focused on global feature matching (Chen et al. 2021; Chen, Xu, and Luo 2018) but missed fine-grained details, followed by explicit local alignments (Chen et al. 2022; Zhu et al. 2021) targeting region correspondences. Most recently, implicit alignment methods (Farooq et al. 2022; Jiang and Ye 2023) have emerged to capture cross-modal relationships without explicit region-level supervision. Despite these advancements in accuracy and efficiency, nearly all existing methods rely heavily on visible-light images. This dependence leads to a sharp performance drop in low-light conditions, where visible-light image information is severely degraded or lost. To address the limitation in low-light conditions, (Deng et al. 2026) are the first to propose the text-to-visible & infrared person retrieval task. They improve robustness by decoupling the alignment between text and RGBT, yet overlook the modeling of cross-modal and cross-view associations. For this reason, we propose DIKDNet, which optimizes CLIP's TIR feature extraction and strengthens cross-modal and cross-view collaborative representation capabilities to meet retrieval needs in complex scenarios.

Knowledge Distillation

Knowledge distillation (Hinton, Vinyals, and Dean 2015) is a model compression and cross-network knowledge transfer technology that aims to transfer knowledge from large or complex teacher models to small, simple student models. It has been widely applied in tasks such as visual recognition (Huang et al. 2024; Yang et al. 2022) and multi-

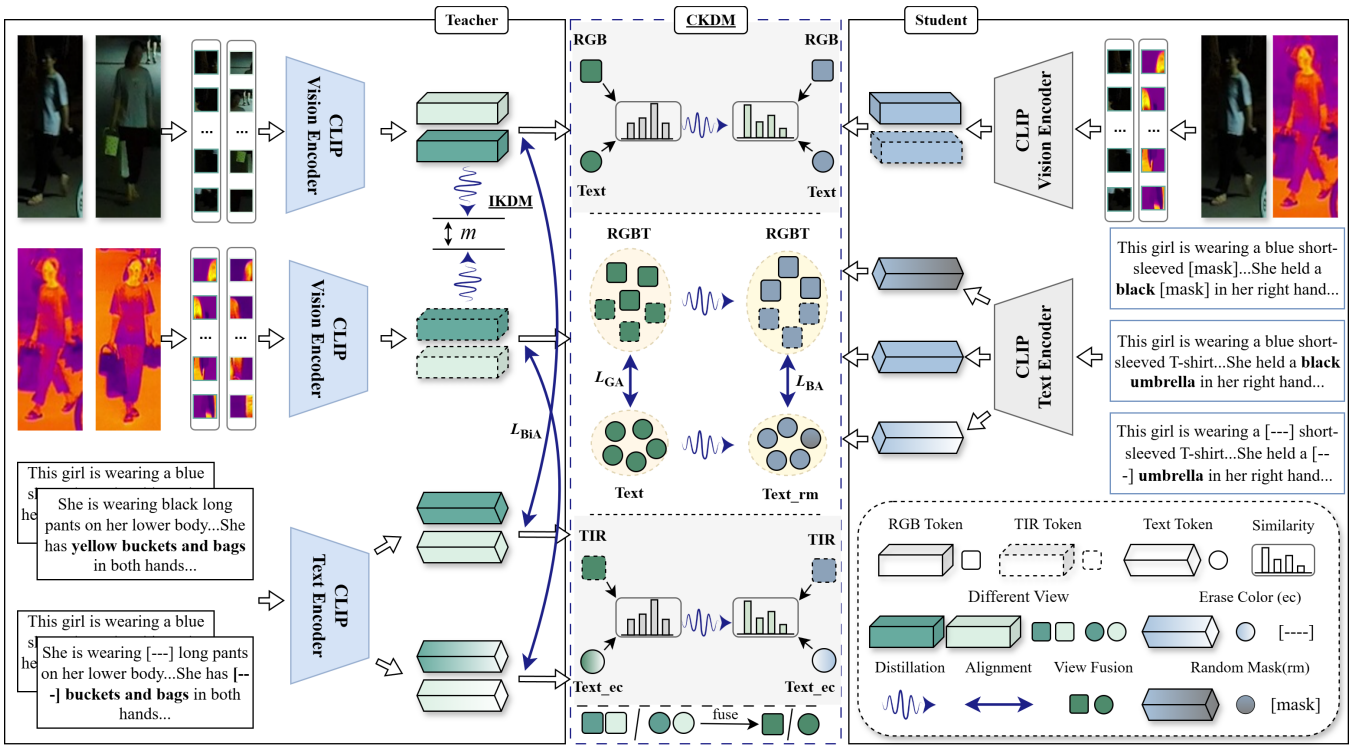

Figure 2: The dual-teacher interactive distillation network framework is an online distillation architecture. IKDM aims to force the TIR branch output to mimic the features of the visible-light branch output, and enhance its infrared extraction capability while preserving discriminative features by controlling the minimum distance between features. CKDM collaboratively transfers cross-view multimodal representations and cross-modal similarity relationships, which enhances the network's ability for multimodal and multi-view collaborative representation and further improves retrieval performance.

