remote sensing

MDPI

Article

A Re-Identification Framework for Visible and Thermal-Infrared Aerial Remote Sensing Images with Large Differences of Elevation Angles

Chunhui Zhao [1], Wenxuan Wang [1], Yiming Yan [1,2,*], Baoyu Ge [1,2], Wei Hou [1,2] and Fengjiao Gao [3]

[1] Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information, College of Information and Communication Engineering, Harbin Engineering University, Harbin 150009, China; zhaochunhui@hrbeu.edu.cn (C.Z.); wangwenxuan@hrbeu.edu.cn (W.W.); gebaoyu@spacestar.com.cn (B.G.); houwei@spacestar.com.cn (W.H.)
[2] Harbin Space Star Data System Science and Technology Co., Ltd., Harbin 150080, China
[3] Intelligent Manufacturing Research Institute, Heilongjiang Academy of Sciences, Harbin 150001, China; gaofengjiao@haai.org.cn
* Correspondence: yanyiming@hrbeu.edu.com

Abstract: Visible and thermal-infrared re-identification (VTI-ReID) based on aerial images is a challenging task due to the large range of elevation angles, which exacerbates the modality differences between different modalities. The substantial modality gap makes it challenging for existing methods to extract identity information from aerial images captured at wide elevation angles. This limitation significantly reduces VTI-ReID accuracy. This issue is particularly pronounced in elongated targets. To address this issue, a robust framework for extracting identity representation (RIRE) is proposed, specifically designed for VTI-ReID in aerial cross-modality images. This framework adopts a mapping method based on global representation decomposition and local representation aggregation. It effectively extracts features related to identity from aerial images and aligns the global representations of images captured from different angles within the same identity space. This approach enhances the adaptability of the VTI-ReID task to elevation angle differences. To validate the effectiveness of the proposed framework, a dataset group for elongated target VTI-ReID based on unmanned aerial vehicle (UAV)-captured data has been created. Extensive evaluations of the proposed framework on the proposed dataset group indicate that the framework significantly improves the robustness of the extracted identity information for elongated targets in aerial images, thereby enhancing the accuracy of VTI-ReID.

Keywords: UAV remote sensing; cross-modality re-identification; visible and thermal remote sensing; elevation angle differences

Academic Editors: Dongyu Li, Rui Li, Yinxia Cao and Haoyang Yang

Received: 17 April 2025
Revised: 10 May 2025
Accepted: 2 June 2025
Published: 5 June 2025

Citation: Zhao, C.; Wang, W.; Yan, Y.; Ge, B.; Hou, W.; Gao, F. A Re-Identification Framework for Visible and Thermal-Infrared Aerial Remote Sensing Images with Large Differences of Elevation Angles. *Remote Sens.* **2025**, *17*, 1956. https://doi.org/10.3390/rs17111956

1. Introduction

Visible and thermal-infrared re-identification (VTI-ReID) aims to match and associate data of the same target captured by multiple cross-modality cameras across different times and locations to re-identify target identities [1–6]. Due to its significance in all-weather surveillance and public security, it has garnered increasing attention in recent years. Unmanned Aerial Vehicles (UAVs) overcome the limitations of traditional fixed surveillance sensors and provide enhanced monitoring capabilities [7–11]. However, aerial images often exhibit wide variations in perspective angles, which significantly exacerbates the visual representation discrepancies between visible and thermal-infrared modalities [12–17], as illustrated in Figure 1. For instance, a vessel target captured from a top-down perspective

may emphasize roof features (e.g., deck structures), while a side-view angle highlights lateral characteristics (e.g., hull contours), leading to substantial differences in shape, texture, and other visual attributes. This results in insufficient identity-discriminative information for recognition, particularly critical for slender targets such as humans and vessels [18–21]. Existing VTI-ReID methods, primarily designed for ground-based fixed-view scenarios, struggle to address the cross-modality discrepancy challenges exacerbated by large elevation angle variations in aerial contexts.

Figure 1. The aerial images exhibit significant differences in elevation angles, which are more pronounced in elongated targets such as humans and vessels: (**a**) shows person data captured by traditional fixed surveillance cameras, while (**b**) displays aerial Person surveillance data captured by UAVs, demonstrating a wider range of elevation angles in the aerial person images. The comparison between (**c**,**d**) showing significant differences in elevation angles of vessel targets in aerial images.

Specifically, existing research has two main limitations: the first type of method synthesizes cross-modality images through generative adversarial networks (GANs) to align feature spaces [22–28], but due to the lack of corresponding image pairs, significant changes in target perspectives in aerial scenes can lead to severe noise in the synthesized images. The second type of method attempts to map global features to a shared embedding space [29–31], but large elevation angle variations can disrupt the stability of the target global features, and relying solely on cross-modality global features can make accurate mapping challenging. Given this issue, we attempt to investigate a VTI-ReID framework for aerial images with large elevation angle variations.

