TokenMatcher: Diverse Tokens Matching for Unsupervised Visible-Infrared Person Re-Identification

Xiao Wang[1,2], **Lekai Liu**[1,3], **Bin Yang**[2*], **Mang Ye**[2], **Zheng Wang**[2], **Xin Xu**[1,3*]

[1]School of Computer Science and Technology, Wuhan University of Science and Technology
[2] School of Computer Science, Wuhan University
[3] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, China
{wangxiao2021,liulekai,xuxin}@wust.edu.cn,{yangbin_cv,yemang,wangzwhu}@whu.edu.cn

Abstract

Unsupervised visible-infrared person re-identification (US-VI-ReID) seeks to match infrared and visible images of the same individual without the use of annotations. Current methods typically derive cross-modal correspondences through a single global feature matching process for generating pseudo labels and learning modality-invariant features. However, this matching approach is hindered by both intra-modality and inter-modality discrepancies, which result in imprecise measurements. As a consequence, the clustering of individuals with single global feature is often incomplete and unreliable, leading to suboptimal performance in cross-modal clustering tasks. To address these challenges and to extract cross-modality discriminative identity information, we propose a TokenMatcher, which encompasses three key components: Diverse Tokens Matching (DTM), Diverse Tokens Neighbor Learning (DTNL), and the Homogeneous Fusion (HF) Module. DTM utilizes multiple class tokens within the visual transformer framework to capture diverse embedding representations, thereby facilitating the integration of fine-grained information essential for reliable cross-modality correspondences. DTNL enhances the intra-modality and inter-modality consistency among diverse tokens by refining neighborhood sets with insights from neighboring tokens and camera information, promoting robust neighborhood learning and fostering discriminative identity information. Additionally, the HF module consolidates clusters of the same identity while effectively separating those of different identities. Extensive experiments conducted on the publicly available SYSU-MM01 and RegDB datasets demonstrate the efficacy of the proposed method.

Code — https://github.com/liulekai123/TokenMatcher

1 Introduction

Person re-identification (ReID) focuses on matching images of the same individual across disparate camera views (2021; 2024; 2022; 2023c; 2024; 2021). The increasing prevalence of smart surveillance cameras equipped with infrared (IR) capabilities is noteworthy, particularly as burglary incidents are significantly higher at night compared to daytime (2020; 2024; 2024; 2024; 2023; 2023). Consequently, the visible-infrared person re-identification (VI-ReID) task has garnered considerable attention within the research community (2019; 2020; 2020; 2020; 2020). Despite the impressive performance demonstrated by numerous studies in VI-ReID (Ye et al. 2023; Wu et al. 2021; Hu, Yang, and Ye 2024; Liu, Ye, and Du 2024), these approaches typically rely on extensive labeled cross-modal datasets. The manual annotation of such large volumes of cross-modal data is both labor-intensive and costly, thereby constraining its practical applicability in industrial settings. To address this challenge, unsupervised visible-infrared ReID (US-VI ReID) has been proposed, which often derives identity labels through clus-

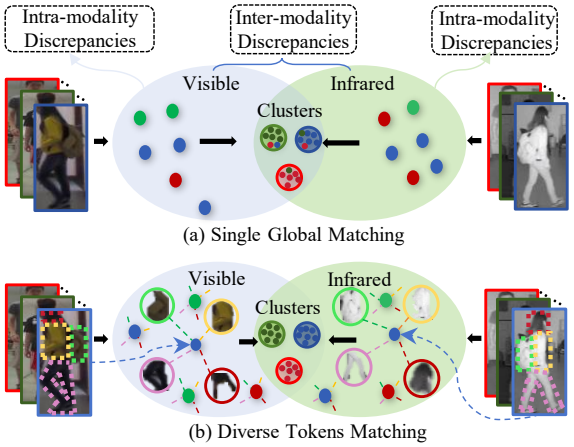

Figure 1: **Idea illustration.** (a) Previous research primarily concentrated on establishing cross-modality correspondences through a single global matching approach. However, this method is impeded by both intra-modality and inter-modality discrepancies, resulting in the use of coarser global features that are inadequate for capturing fine-grained details. Consequently, this leads to ambiguous correspondences and suboptimal clustering outcomes. (b) In this paper, we advance the discourse by emphasizing the fine-grained representation of diverse tokens for matching and propagating cross-modality labels. This approach is inherently more comprehensive and reliable, enhancing the accuracy of identity recognition across modalities.

*Corresponding Author

tering techniques, thereby mitigating the need for expensive cross-modal annotations.

Existing works (2023; 2022; 2022; 2023b; 2021; 2023b) utilize single global feature matching to establish person identities for cross-modality correspondences. However, these approaches struggle to address the discrepancies between the two modalities, namely intra-modality and inter-modality discrepancies. These discrepancies lead to a coarse granularity in the single global feature matching approach (Wu and Ye 2023; Yang et al. 2022; Wang et al. 2022), making it incapable of accurately representing individuals (as illustrated in Figure 1 (a)). Consequently, the cross-modality similarity derived from this method lacks sufficient reliability, ultimately impacting the accuracy of unsupervised visible-infrared person re-identification (US-VI-ReID).

To enhance the reliability of cross-modality correspondences with fine-grained person representation, we introduce TokenMatcher (as shown in Figure 1 (b)) for unsupervised visible-infrared person re-identification (US-VI-ReID). TokenMatcher comprises three key modules: Diverse Tokens Matching (DTM), Diverse Tokens Neighbor Learning (DTNL), and Homogeneous Fusion (HF). DTM utilizes multiple class tokens to represent various embedding representations within the visual transformer (Li et al. 2023). These diverse and compact transformers encapsulate more fine-grained information. By leveraging fine-grained features from multiple embedding spaces, we gain a comprehensive understanding of detailed information, which forms a critical foundation for establishing reliable cross-modality correspondences. DTNL focuses on exploiting the intra-modality and inter-modality consistency among diverse tokens in a dynamic manner during each training iteration. It refines the neighborhood set by leveraging information from neighboring tokens and enhancing this data with available camera information, thereby facilitating robust neighborhood learning. This approach cultivates discriminative identity information, ensuring that the learned representations are more reliable. In HF module, we implement a mechanism that pulls split clusters with the same identity closer together while simultaneously pushing apart clusters corresponding to different identities. This approach promotes the formation of more cohesive identity representations across modalities, enhancing the overall effectiveness of the US-VI-ReID system.