modal representation learning (Li et al. 2024). In early studies, (Hinton, Vinyals, and Dean 2015) achieve knowledge transfer using KL divergence, while (Romero et al. 2014) do so using mean squared error. With the development of multimodal tasks, researchers have begun to explore multimodal knowledge distillation. Some works (Yang et al. 2024) transfer knowledge from CLIP to downstream tasks through methods such as feature distillation and relation distillation. In cross-modal distillation, works like (Gupta, Hoffman, and Malik 2016; Tian, Krishnan, and Isola 2019) have achieved distillation from RGB to depth images by using RGB texture knowledge to guide depth feature extraction, while (Zha et al. 2022) treat RGB as a teacher for TIR target tracking tasks, improving TIR's target representation ability. These efforts collectively verify the value of RGB knowledge in enhancing the feature representation of other modalities. In contrast, this paper realizes cross-modal distillation through interactive distillation with a boundary constraint strategy, enabling the TIR branch to retain its unique structural attributes while receiving semantic knowledge from RGB. Then, through the collaborative knowledge distillation module, the cross-modal similarity relations and cross-view multimodal representations learned by the dual teachers are transferred to the student network, allowing it to maintain robust cross-modal and cross-view alignment even during single-view inference.

Dual-Teacher Interactive Knowledge Distillation Network

Overview

As illustrated in Fig. 2, DIKDNet is an online distillation framework designed to enhance the one-to-one retrieval performance of text-to-visible & infrared person by transferring knowledge from a teacher network to a lightweight student network. Both the teacher and student networks use CLIP's pre-trained models for image and text feature extraction, leveraging its mature encoding capabilities for visible images and text. The teacher network is further equipped with dedicated RGB and TIR image encoders, a text encoder, and a view fusion module, focusing on learning cross-modal and cross-view collaborative representations. Specifically, it leverages multi-view RGBT images and corresponding text to model associations between modalities and across views. In contrast, the student network adopts a shared image encoder and inherits the core capabilities of the teacher through collaborative distillation, maintaining performance while ensuring inference efficiency.

Dual-Teacher

To enhance cross-modal and cross-view collaborative representation, the teacher network takes multi-view RGBT images and text (including original text and color-erased text)

as input. After feature extraction by encoders, features from different views of the same modality are integrated to obtain F^f_{rgb}, F^f_{tir}, F^f_{txt}, and F^f_{txtnc}. Specifically, for image feature fusion, the image feature is first added to the image feature from a different view of the same ID, and then processed through a 4-layer Transformer encoder to generate the fused image feature F^f_{img}. For text feature fusion, simply sum multi-view text features to obtain the fused text feature F^f_{txt}. To enhance the dual teachers' ability to model associations between modalities and across views, we design bidirectional alignment and global alignment. For cross-modal alignment, we adopt the SDM loss (Jiang and Ye 2023), which enforces bidirectional consistency between image and text features. Using the SDM loss, we perform bidirectional alignment between fused and unfused features to constrain the view fusion effect. The bidirectional alignment loss is defined as:

$$\mathcal{L}_{\text{BiA}} = \mathcal{L}_{\text{rgb-txt}} + \mathcal{L}_{\text{tir-txtnc}}, \tag{1}$$

where $\mathcal{L}_{\text{rgb-txt}}$ is computed as the sum of two contrastive losses using the SDM loss: one between the fused RGB feature F^f_{rgb} and the unfused text feature F_{txt}, and the other between the unfused RGB feature F_{rgb} and the fused text feature F^f_{txt}. Similarly, $\mathcal{L}_{\text{tir-txtnc}}$ is constructed following the same mechanism to enforce alignment between the TIR and color-erased text modalities.