Therefore, a robust VTI-ReID framework for extracting identity representation (Robust Identity Representation Extraction framework, RIRE) is proposed. This framework adopts a mapping method based on global representation decomposition and local representation aggregation to extract features related to target identity from aerial images, with the intention of mapping the global representations of images with different elevation angles to the same identity space, thereby enhancing the adaptability of VTI-ReID to elevation angle

differences. To verify the effectiveness of our framework, we constructed an aerial VTI-ReID dataset group for person and vessels that are greatly affected by elevation angle changes. Specifically, the dataset group comprises the aerial human VTI-ReID dataset Beach Aerial Infrared Visible Person dataset (BA-VIP) and the aerial vessels dataset AVC-ReID [32]. The extensive evaluation of the proposed dataset group shows that our framework significantly improves the robustness of the identity information in aerial image VTI-ReID, especially under high elevation angle variations.

In summary, the main contributions are as follows:

- A VTI-ReID framework for aerial images with large elevation angle differences has been proposed.
- To alleviate large elevation angle differences, a method for identity-feature extraction based on global representation decomposition and local feature aggregation is proposed.
- To verify the effectiveness of the proposed framework, a dataset group consisting of aerial visible and thermal-infrared images was proposed. Extensive experiments have shown that the RIRE framework provides relatively reliable cross-modality identity correspondence, and our proposed framework has significantly improved the accuracy of aerial VTI-ReID.

2. Related Works

Since the introduction of the VTI-ReID task in [33], methods for VTI-ReID have experienced rapid development. However, most cross-modality Re-ID datasets currently focus on near-infrared images, which, although cost-effective, do not possess the reflective thermal characteristics of thermal-infrared images. Therefore, research on Re-ID based on visible and thermal-infrared images is of the same importance. The significant feature differences between visible and thermal-infrared images of the same target from different perspective angles make it challenging to project cross-modality images into a shared space directly. Existing methods typically use "feature-level methods" or "image-level methods" to map target images from different perspectives into the same feature space. However, these methods are ineffective in addressing the significant elevation angle differences in aerial images, especially on elongated targets. In the following sections, these two methods will be introduced separately.

2.1. Image-Level Methods

In an attempt to lessen the modality disparity between VIS and IR images in the image space, image-level approaches convert one modality into another. Some methods enhance the network's insensitivity to modalities by fusing images from different modalities, making it more sensitive to the information in the image that can be used for discriminating identities. To bridge the modality gap, Wang et al. [34] suggested an end-to-end alignment generative adversarial network (AlignGAN) that integrates feature alignment with pixel alignment. Wang et al. [35] introduced a Dual-Domain Difference Reduction Learning (D2RL) scheme that handles appearance and different modalities separately through a modality transformation network. Hao et al. [36] introduced an identity-aware marginal center aggregation strategy to extract centralized features while maintaining diversity under the constraint of marginality. Fusion Pattern Collaborative Learning (SMCL) is a model that Wei et al. [37] proposed. It combines the challenges of auxiliary distribution similarity learning (ADSL) and enhanced homogeneity learning (CEHL) to automatically construct a new modality with heterogeneous image features. The model projects heterogeneous features onto a unified space to enhance recognition ability and amplify inter-class differences. Ye et al. [38] uniformly generate color-insensitive images by randomly swapping color channels, continuously improving the model's robustness to color variations. These

techniques for aligning cross-modality images frequently entail sophisticated generative models. Although successful, generating cross-modality images inevitably comes with noise due to the lack of VIS-IR image pairs. Recently, X-modality [39] and its variants (such as SMCL [37] and MMN [40]) have utilized lightweight networks to obtain auxiliary modalities to assist in cross-modality search. Nonetheless, a modality gap persists between the VIS-IR modality and these auxiliary modalities. These problems are the reason that such methods do not achieve very ideal accuracy at present.

Furthermore, Kim et al. [41] introduced a data augmentation technique tailored for VTI-ReID. This technique enhances the performance of part-based VTI-ReID models by synthesizing augmented samples through the fusion of part descriptors from various modalities. This method provides a new idea for solving the problem of cross-modality Re-ID at the data enhancement level.

2.2. Feature-Level Methods

Feature-level methods aim to learn discriminative representations from cross-modality data for identity recognition. They can be further divided into network-based approaches and metric-based approaches.

Network-based approaches aim to find a modality-shared feature space where the modality gap can be minimized. A dynamic dual-attention aggregation learning approach was presented by Ye et al. [42] that simultaneously examines cross-modality contextual clues and intra-modality hierarchy clues in VTI-ReID. A hierarchical cross-modality disentanglement technique was presented by Choi et al. [43]. It uses just the id-discriminative factor for robust cross-modality matching, automatically separating the id-discriminative and id-exclusion factors from visible thermal images. A cross-modality-specific feature transfer technique was introduced by Lu et al. [44]. It predicts the affinity of various modality samples based on common characteristics and then transfers modality-specific features as well as shared features between and within modalities. To alleviate modality discrepancy, Wu et al. [45] proposed a combination modality and pattern alignment network that extracts discriminative features by utilizing both modality alleviation and pattern alignment modules. To learn cross-modality measurements in two ways, Liu et al. [46] improved them even more using memory-based augmentation. A diversified embedding expansion network was presented by Zhang et al. [47], it efficiently creates varied embeddings to learn useful feature representations and lessen modality disparity. To separate the correlated modality-shared information in two orthogonal subspaces, Feng et al. [48] suggested a shape-erasure feature learning paradigm that involves simultaneously learning shape-related features in one subspace and shape-erased features in the orthogonal counterpart. Through the optimization of the conditional mutual information between erased features and excluded body shape information, this method directly increases the diversity of learned representations.