The contributions are summarized as follows:

- We introduce a TokenMatcher, a novel framework designed to extract reliable cross-modality fine-grained person features, facilitating accurate cross-modality correspondences.
- We present the Diverse Tokens Neighbor Learning (DTNL) module, which identifies reliable neighbors. This capability allows the model to effectively capture modality-invariant and discriminative features.
- We propose the Homogeneous Fusion (HF) module, which aims to minimize the differences between various camera views, thereby drawing clusters with the same identity closer together.
- Experiments on SYSU-MM01 and RegDB datasets demonstrate the superiority of our method compared with existing US-VI-ReID methods.

2 Related Work

2.1 Supervised Visible-Infrared Person ReID

The objective of Visible-Infrared Person Re-Identification (VI-ReID) is to accurately match images of the same individual across visible and infrared modalities (Huang et al. 2023). Previous approaches, such as Diverse Embedding Expansion Network (DEEN) (Zhang and Wang 2023), have introduced augmentation networks within embedding spaces that effectively learn feature representations while minimizing the modality gap. In contrast to the generative adversarial network (GAN) models (Karras et al. 2018; Miyato et al. 2018) focused on reducing modality gaps, Grayscale Enhancement Colorization Network (GECNet) (Zhong et al. 2021) takes a different approach by generating intermediate grayscale images. It learns correspondences between infrared and visible images to effectively colorize infrared images. Channel Augmentation (CAJ) (Ye et al. 2023) offers an efficient means of generating intermediate modalities by randomly swapping image channels.

Despite the promising results achieved by these methods in VI-ReID tasks, a significant limitation remains: they often rely on a large number of manually annotated cross-modality images. This is not only time-consuming but also costly, which poses a substantial barrier to the practical application of these techniques in real-world scenarios.

2.2 Unsupervised Single-Modality Person ReID

Unsupervised single-modality person re-identification (ReID) aims to retrieve pedestrian images using unlabeled visible data (Li et al. 2018; Fu et al. 2019; Delorme et al. 2020). The CAP method (Wang et al. 2021) introduces a novel approach by partitioning each cluster into multiple camera proxies, effectively addressing significant intra-ID variance and generating more reliable pseudo-labels for learning. Recently, the cluster-contrast method (Dai et al. 2022) has shown outstanding performance by incorporating a cluster-level memory dictionary and employing a momentum updating strategy. ICE (Chen, Lagadec, and Bremond 2021) enhances the effectiveness of previous class-level contrastive ReID methods by introducing inter-instance contrastive encoding, which leverages pairwise similarity scores among instances. Additionally, CALR (Li et al. 2024) implements intra-camera training to derive reliable local pseudo-labels within each camera, which are then used to refine the global pseudo-labels. Nevertheless, due to the substantial cross-modality variance, these approaches struggle to produce dependable cross-modality pseudo-labels and are limited in their ability to effectively learn modality-invariant features when applied to unsupervised visible-infrared ReID (US-VI-ReID).

2.3 Unsupervised Visible-Infrared Person ReID

Unsupervised Visible-Infrared Person Re-identification (US-VI-ReID) focuses on extracting modality-invariant features despite the presence of noisy pseudo-labels. The

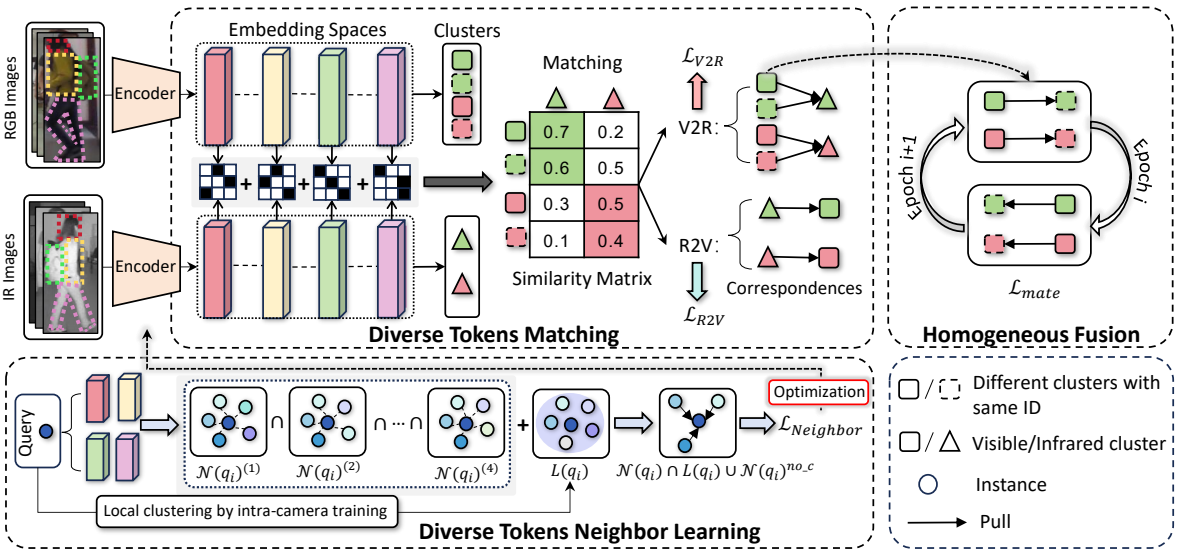

Figure 2: **The pipeline of TokenMatcher.** It contains the Diverse Tokens Matching (DTM), Diverse Tokens Neighbor Learning (DTNL), and Homogeneous Fusion (HF) Modules. By leveraging complementary information among diverse tokens, TokenMatcher enhances the robustness of the learned representations and effectively establishes neighborhood relationships.