On this basis, global alignment is designed to reduce the distance between multi-view fused RGBT comprehensive features and text features, which is realized by strengthening associations at a higher level. The multimodal fused feature (the sum of F^f_{rgb} and F^f_{tir}, denoted as F^f_{rgbt}) is aligned with F^f_{txt} via the SDM loss:

$$\mathcal{L}_{\text{GA}} = \mathcal{L}_{\text{sdm}}\left(F^f_{\text{rgbt}}, F^f_{\text{txt}}\right). \tag{2}$$

The overall optimization function of the Dual-Teacher Network integrates the IKDM, bidirectional alignment, and global alignment losses to comprehensively learn cross-modal semantic associations, cross-view feature invariance, and multimodal global consistency:

$$\mathcal{L}_{\text{Dual-Teacher}} = \mathcal{L}_{\text{IKDM}} + \mathcal{L}_{\text{BiA}} + \mathcal{L}_{\text{GA}}. \tag{3}$$

Collaborative Knowledge Distillation Module

While teacher networks can build associative knowledge using multi-view RGBT images and texts of the same identity during training, only single-view query texts and retrieval images are available in inference. Thus, we transfer cross-modal and cross-view collaborative representation capabilities from teachers to students via the Collaborative Knowledge Distillation Module. CKDM transfers core knowledge from the teachers to the student through two pathways: cross-view multimodal representation transfer and cross-modal similarity relationship transfer.

Cross-View multimodal Representation Transfer (CVRT). This approach focuses on transmitting the discriminative feature patterns learned by the teachers. The teacher network has already mastered RGB texture features, TIR structural features, and their fused representations during multi-view training. By forcing the student to mimic the teachers' output features, this transfer ensures the student inherits the teachers' feature representation capability. Specifically, it minimizes the mean squared error (MSE) between the student's and teachers' features:

$$\mathcal{L}_{\text{CVRT}} = \mathcal{L}_{\text{mse}}\left(F^s_{\text{rgbt}}, F^f_{\text{rgbt}}\right) + \mathcal{L}_{\text{mse}}\left(F^s_{\text{txtmask}}, F^f_{\text{txt}}\right), \tag{4}$$

where F^f_{rgbt} denotes the fused RGBT feature of the teacher network, and F^f_{txt} denotes the teacher's text feature.

Cross-Modal Similarity relationship Transfer (CMST). This mechanism is dedicated to transmitting the semantic association logic modeled by the teachers. The teacher network has formed stable cross-modal alignment patterns through multi-view training. By aligning the student's and teachers' cross-modal similarity distributions, this transfer ensures the student inherits the teachers' cross-modal association logic. Specifically, we compute similarity distributions for two critical cross-modal pairs from both the teacher and student networks, where the similarity metric $\text{sim}(\cdot,\cdot)$ is defined as cosine similarity. For RGB-text pairs, the teacher network's similarity is defined as $S_1 = \text{sim}(F^f_{\text{rgb}}, F^f_{\text{txt}})$ and the student network's counterpart is $S_2 = \text{sim}(F^s_{\text{rgb}}, F^s_{\text{txt}})$. For TIR-color-erased text pairs, the teacher's similarity is $S_3 = \text{sim}(F^f_{\text{tir}}, F^f_{\text{txtnc}})$ while the student's is $S_4 = \text{sim}(F^s_{\text{tir}}, F^s_{\text{txtnc}})$. The cross-modal similarity relationship transfer loss is defined by minimizing the MSE between these distributions:

$$\mathcal{L}_{\text{CMST}} = \|S_1 - S_2\|_2^2 + \|S_3 - S_4\|_2^2. \tag{5}$$

This loss ensures the student learns bidirectional associations between RGB-text and TIR-textnc, strengthening cross-modal and cross-view semantic consistency. The total distillation loss of CKDM integrates the two transfer pathways:

$$\mathcal{L}_{\text{CKDM}} = \mathcal{L}_{\text{CVRT}} + \mathcal{L}_{\text{CMST}}. \tag{6}$$

Moreover, to ensure the student network's basic cross-modal alignment capability, we design a basic alignment as follows: RGB and TIR images are fed into the shared encoder to obtain features, F^s_{rgb} and F^s_{tir}, which are then fused into a unified image representation F^s_{rgbt}. For text processing, random masking is adopted as an augmentation method, and the processed text is encoded into a text representation F^s_{txtrm} via the text encoder. The fused image representation and text representation achieve basic cross-modal alignment through the SDM loss. Thus, the basic alignment loss of the student network is defined as:

$$\mathcal{L}_{\text{BA}} = \mathcal{L}_{\text{sdm}}\left(F^s_{\text{rgbt}}, F^s_{\text{txtrm}}\right). \tag{7}$$