Metric-based approaches aim to reduce the distance between different modality data in the feature space by designing loss functions. Ye et al. [49] proposed a dual-path network with a novel bidirectional dual-constraint top-level loss to learn discriminative feature representations, simultaneously handling variations between modalities and within modalities to ensure the discriminability of the learned representations. Hao et al. [50] mapped the extracted features onto a hypersphere manifold, where the difference between two samples was computed based on angles. In contrast to angle-based measurements in [34], Feng et al. [51] used Euclidean constraints to narrow the cross-modality gap. Subsequently, by adding a bidirectional center-constrained ranking loss, Ye et al. [52] expanded on this and enhanced the effectiveness of image-based person re-identity. A cross-center loss was introduced by Tan et al. [53] to investigate more compact intra-class

distributions. A loss paradigm known as sparse pairwise loss was presented by Zhou et al. [54]. It combines an adaptive positive mining technique to dynamically adjust to various intra-class variances. However, due to significant feature differences between VIS and IR images from different observation perspectives, it is difficult to directly project cross-modality images into a shared space. Therefore, it is necessary to design a more targeted network architecture and effective constraints to improve the accuracy and identity robustness of cross-modality multi-perspective image feature mapping.

3. Datasets

To evaluate the effectiveness of our proposed VTI-ReID framework for cross-modality aerial data with large elevation angle differences, we have produced a visible and thermal-infrared elongated target dataset group, which includes two types of typical elongated targets: humans and vessels, which are named the BA-VIP dataset and AVC-ReID dataset, respectively. Table 1 tabulates the comparison between the proposed dataset group and existing related cross-modality Re-ID datasets; the range of elevation angle is evaluated by visually observing the images in the dataset, where the bolded dataset represents the dataset group we used. Next, the proposed dataset group is introduced in detail.

Table 1. Comparison between the proposed dataset group and existing related image Re-ID datasets.

Datasets	Types of Targets	Range of Elevation Angle	Modality	Identity	Visible	Infrared
RegDB [55]	Person	0–20/160–180	Visible and Thermal-infrared	412	4120	4120
SYSU-MM01 [34]	Person	0–10/170–180	Visible and Near-infrared	491	26,061	12,210
LLCM [47]	Person	0–20/160–180	Visible and Near-infrared	1064	25,626	21,141
AVC-ReID [32]	Vessel	10–170	Visible and Thermal-infrared	138	3071	2530
BA-VIP (Ours)	Person	60–120	Visible and Thermal-infrared	205	43,843	26,795

3.1. Beach Aerial Infrared Visible Person Dataset

In this paper, we constructed the BA-VIP dataset with aerial images for human VTI-ReID. To our knowledge, this is the first visible and thermal-infrared person Re-ID dataset captured by UAV. BA-VIP provides richer observation angles and higher difficulty in VTI-ReID. The BA-VIP dataset was captured and compiled in May 2023 at the First Sea Bathing Beach in Yantai City, Shandong Province, China, using the DJI M300RTK [56] UAV paired with the H20T sensor [57], which made by Shenzhen DJI Innovation Technology Co., Ltd. in China. The UAV flew at an altitude of 15–20 m above sea level, with wide-angle lenses used to capture visible light images and dual digital zoom infrared lenses used to capture thermal-infrared images, with a sensor elevation angle range of 60–120 degrees.

In detail, our BA-VIP dataset consists of 43,843 visible images and 26,795 thermal-infrared images of 205 unique identities. These images were extracted from 446 video segments captured by the UAV during 24 take-off and landing sessions, with frames sampled at 5-frame intervals. The same person may appear in different videos captured during various take-offs and landings. To ensure data quality, persons identified in the sensor were manually annotated and images were retained if their row pixels were greater than 50, column pixels were greater than 50, or total pixels were greater than 1200 because we have found that under current conditions, it can be guaranteed that there will be no problem of low infrared image quality for the person due to too far distance. Our dataset is entirely manually annotated by humans, avoiding label noise.

The BA-VIP dataset was captured by UAV in a seaside amusement area, containing rich variations in person poses and observation perspectives. For thermal-infrared images, the pseudo-color represents the relative temperature of different targets in a single image; therefore, the person targets appear whitish or blackish in the infrared image depending

on the contrast between the person's temperature and the environmental temperature. Therefore, based on the ground temperature, the background can be classified into three types: moist beach, dry beach, and sidewalk. In the wet beach background, the person targets in the thermal-infrared image appear whitish, in the dry beach background, the person targets in the thermal-infrared image are similar in color to the background or appear blackish, and in the sidewalk background, the person targets in the thermal-infrared image appear blackish, as shown in Figure 2.