Augmented Dual-Contrastive Aggregation (ADCA) method (Yang et al. 2022) has established a robust baseline for Unsupervised Learning in Visible-Infrared ReID (USL-VI-ReID). Additionally, strategies such as those implemented in Optimal Transport Label Assignment (OTLA) (Wang et al. 2022) and its extension DOTLA (Cheng et al. 2023b) utilize optimal transport methodologies to transfer pseudo-labels between modalities. Furthermore, approaches like PGM (Wu and Ye 2023) and MBCCM (Cheng et al. 2023a) employ graph matching techniques to extract cross-modality correspondences. In contrast, the approach delineated in (Yang et al. 2023a) synthesizes online feature interaction learning with offline cross-modality label refinement to establish robust correspondences. (Yang, Chen, and Ye 2024) combines shallow and deep features through collaborative learning and label association in order to accurately capture modality-invariant features.

However, the aforementioned methods often struggle to effectively represent fine-grained features of individuals. The exclusive reliance on coarser global features proves inadequate for accurately learning cross-modality representations, due to the interference caused by various discrepancies. In response, our approach emphasizes the comprehensive exploration of distinguishable individual characteristics and collaboratively leverages these insights for enhanced feature learning and cross-modality label association.

3 Methodology

The framework of our proposed methods is illustrated in Figure 2. To establish reliable cross-modality correspondences, we introduce the Diverse Tokens Matching (DTM) component. This module is designed to effectively identify and match diverse tokens across modalities. Building on this, the Diverse Tokens Neighbor Learning (DTNL) component leverages the connections between multiple embedding spaces and camera information to facilitate reliable neighbor learning. Finally, the Homogeneous Fusion (HF) module is proposed to integrate different clusters sharing the same ID identified by DTM. This fusion process aims to further minimize the intra-modality gap, enhancing the overall robustness and accuracy of the feature representations.

3.1 Baseline

Modality-Specific Memory Initialization. At the beginning of each training epoch, we construct modality-specific cluster memories using the following equation:

$$\phi_k^e = \frac{1}{|\mathcal{H}_k^e|} \sum_{u_n^e \in \mathcal{H}_k^e} u_n^e, \tag{1}$$

where u_n^e represents an instance feature concatenated with each class token. The term $|\mathcal{H}_k^e|$ denotes the number of instances within the k-th cluster set The variable e can take on values $\{v,r\}$, indicating visible or infrared modality, respectively. Throughout the training process, we update both memories using a momentum updating strategy as described in (Yang et al. 2022)

Contrastive Loss. Given visible and infrared query features (q^v and q^r). We can calculate the contrast loss by the following equations:

$$\mathcal{L}_{id}^v = -\log \frac{\exp\left(q^v \cdot \phi_+^v/\tau\right)}{\sum_{k=0}^{N^v} \exp\left(q^v \cdot \phi_k^v/\tau\right)}, \tag{2}$$

$$\mathcal{L}_{id}^r = -\log \frac{\exp\left(q^r \cdot \phi_+^r/\tau\right)}{\sum_{k=0}^{N^r} \exp\left(q^r \cdot \phi_k^r/\tau\right)}, \tag{3}$$

$$\mathcal{L}_{id} = \mathcal{L}_{id}^v + \mathcal{L}_{id}^r, \tag{4}$$

where N^v, N^r indicates the number of visible or infrared clusters. ϕ_+ is the positive cluster memory corresponding to the pseudo-label of q. τ is a temperature factor.

Self-Diverse Constraint Loss. Following DC-Former (Li et al. 2023), multiple class tokens are concatenated with patch embeddings, followed by the addition of positional embeddings. These combined embeddings are fed into the transformer encoder. Then a self-diverse constraint is applied to these class tokens in the final transformer layer to push them far way from each other, thereby leading to diverse representation spaces. The self-diverse constraint loss can be expressed by the following equations:

$$\begin{aligned} \mathcal{L}_{SDC} &= \sum_i \sum_j \omega_{ij}\nu_{ij},\ i<j,\ i,j=1,...,N \\ \omega_{ij} &= \frac{exp(\nu_{ij})}{\sum_m \sum_n exp(\nu_{mn})},\ m<n,\ m,n=1,...,N \end{aligned} \tag{5}$$

where $\nu_{ij} = |\cos(f_i, f_j)|$, and f_i, f_j indicates any two class tokens. N is the number of class tokens.

3.2 Diverse Tokens Matching

We integrate various embedding spaces to compute cross-modality similarity, thereby generating reliable similarity matrices. These matrices are subsequently employed to identify cross-modality correspondences using our proposed cross-modality correlation algorithm.

Corresponding Calculate Similarity Matrix. Two modalities can obtain N token features each through modal, so for a pair of cross-modality clusters we can get $N * N$ cross-modalities similarities. How can we best utilize these $N * N$ similarities? We observe that the similarity computed between corresponding embedding spaces across modalities is more accurate than that derived from non-corresponding spaces. This is because different embedding spaces emphasize distinct aspects. To address this, we compute similarity using the sum of corresponding embedding spaces of different modalities. We use $S = \{S(i,j)\}$ to denote the similarity matrix with each element illustrating the similarity of cluster ϕ_i^v and ϕ_j^r. The Similarly Matrix is computed by:

$$\begin{aligned} S(i,j) &= \sum_{n=1}^{N} exp(sim(\phi_i^{v(n)}, \phi_j^{r(n)})), \\ sim(\phi_i^{v(n)}, \phi_j^{r(n)}) &= \frac{\phi_i^{v(n)} \cdot \phi_j^{r(n)}}{||\phi_i^{v(n)}|| \times ||\phi_j^{r(n)}||}, \\ \phi_k^{e(n)} &= \frac{1}{|\mathcal{H}_k^e|} \sum_{u_l^e \in \mathcal{H}_k^e} u_l^{e(n)},\ n=1,...,N \end{aligned} \tag{6}$$

where $u_l^{e(n)}$ is an instance feature of the n-th class token. The total similarity between a pair of clusters is computed by aggregating the similarities obtained from the same embedding spaces across different modalities. This approach is more comprehensive and reliable than relying on similarities derived from a single embedding space.

Cross-Modality Matching Algorithm. To address the issue where the same pedestrian may be divided into different clusters due to camera differences, we propose a cross-modality labels matching algorithm. This algorithm aims to ensure that clusters representing the same ID have consistent cross-modality labels whenever possible.