Interactive Knowledge Distillation Module

To further enhance the representation capability of the visual encoder for the TIR modality while preserving its modality-specific attributes, we design an Interactive Knowledge Distillation Module (IKDM), which introduces a boundary-constrained distillation strategy between two modality-

specific teacher networks, effectively improving the representation of TIR modality information. Considering the significant cross-modal domain gap between RGB and TIR, directly forcing the TIR modality to fully mimic the output of the RGB teacher network may lead to overfitting and the loss of modality-specific information. Thus, we introduce a boundary constraint parameter m in the distillation process, which allows TIR features to moderately inherit RGB semantic representations while retaining their inherent modality-specific characteristics. Specifically, given the feature outputs of the RGB teacher Frgb and the TIR teacher Ftir, the proposed Interactive Knowledge Distillation loss is defined as: The IKDM loss is defined as:

$$\mathcal{L}_{\text{IKDM}} = \left\lfloor m - \|F_{\text{rgb}} - F_{\text{tir}}\|_2^2 \right\rfloor_+, \tag{8}$$

where $\lfloor \cdot \rfloor_+$ represents non-negative truncation, and m is a hyperparameter controlling the upper bound of feature similarity. This balances semantic consistency with RGB and the uniqueness of the TIR modality.

Experiments

In the following section, we first present the datasets, evaluation metrics, and implementation details employed in the experiments. For comparative experiments, we extend a subset of traditional text-to-image person retrieval methods by incorporating a TIR branch, and conduct experiments on both the RGBT-PEDES dataset and the RGBNT201-PEDES dataset. Finally, we perform ablation studies to assess the effectiveness of each component in our proposed method.

Datasets and Metrics

RGBT-PEDES is the first benchmark dataset incorporating thermal modality into text-image person retrieval, including 1,822 persons and 7,987 manually annotated descriptions. Each identity has multiple paired RGBT images from different viewpoints, with 1-2 text descriptions attached to each image pair. The training set contains 1,266 identities, 3,250 pairs of RGBT person images, and 5,497 text descriptions, while the test set includes 553 identities, 1,473 image pairs, and 2,490 descriptions.

RGBNT201-PEDES is an extended dataset derived from the original RGBNT201 dataset (Zheng et al. 2021). The base RGBNT201 dataset contains 201 identities, where each identity includes at least 20 non-consecutive image triplets covering three modalities (RGB, NIR, TIR), totaling 4,787 images. Subsequently, Wang et al. (Wang et al. 2025) generated modality-specific text descriptions for each image in RGBNT201, resulting in separate descriptions for RGB, NIR, and TIR images. To adapt these descriptions to our task, where RGB and TIR images of the same identity must share a unified text. We used a Qwen7b model (distilled by DeepSeek-R1) to merge the RGB and TIR descriptions into a single unified text. Through this process, the RGBNT201-PEDES dataset was constructed. Its training set contains 171 identities, 3,951 pairs of RGBT person images and text descriptions, and the test set includes 30 identities, 836 image pairs, and descriptions.

Metrics. We use Rank-k (k=1, 5, 10) as the primary evaluation metric, which measures the probability of finding at least one matching person image within the top-k candidate list when given a text description as a query. Meanwhile, Mean Average Precision (mAP) is used for comprehensive evaluation.

Implementation Details

Experiments are conducted on PyTorch using a single RTX A6000 48GB GPU. The image encoder referenced throughout is CLIP-ViT-B/16, and the text encoder is a 12-layer Transformer. Notably, instead of training the model from scratch or using the original CLIP weights, we initialize both encoders with pre-trained weights from (Tan et al. 2024) on the LUPerson-MLLM dataset. All input images are resized to 384 × 128 pixels, with each textual description limited to a maximum length of 77 tokens. Image augmentation is applied via random horizontal flipping, random cropping, and random erasing. Following IRRA (Jiang and Ye 2023), our model is trained for 60 epochs using the Adam optimizer, with an initial learning rate of 1×10^{-5} and cosine learning rate decay. The temperature parameter τ in the SDM loss for both teacher and student models is set to 0.02.

Comparison with State-of-the-Art Methods

In this section, we present comparison results of DIKDNet against state-of-the-art methods on two benchmark datasets.

Methods	RGBT-PEDES			
	R-1	R-5	R-10	mAP
NAFS (Ding et al. 2021)	25.83	50.31	60.99	-
TIPCB (Chen et al. 2022)	16.11	36.94	49.07	15.03
LGUR (Shao et al. 2022)	27.63	52.97	64.62	24.49
APTM (Yang et al. 2023)	53.15	78.54	85.94	49.62
IRRA (Jiang and Ye 2023)	55.62	80.14	87.91	52.43
PLIP (Zuo et al. 2024a)	26.41	48.74	59.44	23.04
TBPS-CLIP (Cao et al. 2024)	51.81	71.81	79.75	48.61
CFAM (Zuo et al. 2024b)	55.64	80.60	88.15	52.76
RDE (Qin et al. 2024)	61.57	84.26	89.88	57.81
NAM (Tan et al. 2024)	<u>65.51</u>	<u>86.26</u>	<u>92.04</u>	<u>61.33</u>
DM-Adapter (Liu et al. 2025)	51.41	77.03	87.07	49.15
VFE-TPS (Shen et al. 2025)	55.62	80.28	87.91	53.07
DCAlign (Deng et al. 2026)	58.47	82.01	89.52	54.59
DIKDNet	**69.92**	**88.80**	**93.21**	**64.86**

Table 1: Comparison results on the RGBT-PEDES dataset. The best results are bolded, and the second-best are underlined.