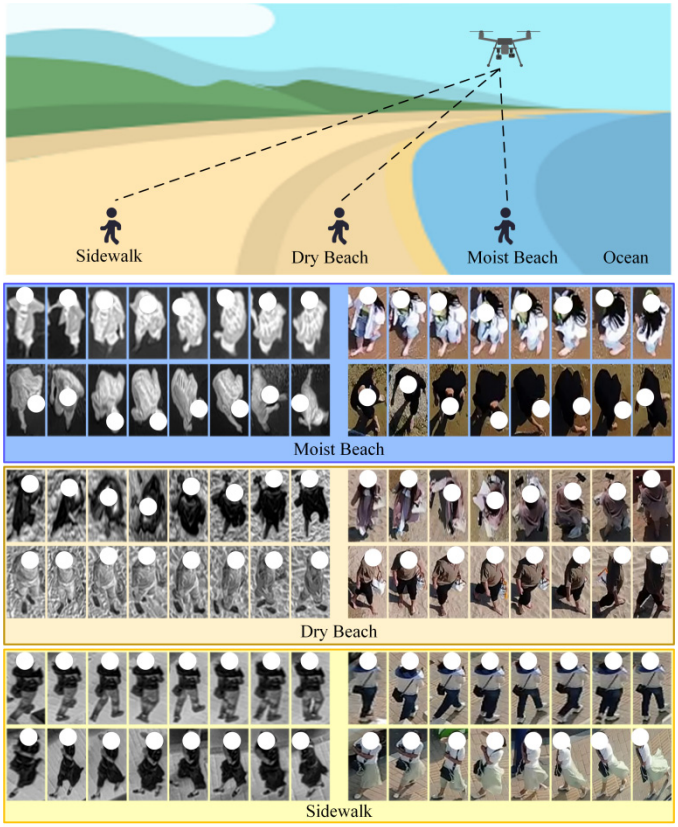

Figure 2. Schematic diagram of the influence of different backgrounds on the person thermal-infrared and visible images in the BA-VIP dataset.

It is worth mentioning that infrared images are usually divided into long-wave infrared (LWIR), medium-wave infrared (MWIR), and short-wave infrared (SWIR) based on their wavelengths. LWIR and MWIR are commonly referred to as thermal-infrared, which generate thermal-infrared images by detecting the heat emitted by an object, displaying its surface temperature. Thermal-infrared images excel in capturing object heat in complete darkness or other lighting conditions, making them advantageous in nighttime, low light, or no light conditions, particularly for warm-blooded objects like humans. The DJI H20T utilizes an uncooled vanadium oxide (VOx) microbolometer to capture thermal-infrared images at a wavelength range of 8–14 μm [57]. Among existing VTI-ReID datasets, except for the RegDB dataset initially not designed for VTI-ReID tasks, our BA-VIP dataset is the only dedicated remote sensing thermal-infrared person dataset created specifically for VTI-ReID tasks.

However, despite the BA-VIP dataset filling the gap in VTI-ReID data from UAV perspectives and demonstrating significant value in advancing research on UAV-based VTI-ReID, it still exhibits several limitations:

(1) (Lack of quantitative observation angle data: Due to the dependency of person imaging angles on UAV flight altitude, UAV-target distance, and other factors, the

BA-VIP dataset currently fails to provide precise quantitative angular measurements. Consequently, it is challenging to analyze the specific impact of elevation angle variations on model performance, which restricts the design of optimization strategies tailored to angular discrepancies.

(2) Limited scene diversity: Constrained by its acquisition environment, the dataset only covers beach and sidewalk scenarios, lacking data from complex urban environments, nighttime conditions, or other geographic regions.

(3) Data scale and imbalance issues: With only 205 identities and an imbalanced distribution between visible (43,843 images) and thermal-infrared (26,795 images) modalities, certain identities may suffer from insufficient samples. This imbalance limits the depth of model training and may lead to inadequate feature learning for low-frequency identities.

In summary, future enhancements are essential to further elevate the dataset's research value and applicability in real-world UAV-based VTI-ReID tasks, such as data augmentation, precise annotation, and multi-scenario expansion.

3.2. Airborne Vessel Cross-Modality Re-Identification Dataset

The AVC-ReID dataset consists of visible and thermal-infrared modality data captured by UAV from 138 vessels, comprising a total of 3071 visible images and 2530 thermal-infrared images, with each identity represented by at least 15 images in each modality. Sample images from the dataset are illustrated in Figure 3. As another visible and thermal-infrared dataset captured by UAV, it exhibits significant variations in resolution and changes in observation perspectives [32].

Figure 3. Schematic diagram of vessel dataset AVC-ReID obtained by UAV from different observation perspectives.

4. Methodology

In this section, we will provide a detailed introduction to our proposed VTI-ReID framework named RIRE for aerial cross-modality images. Figure 4 shows the overview of our framework. To cope with the modality differences of different modality data, the network uses ResNet50 as the backbone, which has the other stage's shared structure and parameters, but not the first stage's parameters. Due to the significant differences in cross-modality image features under different observation perspectives of the target, the network uses the Aggregation of Deep and Shallow Global Features (DSGFAs) module and Global Representation Mapping (GRM) module based on local representation extraction and aggregation to adaptively adjust the feature extracted from different observation perspectives. By extracting identity-related information unrelated to the perspective from images of varying observation perspectives, the network enhances its ability to recognize the target's images from multiple observation perspectives and improves feature consistency among individuals with different modalities of images from different observation perspectives.