Algorithm 1: Cross-modality labels Matching

Input: The similarity matrix $S \in \mathbb{R}^{Y^v \times Y^r}$ (suppose $Y^v > Y^r$)

Output: **V2R** and **R2V** with keys storing clusters and values storing their correspondences.

Initialize and empty array of length Y^r: matched_r. And copy similarity matrix S as $S1$.
while len(**V2R**) $\neq Y^v$ **do**
  Get i and j according to Eq.7
  if matched_r$[j] < \lambda$ **then**
    V2R$[i].append(j)$; **R2V**$[j].append(i)$;
    $S[i,:] = -1$; matched_r$[j] += 1$;
  else
    $S[i,j] = -1$; *// -1 means skip this value*
  end if
end while
Take the subscript in matched_r whose value is not 0 as unmatched_r. And let $S \leftarrow S1$[unmatched_r] $= -1$;
while len(**R2V**) $\neq Y^r$ **do**
  Repeat for 3-9 above. Note that **V2R** would not update.
end while
return V2R and **R2V**

Specifically, we select the most similar pair of cross-modality clusters from the similarity matrix at each step. Assignment stops if the infrared cluster already has more than λ correspondences. The most similar pair of cross-modality clusters is selected using the following equation:

$$(i,j) = \arg\max_{i,j} C,\ i=1,...,Y^v,\ j=1,...,Y^r, \tag{7}$$

where $\arg\max$ get the row and column ordinal numbers (i, j) of the global minimum from the matrix.

By selecting the most similar cross-modality clusters from the current similarity matrix each time, we ensure that multiple clusters associated with the same cross-modality cluster are highly similar. This increases the likelihood that clusters split from the same ID will receive the same cross-modality label. Detailed steps are provided in Algorithm 1.

Modality-Shared Contrastive Learning. We use the modality-shared memory to learn cross-modality features. It consists of *infrared to visible* (R2V) learning and *visible to infrared* (V2R) learning. The $\mathcal{L}_{\text{V2R}}$ can be formulated as follows:

$$\phi_k^{vs} = \frac{1}{|\mathcal{H}_k^v \bigcup \mathcal{H}^r_{\text{V2R}[k]}|}\Big(\sum_{u_n^v \in \mathcal{H}_k^v} u_n^v + \sum_{u_n^r \in \mathcal{H}^r_{\text{V2R}[k]}} u_n^r\Big), \tag{8}$$

$$\mathcal{L}_{\text{V2R}} = -\log \frac{\exp\left(q_i^v \cdot \phi^{rs}_{\hat{y_i^v}}/\tau\right)}{\sum_{k=0}^{N^r} \exp\left(q_i^v \cdot \phi_k^{rs}/\tau\right)}, \tag{9}$$

where ϕ_k^{vs} is the modality-shared memory of visible. $\hat{y_i^v} = \text{V2R}[y_i^v]$ is the cross-modality label for q_i^v. The details of $\mathcal{L}_{\text{R2V}}$ are given in the supplementary material. Modality-Shared Contrastive Learning can be formulated as $\mathcal{L}_{scl} =$

$\mathcal{L}_{\text{V2R}} + \mathcal{L}_{\text{R2V}}$. Subsequently, $\mathcal{L}_{id}$ is reformulated using modality-shared memory.

Discussion. Unlike previous approaches (Yang et al. 2022; Wu and Ye 2023; Wang et al. 2022), which rely solely on a single-view similarity matrix derived from global features to assess cross-modality similarity, which overlooks fine-grained pedestrian features and fail to provide a comprehensive assessment of cross-modality clusters, limiting the accuracy of cross-modality associations. In contrast, our approach leverages multiple embedding spaces to generate several reliable similarity matrices. These matrices capture diverse fine-grained details about pedestrians, offering a robust foundation for accurate cross-modality associations.

3.3 Diverse Tokens Neighbor Learning

Our fundamental principle is to leverage the intrinsic consistency between embedding spaces to constrain inter- and intra-modality neighbors and enhance the robustness against multiple discrepancies.

Intra- and Inter-Modality Neighborhood Sets. Given query q_i^v, we utilize the consistency of the N embedding spaces to define the neighborhood sets. Its intra-modality neighborhood set can be defined as follows:

$$\mathcal{N}^v(q_i^v) = \{\mathcal{N}^v(q_i^v)^{(1)} \cap ... \cap \mathcal{N}^v(q_i^v)^{(N)}\}, \tag{10}$$

where $\mathcal{N}^v(q_i^v)^{(n)}$ is the intra-modality neighborhood set obtained from the n-th embedding space of q_i^v. It can be defined as follows:

$$\begin{aligned}\mathcal{N}^v(q_i^v)^{(n)} =& \{u_j^{v(n)} | sim(q_i^{v(n)}, u_j^{v(n)}) > \\ & \gamma \cdot \max_{j=1...N^v} sim(q_i^{v(n)}, u_j^{v(n)})\}.\end{aligned} \tag{11}$$

In this context, γ specifies the range for selecting neighbors. Similarly, the intra-modality neighbors of q_j^r can be identified as $\mathcal{N}^r(q_j^r)$.

Given query q_i^v, the inter-modality neighborhood set can be defined as follows:

$$\mathcal{N}^r(q_i^v) = \{\mathcal{N}^r(q_i^v)^{(1)} \cap ... \cap \mathcal{N}^r(q_i^v)^{(N)}\}, \tag{12}$$

where $\mathcal{N}^r(q_i^v)^{(n)}$ can be defined as follows:

$$\begin{aligned}\mathcal{N}^r(q_i^v)^{(n)} =& \{u_j^{r(n)} | sim(q_i^{v(n)}, u_j^{r(n)}) > \\ & \gamma \cdot \max_{j=1...N^r} sim(q_i^{v(n)}, u_j^{r(n)})\}.\end{aligned} \tag{13}$$

Similarly, the inter-modality neighbors of q_j^r can be obtained as $\mathcal{N}^v(q_j^r)$. Thus, we can get the neighborhood sets $\mathcal{N}^v(q_i^v)$, $\mathcal{N}^r(q_j^r)$, $\mathcal{N}^r(q_i^v)$ and $\mathcal{N}^v(q_j^r)$ by the constraints of all embedding spaces.