Performance Comparisons on RGBT-PEDES. As shown in Table 1, DIKDNet outperforms all comparative methods on the RGBT-PEDES dataset. Traditional methods (e.g., NAM, VFE-TPS, DM-Adapter) achieve only 55%–65.5% Rank-1, as they lack TIR adaptation. Our IKDM addresses this issue by enhancing TIR feature representation while preserving its modality-specific attributes through a boundary-constrained distillation strategy. DCAlign, the sole existing method tailored to this task, lags significantly

behind DIKDNet. This performance gap arises from its inability to model cross-modal and cross-view associations effectively, whereas our CKDM module effectively addresses these challenges.

Performance Comparisons on RGBNT201-PEDES. As shown in Table 2, DIKDNet achieves more modest gains over traditional methods. This limited improvement relates to the dataset's intrinsic characteristics: images are captured from consecutive video frames, resulting in small intra-identity variations in viewpoint. Our framework, which excels at leveraging cross-view diversity, struggles to demonstrate full potential here. The narrow view variance reduces opportunities to utilize its cross-modal and cross-view enhancement mechanisms. Despite this, DIKDNet still outperforms baselines, validating its generalizability.

Methods	RGBNT201-PEDES			
	R-1	R-5	R-10	mAP
NAFS (Ding et al. 2021)	23.44	38.27	49.52	-
TIPCB (Chen et al. 2022)	21.77	37.67	48.80	17.24
LGUR (Shao et al. 2022)	15.91	33.37	43.18	13.43
APTM (Yang et al. 2023)	37.56	52.63	61.84	27.85
IRRA (Jiang and Ye 2023)	37.92	52.75	61.84	29.10
PLIP (Zuo et al. 2024a)	8.01	14.59	19.13	6.62
TBPS-CLIP (Cao et al. 2024)	26.19	41.02	52.39	24.21
CFAM (Zuo et al. 2024b)	39.11	54.54	62.56	29.56
RDE (Qin et al. 2024)	38.52	56.22	65.55	27.27
NAM (Tan et al. 2024)	43.30	58.85	66.86	32.53
DM-Adapter (Liu et al. 2025)	37.08	51.67	60.41	28.15
VFE-TPS (Shen et al. 2025)	38.87	53.46	62.08	30.14
DIKDNet	**44.50**	**59.33**	**67.58**	**35.26**

Table 2: Comparison results on the RGBNT201-PEDES datasets. The best results are bolded, respectively.

Ablation Analysis

To evaluate the effectiveness of each component in DIKDNet, we conducted a comprehensive set of ablation experiments under the same experimental settings. The baseline refers to a student network that uses a shared encoder to extract RGB and TIR features. These features are fused into a unified representation, which is then constrained with text features via the SDM loss to achieve cross-modal alignment.

Effectiveness of CKDM. The baseline method (No.0 in Table 3), without any knowledge distillation components, provides a reference point for evaluating CKDM's contributions. Incorporating CVRT alone (No.1) boosts results, confirming its value in transferring cross-view robustness, ensuring retrieval consistency across varying viewpoints. Using CMST alone (No.2) validates its role in distilling cross-modal and cross-view similarity knowledge from teachers to the student, enhancing cross-view text-image alignment as designed. Combining both components (No.3) yields gains exceeding individual contributions, reflecting their synergy: CVRT strengthens cross-view consistency, CMST enhances cross-modal alignment, and together they enable the student to inherit the teachers' dual capabilities, underscoring CKDM's collaborative design.

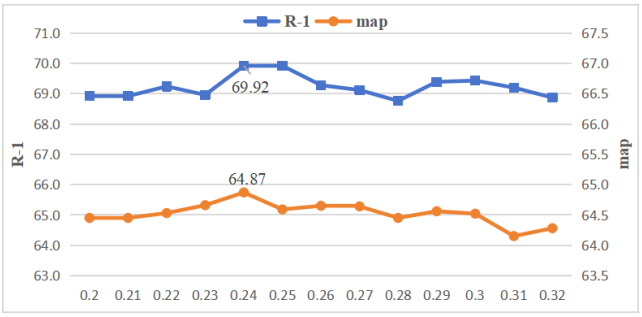

Figure 3: Impact of Hyperparameter m in IKDM on the RGBT-PEDES Dataset

Effectiveness of IKDM. Adding IKDM to CKDM (No.4) further improves all metrics, highlighting its role in enhancing TIR features. Through boundary-constrained distillation between dual-teachers, IKDM enriches TIR with RGB's semantic details while preserving TIR's structural integrity. This confirms IKDM complements CKDM by first strengthening teacher features to lay a better foundation for CKDM's knowledge transfer, forming a cohesive framework.