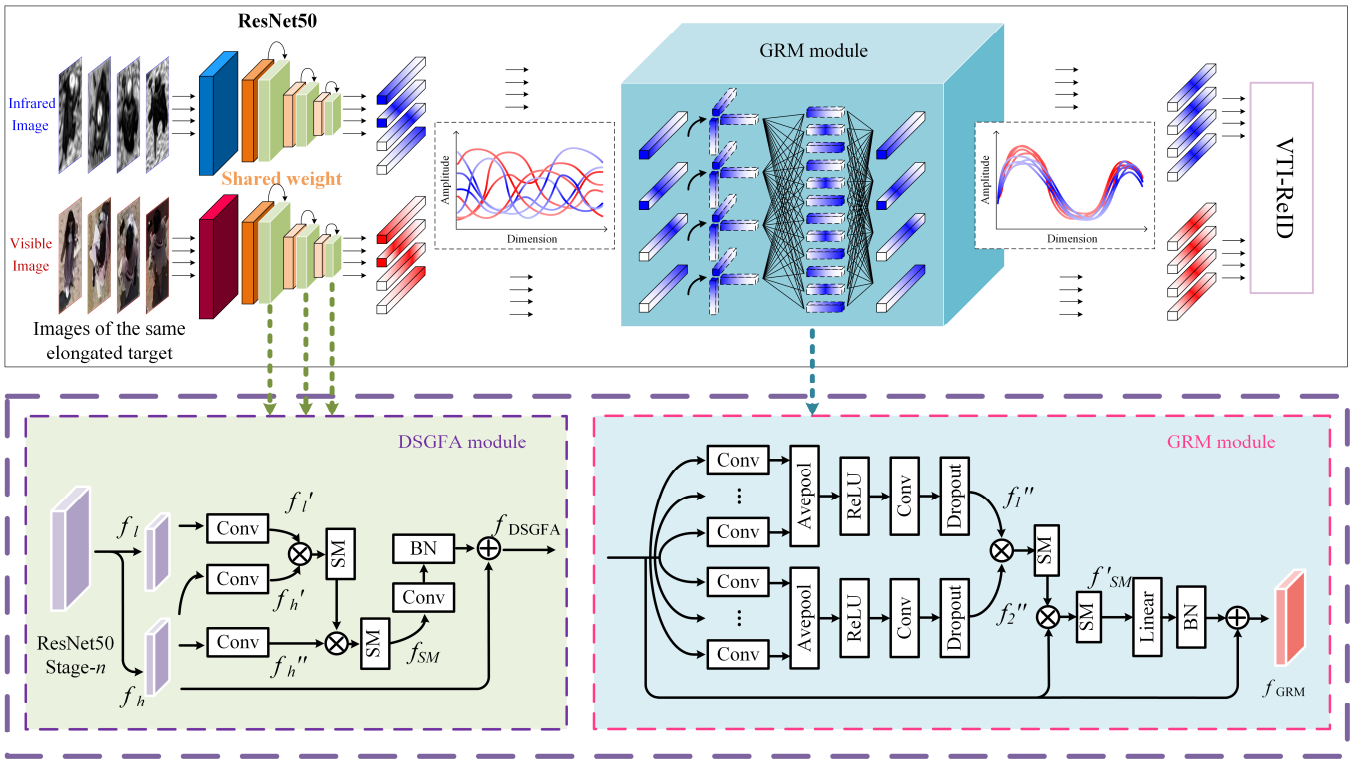

Figure 4. Schematic diagram of RIRE framework's structure. Blue represents the infrared branch, red represents the visible branch, light green squares represent the DSGFA module, and light blue squares represent the GRM module.

4.1. Aggregation of Deep and Shallow Global Features Module (DSGFA)

With the increase in network depth, information needs to be propagated and extracted through multiple layers; however, this process may lead to issues such as gradient vanishing and information loss, resulting in a decrease in network performance. Furthermore, the enlargement of receptive field may cause imprecise extraction of detailed information. Therefore, the DSGFA module is proposed. The function of the DSGFA module is to perform adaptive aggregation of deep and shallow features to ensure that identity-related information is not lost during the convolutional process. The DSGFA layer takes both the input and output of its preceding part of the ResNet50 network (STAGE-*n*) as input, where the input from STAGE-*n* provides low-level feature maps $f_l \in R^{C_l \times H_l \times W_l}$ to the DSGFA layer, while the output from STAGE-*n* provides high-level feature maps $f_h \in R^{C_h x H_h x W_h}$. Here, *C*, *W*, and *H* represent the number of channels, width, and height of the features, respectively. As high-level features typically have better information representation capabilities, the focus of self-feature aggregation is mainly on these high-level features.

In the DSGFA module, the low-level feature f_l is passed through a 1×1 convolutional block $\varphi^1_{1\times1}$, resulting in $f_l{}'$. The high-level feature f_h is separately passed through two 1×1 convolutional blocks $\varphi^2_{1\times1}$ and $\varphi^3_{1\times1}$, resulting in $f_h{}'$ and $f_h{}''$, respectively, as shown in (1).

$$\begin{aligned} f_l{}' &= \varphi^1_{1\times1}(f_l) \\ f_h{}' &= \varphi^2_{1\times1}(f_h) \\ f_h{}'' &= \varphi^3_{1\times1}(f_h) \end{aligned} \tag{1}$$

Then, the matrix multiplication operation is performed between $f_l{}'$ and $f_h{}''$. To prevent the inner product from becoming too large, the result is divided by the dimension of the embedding. This step is similar to softmax, except that it does not force the output to be constrained between 0 and 1; therefore, we use *SM* to represent this step, as shown in (2).