Refinement of Neighborhood Sets. Inspired by (Li et al. 2024), we employ locally reliable pseudo-labels to deal with global noise. Initially, outside of this framework, we learn independently for each camera via the baseline and store reliable local clusters. Given query q_i^v, we can get neighbor sets $\mathcal{N}^v(q_i^v)$, and corresponding local cluster $L(q_i^v)$ generated by the intra-camera training, and corresponding camera label c. The process of refining the neighbor set $\mathcal{N}^v(q_i^v)$ is represented as:

$$\mathcal{N}^v(q_i^v)_{refine} = \mathcal{N}^v(q_i^v) \cap L(q_i^v) \cup \mathcal{N}^v(q_i^v)^{no\text{-}c}, \tag{14}$$

where $\mathcal{N}^v(q_i^v)^{no\text{-}c}$ is the set of all the $\mathcal{N}^v(q_i^v)$ samples except samples captured from camera c. Similarly, $\mathcal{N}^r(q_i^v)$, $\mathcal{N}^r(q_i^r)$ and $\mathcal{N}^v(q_i^r)$ can be refined by the same way. This approach allows us to effectively remove impurities within the same camera's neighborhood sets, thereby improving their overall reliability.

Neighbor Learning. Given query q_i^v and q_i^r, we use the formula to calculate visible-visible neighbor learning:

$$\mathcal{L}_{v\to v} = -K \sum_{i=1}^{N^b} \sum_{j \in \mathcal{N}^v(q_i^v)} \log \frac{\exp\left(sim(q_i^v, u_j^v)/\tau\right)}{\sum_{k=1}^{N^v} \exp\left(sim(q_i^v, u_k^v)/\tau\right)}, \tag{15}$$

where $K = \frac{1}{N^b \cdot |\mathcal{N}^v(q_i^v)|}$. N^b is the batch size of query q_i. Similarly, we can obtain neighbor learning for visible-infrared $\mathcal{L}_{v\to r}$, infrared-infrared $\mathcal{L}_{r\to r}$ and infrared-visible $\mathcal{L}_{r\to v}$. The following combination denotes the final optimization for Diverse Tokens Neighbor Learning:

$$\mathcal{L}_{neighbor} = \mathcal{L}_{v\to v} + \mathcal{L}_{v\to r} + \mathcal{L}_{r\to r} + \mathcal{L}_{r\to v}. \tag{16}$$

Discussion. In contrast to the neighborhood learning approach presented in (Yang, Chen, and Ye 2024), we leverage the intrinsic consistency of various fine-grained information about pedestrians to derive the neighborhood sets. Additionally, we refine these sets using available camera information. Our method enables the extraction of more reliable neighborhood sets, facilitating effective feature alignment.

3.4 Homogeneous Fusion

Due to our excellent cross-modality correlation (DTM), there is a high similarity between multiple visible clusters associated with the same infrared cluster, and they are probably those split clusters with the same ID. On this basis, we aggregate the different clusters with the same ID to further alleviate the intra-modality gap with the Homogeneous Fusion (HF) module. Given visible query q^v, suppose that $len(\text{R2V}[\text{V2R}[y^v]]) = M$ and satisfy $M > 1$. Thus, we can get multiple different visible clusters with the same cross-modality label that can be denoted as $q^v(i) = \text{R2V}[\text{V2R}[y^v]][i], i < M$. We use the alternate contrastive learning scheme (Wu and Ye 2023). The process of homogeneous fusion using modality-specific memory ϕ_k^v is formulated as follows:

$$\begin{aligned}\mathcal{L}_{i\to j} &= -\log \frac{\exp\left(q^v(i) \cdot \phi_{y^v(j)}^v/\tau\right)}{\sum_{k=0}^{N^v} \exp\left(q^v(i) \cdot \phi_k^v/\tau\right)}, \\ \mathcal{L}_{mate} &= \begin{cases} \sum_{i=0}^{M-2} \mathcal{L}_{i\to(i+1)}, & if Epoch\%2 = 0 \\ \sum_{i=0}^{M-2} \mathcal{L}_{(i+1)\to i}, & if Epoch\%2 = 1 \end{cases}.\end{aligned} \tag{17}$$

The total training loss $\mathcal{L}$ can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{id} + \mathcal{L}_{SDC} + \beta_1 \mathcal{L}_{scl} + \beta_2 \mathcal{L}_{neighbor} + \beta_3 \mathcal{L}_{mate}. \tag{18}$$

4 Experiments

4.1 Datasets and Evaluation Protocol

Datasets. We evaluate the proposed methods on the SYSU-MM01 (Wu et al. 2017) and RegDB (Nguyen et al. 2017)