No.	CKDM		IKDM	R-1	R-5	R-10	mAP
	CVRT	CMST					
0				65.66	86.51	92.77	61.45
1	✓			67.51	87.87	93.01	62.75
2		✓		66.75	86.83	92.89	63.23
3	✓	✓		69.32	88.19	93.17	64.64
4	✓	✓	✓	69.92	88.80	93.21	64.86

Table 3: The ablation experiments on RGBT-PEDES demonstrate the necessity of boundary constraint.

Impact of Hyperparameter m in IKDM. The hyperparameter m, whose range is determined by observing that the distillation loss drops from 0.32 to 0.2 and then plateaus, controls the target similarity degree between RGB and TIR features in the teacher network. As shown in Fig. 3, if m is too large, the constraint will be too loose, resulting in weak supervision. if m is too small, it may cause TIR features to be excessively similar to RGB features, thereby losing their unique characteristics.

Necessity of Boundary Constraint in IKDM. Table 4 presents ablation experiments on the impact of IKDM's boundary constraint on model performance in RGBT-PEDES. As shown in the table, when the boundary constraint (a key component of IKDM) is removed, the model performance even drops below that of No. 3 in Table 3. This indicates that without IKDM's boundary control, TIR features tend to overfit to RGB features, resulting in performance degradation, thus confirming the necessity of this constraint within IKDM.

Ablation Studies on Distillation Design. Table 5 presents ablation studies on key distillation design, validating their impact on model performance. Unified Teacher (shared

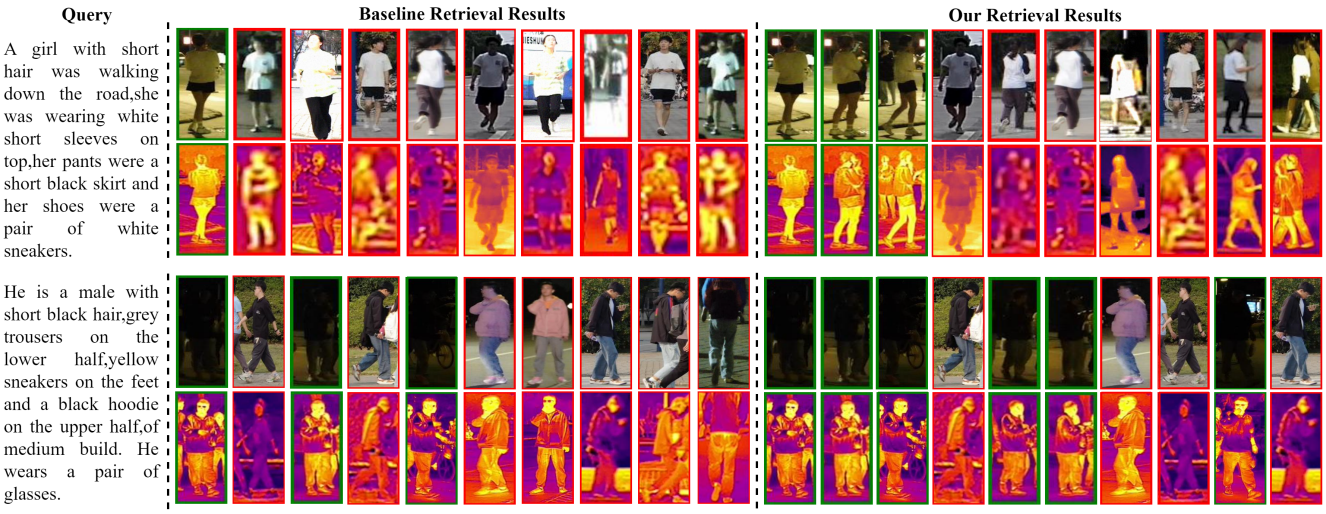

Figure 4: The results show that our method outperforms the baseline in retrieval accuracy, not only retrieving correct results under various lighting conditions but also capturing more viewpoints.

Methods	RGBT-PEDES			
	R-1	R-5	R-10	mAP
Baseline	65.66	86.51	92.77	61.45
W/O boundary constraint	68.79	88.23	93.01	64.29
W boundary constraint	69.92	88.80	93.21	64.86

Table 4: Ablation studies on the boundary constraint for RGBT-PEDES, where W/O boundary constraint denotes the model without boundary constraint and W boundary constraint denotes the model with boundary constraint.