Among (2), **R** represents the feature matrix, and $m \times n$ represent the size of the feature matrix. Subsequently, the result is multiplied by f_h'' again, and the same normalization by the embedding dimension of a is applied, as shown in (3).

Finally, the obtained result is passed through a 1×1 convolutional block $\varphi^4_{1\times1}$ and a batch normalization layer *BN*. This yields the output f_{DSGFA} of the DSGFA module. The output of the DSGFA module is shown in (4):

$$SM : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}, SM(\mathbf{R}) = \begin{cases} \frac{1}{m}\mathbf{R}, \text{if } m \geq n \\ \frac{1}{n}\mathbf{R}, \text{if } m < n \end{cases} \tag{2}$$

$$f_{SM} = SM(SM(f_l{}' \times f_h{}') \times f_h{}'')) \tag{3}$$

$$f_{DSGFA} = BN(\varphi^4_{1\times1}(f_{SM})) \tag{4}$$

4.2. Global Representation Mapping Module (GRM)

The GRM module decomposes the target global representation extracted from cross-modality images to obtain local representations, and then adaptively aggregates the local representations to filter out target identity irrelevant information contained in the global representation, thereby enhancing the network's ability to recognize the identity of person images from different UAV observation perspectives. Specifically, a multi-branch convolutional structure is used to diversify and aggregate the input feature f_{DSGFA} in GRM. The structure of the *i*-th branch is as follows: first, *p* separates convolution blocks, namely, ${\varphi^1}_{3\times3}$,..., ${\varphi^p}_{3\times3}$, are used to fuse the input feature *f* and extract potential embedding information, as shown in (5).

Then, the F_{ReLU} activation layer is applied to provide the embedding exploration network with non-linear representation capability. Subsequently, a 1×1 convolutional block $\varphi_{1\times1}$ is used to map the virtual embeddings, and finally, a dropout module *D* is applied to improve the generalization performance of the virtual embedding exploration network, as shown in (6); *p* represents the probability that the features at the current position are set to zero, and we take *p* equal to 0.01. The output of the *i*-th branch is shown in (7). Here, we use orthogonal loss to constraints between different feature branches f_i'':

$$f_i{}' = [\phi^{i\_1}_{3\times3}(f) + \ldots\ldots + \phi^{i\_p}_{3\times3}(f)]/p \tag{5}$$

$$D : f_{m\times n}{}' = p f_{m\times n} \tag{6}$$

$$f_i{}'' = D(\phi_{1\times1}(F_{ReLU}(f_i{}'))) \tag{7}$$

In the aggregation process after diversified feature representation in the GRM, the virtual embeddings f_1'' and f_2'' are first subjected to matrix multiplication. To prevent the inner product from becoming too large, the result is divided by the dimension of the embeddings by *SM*. Then, the result is multiplied by *f*, and the same normalization by the embedding dimension is applied, as shown in (8). Afterwards, a linear mapping layer *L* and a batch normalization layer *BN* are employed. Finally, the result is added to *f*, resulting in f_{GRM}, which represents the aggregation of features. The complete output of the GRM layer is shown in (9):

$$f_{SM}{}' = SM(SM(f_1'' \times f_2'') \times f_{\mathrm{DSGFA}}) \tag{8}$$

$$f_{\mathrm{GRM}} = f_{\mathrm{DSGFA}} + BN(L(f_{SM}{}')) \tag{9}$$

After passing through the GRM module, the features will be used to calculate the loss function and perform similarity measurement to obtain VTI-ReID results. In our method, unless otherwise specified, we use cross entropy loss L_{cro} [58], triplet loss L_{tri} [59],

center guided pair mining loss L_{cen} [47] and orthogonal loss L_{orth} as the basic loss functions. In the proposed framework, L_{cro} denotes the cross-entropy loss between features and labels, L_{tri} represents the triplet loss, L_{cen} minimizes the distance between the original embedding and the generated embedding by cross-modality alignment and intra-class constraints to reduce modality differences [47]. L_{orth} enforces orthogonality constraints during global representation mapping, ensuring that distinct sub-features encapsulate mutually independent information. The total loss function is shown in (10). Among (10), λ_{cro}, λ_{tri}, λ_{cen} and λ_{orth}, respectively represent the weights of different loss functions in the total loss function:

$$L = \lambda_{cro} L_{cro} + \lambda_{tri} L_{tri} + \lambda_{cen} L_{cen} + \lambda_{orth} L_{orth} \tag{10}$$

5. Results

This section will demonstrate the effectiveness of our framework through experiments. First, the parameter settings and assessment measures used in our study are initially introduced in this section, and then the progressiveness of our method by comparing the accuracy with other advanced algorithms is shown. Next, we validate the effectiveness of each part of our method through ablation experiments. Finally, we visualize the VTI-ReID results and feature distribution, once again proving the effectiveness of our method. In this section, we provide a detailed explanation of our experimental design and findings and finally present the results in visual form.