	Methods	Venue	SYSU-MM01						RegDB					
			All Search			Indoor Search			Visible to Infrared			Infrared to Visible		
			$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP
Supervised	AGW (Ye et al. 2021)	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
	MSO (Gao et al. 2021)	MM-21	58.70	56.42	-	63.07	70.31	-	73.6	66.9	-	74.6	67.5	-
	FMCNet (Zhang et al. 2022)	CVPR-22	66.34	62.51	-	68.15	74.09	-	89.12	84.43	-	88.38	83.86	-
	MAUM (Liu et al. 2022)	CVPR-22	71.68	68.79	-	76.97	81.94	-	87.87	85.09	-	86.95	84.34	-
	DEEN (Zhang and Wang 2023)	CVPR-23	74.7	71.8	-	80.3	83.3	-	91.1	85.1	-	89.5	83.4	-
	PMCM (Qian, Lin, and Du 2023)	IJCAI-23	75.54	71.16	-	81.52	84.33	-	93.09	89.57	-	91.44	87.15	-
	PartMix (Kim et al. 2023)	CVPR-23	77.78	74.62	-	81.52	84.38	-	84.93	82.52	-	85.66	82.27	-
	DEN (Kim, Gwon, and Seo 2024)	WACV-24	76.36	71.3	-	83.56	84.65	-	95.34	90.21	-	94.98	90.24	-
Unsupervised	OTLA (Wang et al. 2022)	ECCV-22	29.9	27.1	-	29.8	38.8	-	32.9	29.7	-	32.1	28.6	-
	H2H (Liang et al. 2021)	TIP-21	30.15	29.40	-	-	-	-	23.81	18.87	-	-	-	-
	ADCA (Yang et al. 2022)	MM-22	45.51	41.73	28.29	50.60	59.11	55.17	67.20	64.05	52.67	68.48	63.81	49.62
	CHCR (Pang et al. 2023)	TCSVT-22	59.47	59.14	-	-	-	-	69.31	64.74	-	69.96	65.87	-
	DOTLA (Cheng et al. 2023b)	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
	MBCCM (Cheng et al. 2023a)	MM-23	53.14	48.16	32.41	55.21	61.98	57.13	83.79	77.87	65.01	82.82	76.74	61.73
	PGM (Wu and Ye 2023)	CVPR-23	57.27	51.78	34.96	56.23	62.74	58.13	69.48	65.41	52.97	69.85	65.17	-
	CMAM(Wu, Lin, and Zheng 2024)	TCSVT-24	62.0	58.2	-	67.6	72.7	-	89.1	74.0	-	89.0	74.0	-
	DCCL (Yang et al. 2023a)	TIFS-23	63.18	58.62	42.99	66.67	71.82	67.46	78.28	71.98	58.79	78.28	71.30	55.23
	GUR (Yang, Chen, and Ye 2023)	ICCV-23	63.51	61.63	47.93	71.11	76.23	72.57	76.91	70.23	58.88	75.00	69.94	56.21
	SDCL (Yang, Chen, and Ye 2024)	CVPR-24	64.49	**63.24**	**51.06**	**71.37**	**76.90**	**73.50**	86.91	78.92	62.83	85.76	77.25	59.57
	Ours	-	**65.07**	62.79	50.18	68.97	74.89	71.23	**92.96**	**86.32**	**72.93**	**91.82**	**85.17**	**70.52**

Table 1: **Comparison with SOTA on SYSU-MM01 and RegDB.** Rank at r accuracy(%), mAP (%) and mINP (%) are reported.

					SYSU-MM01						RegDB					
	Components				All Search			Indoor Search			Visble to Infrared			Infrared to Visible		
Index	Baseline	DTM	DTNL	HF	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP	$r1$	mAP	mINP
1	✓				51.27	51.48	39.92	54.35	63.38	59.97	53.65	49.87	35.68	52.32	47.87	33.95
2	✓			✓	53.54	53.81	42.43	56.44	65.35	62.00	59.49	55.10	40.47	58.00	52.88	38.05
3	✓		✓		54.35	52.89	40.38	57.40	64.97	61.20	55.62	51.68	37.67	53.81	49.52	35.68
4	✓	✓			62.16	61.46	50.18	67.63	74.28	71.05	91.93	85.43	72.19	91.04	84.42	69.86
5	✓	✓	✓		64.52	62.36	49.83	68.25	74.83	70.53	92.80	86.12	72.76	91.79	84.97	70.20
7	✓	✓	✓	✓	**65.07**	**62.79**	**50.18**	**68.97**	**74.89**	**71.23**	**92.96**	**86.32**	**72.93**	**91.82**	**85.17**	**70.52**

Table 2: **Ablation studies on the SYSU-MM01 and RegDB.** Rank at r accuracy (%), mAP (%) and mINP (%) are reported.

datasets. Further details and explanations can be found in the supplementary materials.

Evaluation Protocols. We follow widely used protocols (Ye et al. 2021) to assess the two datasets, where mean precision (mAP), cumulative matching characteristic (CMC), and mean Inverse Negative Penalty (mINP) are adopted.

Implementation Details. The model is trained over 100 epochs. For the first 50 epochs, we use the baseline method for learning. In the subsequent 50 epochs, we incorporate the proposed framework. Additional details on the settings are provided in the supplementary materials.

4.2 Comparison with State-of-the-art Methods

In Table 1, our proposed methods are compared with supervised and unsupervised VI-ReID methods on two datasets.

Comparison with Unsupervised Methods. As shown in Table 1, our methods outperform current advanced unsupervised approaches. Specifically, we achieve 65.07% and 92.96% rank-1 accuracy on SYSU-MM01 (all search) and RegDB (visible to infrared), respectively. Compared to the best current method, SDCL (Yang, Chen, and Ye 2024), our methods improve rank-1 accuracy by about 0.6% on SYSU-MM01 and by 6% rank-1 accuracy on RegDB.

Comparison with Supervision Methods. Comparisons with advanced supervised methods show that our approach outperforms several, including AGW (Ye et al. 2021) and MSO (Gao et al. 2021), and is competitive with FMCNet (Zhang et al. 2022). These gains are due to our effective use of complementary information among diverse tokens. However, our methods still differ significantly from state-of-the-art fully supervised results, primarily due to the lack of annotation information.

4.3 Ablation Study

In this subsection, we carefully conduct ablation experiments to verify the effectiveness of each component of our methods, and the results are shown in Table 2.

Effectiveness of DTM. Compared to the baseline, the DTM significantly improves by 11% and 38% on the SYSU-MM01 and RegDB datasets, respectively. Our advantages are mainly using different fine-grained features of pedestrians and aligning them to go for cross-modality matching in multiple perspectives.

Effectiveness of DTNL. There is considerable improve-

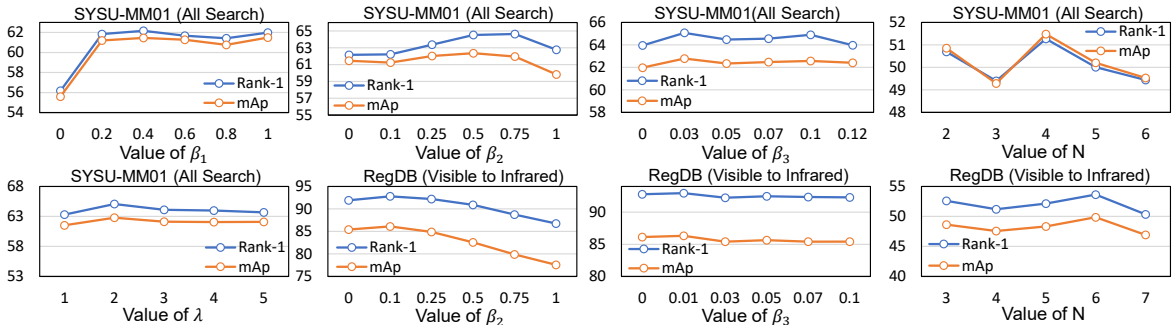

Figure 3: Performance of our framework with different values of β_1, β_2, β_3, λ and N on SYSU-MM01 and RegDB datasets.