RGB-TIR teacher weights) impairs the teacher's modality-specific discriminative capabilities, while Unified View (same-image multi-view input) weakens cross-view robustness. These results thus underscore our design choices of separate modality-specific teacher branches and multi-view input processing, which collectively enhance modality specificity and cross-view invariance.

Qualitative Results

Fig. 4 presents the top-10 retrieval results for two distinct identities, contrasting the baseline method with our approach. In the first identity of the upper group of results, under general lighting conditions where RGB is clear and TIR features are distinguishable, our method retrieves more correct results, demonstrating superior precision. In the second identity of the lower group of results, even under poor lighting conditions where RGB visual details are less obvious, our method still improves retrieval performance by relying on TIR features.

Meanwhile, the baseline method struggles to filter out irrelevant candidates, but our method not only retrieves correct matches but also accommodates different viewing angles. This improvement stems from two key components. Our IKDM bolsters TIR representation for robustness in complex lighting. The CKDM inherits cross-modal and cross-view modeling strengths from teacher networks to adapt to viewpoint variations. In essence, our method excels across lighting conditions and effectively leverages multi-view information for more reliable person retrieval.

Methods	RGBT-PEDES			
	R-1	R-5	R-10	mAP
Baseline	65.66	86.51	92.77	61.45
Unified Teacher	68.88	87.95	93.05	64.59
Unified View	68.43	88.39	93.01	63.67
DIKDNet	69.92	88.80	93.21	64.86

Table 5: The ablation experiments demonstrate the necessity of critical distillation design choices on RGBT-PEDES.

Conclusion

In this paper, we propose the DIKDNet for text-to-visible & infrared person retrieval, aiming to address the issues of insufficient TIR representation and weak cross-modal and cross-view collaborative representation capabilities in existing methods. To enhance TIR feature representation while preserving its modality-specific attributes, we design the IKDM that enables controlled knowledge transfer between RGB and TIR teacher networks through a boundary-constrained strategy. For strengthening cross-modal and cross-view collaborative representation, we introduce the CKDM, which transfers cross-modal similarity relations and cross-view multimodal representations from dual teachers to the student network. These modules jointly align multi-modal images and text into a unified embedding space. Experiments on RGBT-PEDES and RGBNT201-PEDES show DIKDNet outperforms state-of-the-art methods, with ablation studies validating each module's effectiveness.

Acknowledgments

This research is supported in part by the National Natural Science Foundation of China under Grants 62372003, the Natural Science Foundation of Anhui Province under Grants 2308085Y40.

References

Cao, M.; Bai, Y.; Zeng, Z.; Ye, M.; and Zhang, M. 2024. An empirical study of clip for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 465–473.

Chen, T.; Xu, C.; and Luo, J. 2018. Improving text-based person search by spatial matching and adaptive threshold. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1879–1887. IEEE.

Chen, Y.; Huang, R.; Chang, H.; Tan, C.; Xue, T.; and Ma, B. 2021. Cross-modal knowledge adaptation for language-based person search. *IEEE Transactions on Image Processing*, 30: 4057–4069.

Chen, Y.; Zhang, G.; Lu, Y.; Wang, Z.; and Zheng, Y. 2022. TIPCB: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494: 171–181.

Deng, Y.; Chen, Z.; Li, C.; and Tang, J. 2025a. Uncertainty-aware coarse-to-fine alignment for text-image person retrieval. *Visual Intelligence*, 3(1): 6.

Deng, Y.; Li, C.; Chen, Z.; Xu, Z.; and Tang, J. 2026. Decoupled cross-modal alignment network for text-RGBT person retrieval and a high-quality benchmark. *Information Fusion*, 127: 103948.

Deng, Y.; Li, C.; Wang, F.; and Tang, J. 2025b. Learning Hierarchical Cross-modal Association with Intra-modal Context for Text-Image Person Retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2723–2731.

Deng, Y.; Wang, G.; Li, C.; Wang, W.; Zhang, C.; and Tang, J. 2024. Collaborative license plate recognition via association enhancement network with auxiliary learning and a unified benchmark. *IEEE Transactions on Multimedia*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Ding, Z.; Ding, C.; Shao, Z.; and Tao, D. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*.

Farooq, A.; Awais, M.; Kittler, J.; and Khalid, S. S. 2022. Axm-net: Implicit cross-modal feature alignment for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 4477–4485.

Gupta, S.; Hoffman, J.; and Malik, J. 2016. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2827–2836.

Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; and Harada, T. 2017. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5108–5115. IEEE.