5.1. Implementation Details

For person targets, each input image is resized to $3 \times 384 \times 144$ before entering the network; for vessel targets, each input image is resized to $3 \times 224 \times 224$. To increase the network performance during the training phase, the techniques of random horizontal flipping and random erasure [60] are applied to the data. The initial learning rate is set to 1×10^{-2} and then it increases to 1×10^{-1} after 10 epochs with a warm-up strategy. After 10 epochs, it decreases to 0.9×10^{-1}, followed by a further decrease to 0.8×10^{-1} after another 10 epochs, and continues iteratively. Our framework was trained for a total of 100 periods. In each small batch, we randomly selected four VIS images and four IR images from six identities for training. Training was carried out using an SGD optimizer with momentum set to 0.9. In the experiment, we used Rank-1, Rank-10, Rank-20, mAP, and mINP [61] as evaluation indicators. Among them, Rank-1 represents the probability that the best search among all VTI-ReID results is the correct ID. Rank-10 and Rank-20 represent the probability that the top ten and twenty search results in all VTI-ReID results contain the correct ID image. MAP and mINP are evaluations of the distribution of correct results in all search results. In our experiment, unless otherwise specified, we use L_{cro}, L_{tri}, L_{cen}, and L_{orth} as the basic loss functions; λ_{cro}, λ_{tri}, λ_{cen}, and λ_{orth} are set to 1.0, 1.0, 0.8, and 0.01, respectively, depending on [47]. Firstly, we experimented with the impact of different p-value settings on VTI-ReID accuracy. Based on the experimental results, a p-value of 3 in (5) will be selected for the following experiments.

To comprehensively evaluate the practicality of RIRE, we analyze its computational efficiency in terms of model parameters, FLOPs, and inference speed. All tests are conducted on an NVIDIA RTX 3090 GPU with a batch size of 12. In order to comprehensively evaluate the practicality of RIRE, we analyzed its computational efficiency from the aspects of model parameters, FLOP, and inference speed. RIRE contributed 76.88 M of model parameters, with FLOPs of 21.32 GB and 24.29 G for processing standard human images ($3 \times 384 \times 144$) and vascular images ($3 \times 224 \times 224$), respectively. The time required to complete one test code is 8.5 min.

5.2. Comparison with State-of-the-Art Methods

To showcase the effectiveness of our framework, RIRE was compared with extensive state-of-the-art VTI-ReID algorithms using aerial VTI-ReID data from the proposed datasets group; the result is illustrated in Tables 2 and 3. Some of the results are referenced from [32]. The best results are highlighted in red.

Table 2. Person VTI-ReID accuracy between RIRE and the other state-of-the-art methods on the BA-VIP datasets.

Settings		BA-VIP (VIS to IR)					BA-VIP (IR to VIS)				
Method	**Year**	**Rank-1**	**Rank-10**	**Rank-20**	**mAP**	**mINP**	**Rank-1**	**Rank-10**	**Rank-20**	**mAP**	**mINP**
DDAG [42]	2020	55.85%	76.96%	81.85%	41.88%	15.13%	56.58%	73.96%	78.90%	42.20%	14.80%
CM-NAS [62]	2021	56.25%	76.35%	81.48%	44.94%	17.09%	61.45%	77.16%	81.70%	49.89%	21.23%
CAJ [38]	2021	58.71%	75.77%	81.36%	42.55%	14.97%	62.01%	78.10%	82.62%	50.22%	20.76%
DEEN [47]	2023	63.54%	81.79%	86.34%	49.96%	21.11%	72.19%	86.84%	90.22%	56.15%	23.48%
SGIEL [48]	2023	67.05%	82.16%	86.41%	53.70%	22.08%	69.15%	83.24%	86.47%	57.14%	25.78%
RIRE (Ours)		69.07%	84.01%	88.55%	59.54%	33.23%	76.62%	86.41%	88.91%	66.40%	37.52%

Table 3. Vessel VTI-ReID accuracy between RIRE and the other state-of-the-art methods on the AVC-ReID datasets.

Settings		AVC-ReID (VIS to IR)				AVC-ReID (IR to VIS)			
Method	**Year**	**Rank-1**	**Rank-10**	**Rank-20**	**mAP**	**Rank-1**	**Rank-10**	**Rank-20**	**mAP**
DDAG [42]	2020	28.10%	74.39%	90.02%	42.70%	25.26%	72.54%	89.33%	39.98%
CAJ [38]	2021	29.58%	72.50%	87.49%	43.44%	31.13%	75.87%	90.09%	45.44%
AGW [61]	2022	31.41%	78.79%	92.73%	46.54%	33.79%	80.61%	92.72%	49.04%
DART [63]	2022	43.73%	87.46%	96.48%	58.35%	45.30%	87.94%	96.32%	59.73%
DEEN [47]	2023	47.47%	92.07%	98.50%	62.58%	46.55%	89.36%	96.88%	61.00%
TMCN [32]	2024	50.40%	94.53%	99.00%	65.61%	53.09%	94.67%	99.06%	67.76%
RIRE (Ours)		52.91%	94.65%	99.30%	66.99%	59.69%	90.74%	95.87%	69.99%

Based on the results, it is evident that in the field of aerial VTI-ReID, our framework outperforms other state-of-the-art algorithms in all evaluation metrics for the VIS to IR and IR to VIS modes on the human-target dataset BA-VIP. Moreover, in the vessel-target dataset AVC-ReID, our method outperforms other advanced algorithms in all evaluation metrics for the VIS to IR mode, while achieving advantages in the key metrics of Rank-1 and mAP for the IR to VIS mode.