Meathod	All Search			Indoor Search		
	$r1$	mAP	mINP	$r1$	mAP	mINP
single	58.18	57.08	44.98	63.30	70.58	67.02
N*N	63.31	61.62	49.65	67.52	73.95	70.67
Ours	65.07	62.79	50.18	68.97	74.89	71.23

Table 3: Comparison of different methods for calculating cross-modality similarity on SYSU-MM01.

ment in DTNL performance compared to the DTM. DTNL obtains reliable homogeneous and heterogeneous neighborhoods by capturing intrinsic connections in the individual embedding spaces, enhancing the model's robustness.

Effectiveness of HF. HF further improves mAP by +0.43% and +0.2% on the two datasets, respectively. This improvement is due to the DTM method, which ensures consistent cross-modality correspondences for clusters with the same ID, allowing us to obtain a reliable cluster set of different clusters with the same ID.

4.4 Further Analysis

Hyper-parameters Analysis. Hyper-parameter β_1, β_2 and β_3 are weighting parameters to trade-off $\mathcal{L}_{scl}$, $\mathcal{L}_{neighbor}$ and $\mathcal{L}_{mate}$ in Eq.18, as shown in Figure 3. β_1 is set to 0.4 for optimal performance. We also observe that the performance is insensitive to β_2 and β_3, it achieves the best performance on SYSU-MM01 and RegDB datasets when β_2 = 0.5 or 0.1 and β_3 = 0.03 or 0.01. The number of class tokens N is set to 4 for SYSU-MM01 and 6 for RegDB. To prevent a single cluster from matching too many cross-modality clusters, we set λ to 2 in Algorithm 1, reflecting the initial ratio of the two modal clusters.

Corresponding Calculate Similarly Matrix. We compared the Corresponding Calculate Similarity Matrix with two other approaches: one using a single embedding space as in previous work, and another using the sum of $N * N$ similarities from N embedding spaces. As shown in Table 3, computing similarity with multiple embedding spaces is more accurate and captures finer details. In contrast, summing $N * N$ similarities as the final measure overlooks differences between non-corresponding embedding spaces.

Cross-modality Matching Algorithm. We tested our cross-modality correspondence algorithm on the framework proposed by PGM (Wu and Ye 2023), using the same parameters. As shown in Table 4, our method outperforms PGM by 4% in Rank-1 and mAP accuracy. This is due to the higher likelihood that clusters originating from the same ID will share a consistent cross-modality label.

Meathod	Visible to Infrared			Infrared to Visible		
	$r1$	mAP	mINP	$r1$	mAP	mINP
PGM	69.48	65.41	52.97	69.85	65.17	-
Ours	73.00	70.11	59.98	73.85	69.60	57.06

Table 4: Comparison of different cross-modality correspondences methods on RegDB.

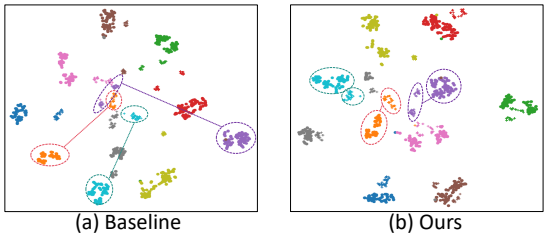

Figure 4: The t-SNE visualization of 10 randomly selected identities. Colors denote identities, with circles for visible modality and crosses for infrared modality.

4.5 Analysis of Visualization

As shown in Figure 4, we visualize t-SNE maps for 10 randomly selected identities from SYSU-MM01. Compared to the baseline, the same identities from different modalities are closer together, demonstrating the effectiveness of the proposed framework. The visualization of similarity distribution is presented in the supplementary materials.

5 Conclusion

In this paper, we use diverse tokens to capture fine-grained pedestrian features and apply them for cross-modality label matching and neighbor learning to effectively address large discrepancies. Additionally, to handle cases where the same pedestrian may split into different clusters with the same ID, we propose a correlation algorithm and a homogeneous fusion module to further enhance the model's robustness. Extensive experiments on two public benchmarks demonstrate that our method outperforms existing approaches.

Acknowledgments

This work is partially supported by National Natural Science Foundation of China under Grant (62176188, 62302351, 62376201, 62066021), Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (GZC20241268, 2024M762479), and Hubei Advanced Postdoctoral Talent Programme (2004HBBHJD070). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

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

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

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

Choi, S.; Lee, S.; Kim, Y.; Kim, T.; and Kim, C. 2020. Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 10254–10263. Computer Vision Foundation / IEEE.

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

Delorme, G.; Xu, Y.; Lathuilière, S.; Horaud, R.; and Alameda-Pineda, X. 2020. CANU-ReID: A Conditional Adversarial Network for Unsupervised person Re-IDentification. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, 4428–4435. IEEE.

Ding, C.; Wang, K.; Wang, P.; and Tao, D. 2022. Multi-Task Learning With Coarse Priors for Robust Part-Aware Person Re-Identification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(3): 1474–1488.

Du, B.; Du, C.; and Yu, L. 2023. Megf-net: multi-exposure generation and fusion network for vehicle detection under dim light conditions. *Visual Intelligence*, 1(1): 28.

Feng, Z.; Lai, J.; and Xie, X. 2020. Learning Modality-Specific Representations for Visible-Infrared Person Re-Identification. *IEEE Trans. Image Process.*, 29: 579–590.

Fu, Y.; Wei, Y.; Wang, G.; Zhou, Y.; Shi, H.; and Huang, T. S. 2019. Self-Similarity Grouping: A Simple Unsupervised Cross Domain Adaptation Approach for Person Re-Identification. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 6111–6120. IEEE.