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.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Huang, L.; Zeng, Y.; Yang, C.; An, Z.; Diao, B.; and Xu, Y. 2024. etag: Class-incremental learning via embedding distillation and task-oriented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12591–12599.

Jiang, D.; and Ye, M. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2787–2797.

Lei, J.; Chen, X.; Zhang, N.; Wang, M.; Bansal, M.; Berg, T. L.; and Yu, L. 2022. Loopitr: Combining dual and cross encoder architectures for image-text retrieval. *arXiv preprint arXiv:2203.05465*.

Li, S.; Xiao, T.; Li, H.; Yang, W.; and Wang, X. 2017a. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE international conference on computer vision*, 1890–1899.

Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; and Wang, X. 2017b. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1970–1979.

Li, Z.; Li, X.; Fu, X.; Zhang, X.; Wang, W.; Chen, S.; and Yang, J. 2024. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26617–26626.

Liu, Y.; Li, Y.; Liu, Z.; Yang, W.; Wang, Y.; and Liao, Q. 2024. Clip-based synergistic knowledge transfer for text-based person retrieval. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7935–7939. IEEE.

Liu, Y.; Liu, Z.; Lan, X.; Yang, W.; Li, Y.; and Liao, Q. 2025. Dm-adapter: Domain-aware mixture-of-adapters for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5703–5711.

Miech, A.; Alayrac, J.-B.; Laptev, I.; Sivic, J.; and Zisserman, A. 2021. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9826–9836.

Qin, Y.; Chen, Y.; Peng, D.; Peng, X.; Zhou, J. T.; and Hu, P. 2024. Noisy-correspondence learning for text-to-image

person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27197–27206.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ring, E.; and Ammer, K. 2012. Infrared thermal imaging in medicine. *Physiological measurement*, 33(3): R33.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Shao, Z.; Zhang, X.; Fang, M.; Lin, Z.; Wang, J.; and Ding, C. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *Proceedings of the 30th acm international conference on multimedia*, 5566–5574.

Shen, W.; Fang, M.; Wang, Y.; Xiao, J.; Li, D.; Chen, H.; Xu, L.; and Zhang, W. 2025. Enhancing visual representation for text-based person searching. *Knowledge-Based Systems*, 309: 112893.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tan, W.; Ding, C.; Jiang, J.; Wang, F.; Zhan, Y.; and Tao, D. 2024. Harnessing the power of mllms for transferable text-to-image person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17127–17137.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

Wang, Y.; Lv, Y.; Zhang, P.; and Lu, H. 2025. Idea: Inverted text with cooperative deformable aggregation for multimodal object re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29701–29710.

Wang, Y.; Zhang, P.; Wang, D.; and Lu, H. 2024. Other tokens matter: Exploring global and local features of Vision Transformers for Object Re-Identification. *Computer Vision and Image Understanding*, 244: 104030.

Yan, S.; Dong, N.; Liu, J.; Zhang, L.; and Tang, J. 2023. Learning comprehensive representations with richer self for text-to-image person re-identification. In *Proceedings of the 31st ACM international conference on multimedia*, 6202–6211.

Yang, C.; An, Z.; Huang, L.; Bi, J.; Yu, X.; Yang, H.; Diao, B.; and Xu, Y. 2024. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15952–15962.

Yang, C.; An, Z.; Zhou, H.; Cai, L.; Zhi, X.; Wu, J.; Xu, Y.; and Zhang, Q. 2022. Mixskd: Self-knowledge distillation from mixup for image recognition. In *European Conference on Computer Vision*, 534–551. Springer.

Yang, S.; Zhou, Y.; Zheng, Z.; Wang, Y.; Zhu, L.; and Wu, Y. 2023. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4492–4501.

Zha, Y.; Sun, J.; Zhang, P.; Zhang, L.; Gonzalez-Garcia, A.; and Huang, W. 2022. Self-supervised cross-modal distillation for thermal infrared tracking. *IEEE MultiMedia*, 29(4): 80–96.

Zheng, A.; Wang, Z.; Chen, Z.; Li, C.; and Tang, J. 2021. Robust multi-modality person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 3529–3537.

Zhu, A.; Wang, Z.; Li, Y.; Wan, X.; Jin, J.; Wang, T.; Hu, F.; and Hua, G. 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM international conference on multimedia*, 209–217.

Zuo, J.; Hong, J.; Zhang, F.; Yu, C.; Zhou, H.; Gao, C.; Sang, N.; and Wang, J. 2024a. Plip: Language-image pre-training for person representation learning. *Advances in Neural Information Processing Systems*, 37: 45666–45702.

Zuo, J.; Zhou, H.; Nie, Y.; Zhang, F.; Guo, T.; Sang, N.; Wang, Y.; and Gao, C. 2024b. Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22010–22019.