Gao, Y.; Liang, T.; Jin, Y.; Gu, X.; Liu, W.; Li, Y.; and Lang, C. 2021. MSO: Multi-feature space joint optimization network for RGB-infrared person re-identification. In *Proceedings of the 29th ACM international conference on multimedia*, 5257–5265.

Hu, Z.; Yang, B.; and Ye, M. 2024. Empowering Visible-Infrared Person Re-Identification with Large Foundation Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Huang, N.; Liu, J.; Miao, Y.; Zhang, Q.; and Han, J. 2023. Deep learning for visible-infrared cross-modality person re-identification: A comprehensive review. *Information Fusion*, 91: 396–411.

Kansal, K.; Subramanyam, A. V.; Wang, Z.; and Satoh, S. 2020. SDL: Spectrum-Disentangled Representation Learning for Visible-Infrared Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.*, 30(10): 3422–3432.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Kim, M.; Kim, S.; Park, J.; Park, S.; and Sohn, K. 2023. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18621–18632.

Kim, S.; Gwon, S.; and Seo, K. 2024. Enhancing diverse intra-identity representation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2513–2522.

Li, P.; Wu, K.; Huang, W.; Zhou, S.; and Wang, J. 2024. Camera-aware Label Refinement for Unsupervised Person Re-identification. *arXiv preprint arXiv:2403.16450*.

Li, W.; Zou, C.; Wang, M.; Xu, F.; Zhao, J.; Zheng, R.; Cheng, Y.; and Chu, W. 2023. Dc-former: Diverse and compact transformer for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1415–1423.

Li, Y.; Yang, F.; Liu, Y.; Yeh, Y.; Du, X.; and Wang, Y. F. 2018. Adaptation and Re-Identification Network: An Unsupervised Deep Transfer Learning Approach to Person Re-Identification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 172–178. Computer Vision Foundation / IEEE Computer Society.

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

Liu, F.; Ye, M.; and Du, B. 2024. Learning a generalizable re-identification model from unlabelled data with domain-agnostic expert. *Visual Intelligence*, 2(1): 28.

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

Lu, A.; Zhang, Z.; Huang, Y.; Zhang, Y.; Li, C.; Tang, J.; and Wang, L. 2024. Illumination Distillation Framework for Nighttime Person Re-Identification and a New Benchmark. *IEEE Trans. Multim.*, 26: 406–419.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

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

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

Peng, Y.; Li, Y.; and Zheng, W. 2024. Revisiting Person Re-Identification by Camera Selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5): 2692–2708.

Qian, Z.; Lin, Y.; and Du, B. 2023. Visible-infrared person re-identification via patch-mixed cross-modality learning. *arXiv preprint arXiv:2302.08212*.

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

Wang, M.; Lai, B.; Huang, J.; Gong, X.; and Hua, X.-S. 2021. Camera-aware proxies for unsupervised person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2764–2772.

Wang, S.; Xu, X.; Chen, H.; Jiang, K.; Wang, Z.; and Tang, K. 2024. Low-Light Salient Object Detection Meets the Small Size. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 1–13.

Wang, S.; Xu, X.; Liu, L.; and Tian, J. 2020. Multi-level feature fusion model-based real-time person re-identification for forensics. *Journal of Real-Time Image Processing*, 17(1): 73–81.

Wang, S.; Xu, X.; Ma, X.; Jiang, K.; and Wang, Z. 2023. Informative Classes Matter: Towards Unsupervised Domain Adaptive Nighttime Semantic Segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 163–172. Ottawa, ON, Canada: Association for Computing Machinery.

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.; and Satoh, S. 2019. Learning to Reduce Dual-Level Discrepancy for Infrared-Visible Person Re-Identification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 618–626. Computer Vision Foundation / IEEE.

Wu, A.; Lin, C.; and Zheng, W.-S. 2024. Asymmetric Mutual Learning for Unsupervised Transferable Visible-Infrared Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*.

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

Wu, Q.; Dai, P.; Chen, J.; Lin, C.-W.; Wu, Y.; Huang, F.; Zhong, B.; and Ji, R. 2021. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4330–4339.

Wu, Z.; and Ye, M. 2023. Unsupervised visible-infrared person re-identification via progressive graph matching and alternate learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9548–9558.

Xu, B.; He, L.; Liao, X.; Liu, W.; Sun, Z.; and Mei, T. 2020. Black Re-ID: A Head-shoulder Descriptor for the Challenging Problem of Person Re-Identification. In Chen, C. W.; Cucchiara, R.; Hua, X.; Qi, G.; Ricci, E.; Zhang, Z.; and Zimmermann, R., eds., *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, 673–681. ACM.

Xu, X.; Wang, S.; Wang, Z.; Zhang, X.; and Hu, R. 2021. Exploring image enhancement for salient object detection in low light images. *ACM transactions on multimedia computing, communications, and applications*, 17(1s): 1–19.

Yan, Z.; Zheng, Y.; Fan, D.-P.; Li, X.; Li, J.; and Yang, J. 2024. Learnable differencing center for nighttime depth perception. *Visual Intelligence*, 2(1): 15.

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

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

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

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

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

Yang, F.; Weng, J.; Zhong, Z.; Liu, H.; Wang, Z.; Luo, Z.; Cao, D.; Li, S.; Satoh, S.; and Sebe, N. 2023c. Towards Robust Person Re-Identification by Defending Against Universal Attackers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4): 5218–5235.

Ye, M.; Chen, S.; Li, C.; Zheng, W.-S.; Crandall, D.; and Du, B. 2024. Transformer for object re-identification: A survey. *International Journal of Computer Vision*, 1–31.

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

Ye, M.; Wu, Z.; Chen, C.; and Du, B. 2023. Channel augmentation for visible-infrared re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhang, Q.; Lai, C.; Liu, J.; Huang, N.; and Han, J. 2022. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7349–7358.

Zhang, Y.; and Wang, H. 2023. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2153–2162.

Zhong, X.; Lu, T.; Huang, W.; Ye, M.; Jia, X.; and Lin, C.-W. 2021. Grayscale enhancement colorization network for visible-infrared person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1418–1430.