For aerial VTI-ReID tasks, Rank-1 and mAP are the two most important evaluation metrics because they respectively reflect the system performance in identifying the correct match at the top rank and in average precision, which is of significant value in practical applications. Specifically, due to the corresponding image pairs between different modality data in the BA-VIP dataset, mAP and mINP evaluation metrics better express the adaptability of our method to different observation perspectives. Our method achieves close to 60% mAP and over 33% mINP in the VIS to IR mode, and over 66% mAP and over 37% mINP in the IR to VIS mode, demonstrating the advancement and effectiveness of our approach. For the AVC-ReID dataset, our method achieves a Rank-1 metric close to 53% and mAP close to 67% in the VIS to IR mode, while in the IR to VIS mode, the Rank-1 metric is close to 60% and mAP close to 70%, proving the superiority and effectiveness of our method. However, in the IR to VIS mode, our method does not lead in the Rank-10 and Rank-20 metrics, which may be due to the uneven distribution of samples, or the increased model complexity amplifying noise or outliers in the training data.

5.3. Ablation Studies

We conducted in-depth research to accurately evaluate the effectiveness of each component of the proposed RIRE on the dataset group. Firstly, we removed the DSGFA module

and GRM module and conducted experiments on the BA-VIP and AVC-ReID datasets to observe the accuracy of VTI-ReID as a baseline for evaluating the contribution of RIRE to Re-ID accuracy. Subsequently, we separately added the DSGFA module and GRM module to the network and observed the changes in VTI-ReID accuracy. The experimental results verified the effectiveness of the two modules working independently. Finally, we combined the DSGFA module and GRM module and applied them to the network to evaluate whether these two modules can synergistically enhance the network's adaptability to perspective and modality differences. The overall settings, including experimental parameters, remain consistent with previous experiments. Tables 4 and 5 show the experimental results, with the highest accuracy highlighted in red. In this section, we used DEEN [47] as the baseline model.

Table 4. The influence on person VTI-ReID accuracy of each component of the proposed RIRE.

Settings (IR to VIS)		BA-VIP (VIS to IR)					BA-VIP (IR to VIS)				
DSGFA	**GRM**	**Rank-1**	**Rank-10**	**Rank-20**	**mAP**	**mINP**	**Rank-1**	**Rank-10**	**Rank-20**	**mAP**	**mINP**
		63.54%	81.79%	86.34%	49.96%	21.11%	72.19%	86.84%	90.22%	56.15%	23.48%
√		64.33%	81.75%	86.73%	50.54%	21.02%	72.59%	87.26%	91.07%	56.61%	24.20%
	√	66.86%	81.45%	85.86%	58.52%	31.78%	75.96%	87.82%	90.72%	65.54%	35.46%
√	√	69.07%	84.01%	88.55%	59.54%	33.23%	76.62%	86.41%	88.91%	66.40%	37.52%

Table 5. The influence on vessel VTI-ReID accuracy of each component of the proposed RIRE.

Settings (IR to VIS)		AVC-ReID (VIS to IR)				AVC-ReID (IR to VIS)			
DSGFA	**GRM**	**Rank-1**	**Rank-10**	**Rank-20**	**mAP**	**Rank-1**	**Rank-10**	**Rank-20**	**mAP**
		40.00%	87.67%	95.47%	55.47%	45.01%	86.18%	94.02%	58.70%
√		41.05%	89.19%	96.98%	56.86%	46.30%	87.18%	95.01%	59.66%
	√	48.72%	91.98%	98.02%	63.01%	58.12%	89.17%	97.44%	68.90%
√	√	52.91%	94.65%	99.30%	66.99%	59.69%	90.74%	95.87%	69.99%

From Tables 4 and 5, it can be seen that compared to not adding these two modules, using the DSGFA module and GRM module separately can improve the accuracy of VTI-ReID, indicating that both modules have effective contributions to the network. The checkmark in the table indicates the use of the current module in the model corresponding to the result. When two modules are added together to the network, RIRE achieves optimal performance, indicating that these two feature filtering layers are not mutually exclusive. The module enhances the adaptability of the network to differences in perspectives and modalities.

5.4. Visualization

To further validate the ability of RIRE to alleviate cross-modality image differences from different observation perspectives, we visualized the VTI-ReID results on both the BA-VIP dataset and the AVC-ReID dataset. In addition, we visualized the distribution of features extracted by the DSGFA and GRM modules based on BA-VIP.

5.4.1. Retrieval Results

To visually demonstrate the effectiveness of RIRE, we presented the retrieval results of RIRE on the BA-VIP dataset and AVC-ReID dataset. For each target's Re-ID result, the retrieval images enclosed in green boxes indicate correct matches, where the retrieved target's image matches the cross-modality image being searched for the same ID, while images enclosed in red boxes indicate incorrect retrievals, as shown in Figures 5 and 6. Experimental visualization shows that our method improves the probability of the first retrieval result being correct (Rank-1) and the overall ability to retrieve all correct results (mAP and mINP). Comparing the correct results identified by our method with the baseline,

the significant difference in observation perspectives does seriously affect the accuracy of VTI-ReID. Our method can enhance the adaptability of the VTI-ReID network to perspective differences; therefore, our method has improved in various evaluation indicators.

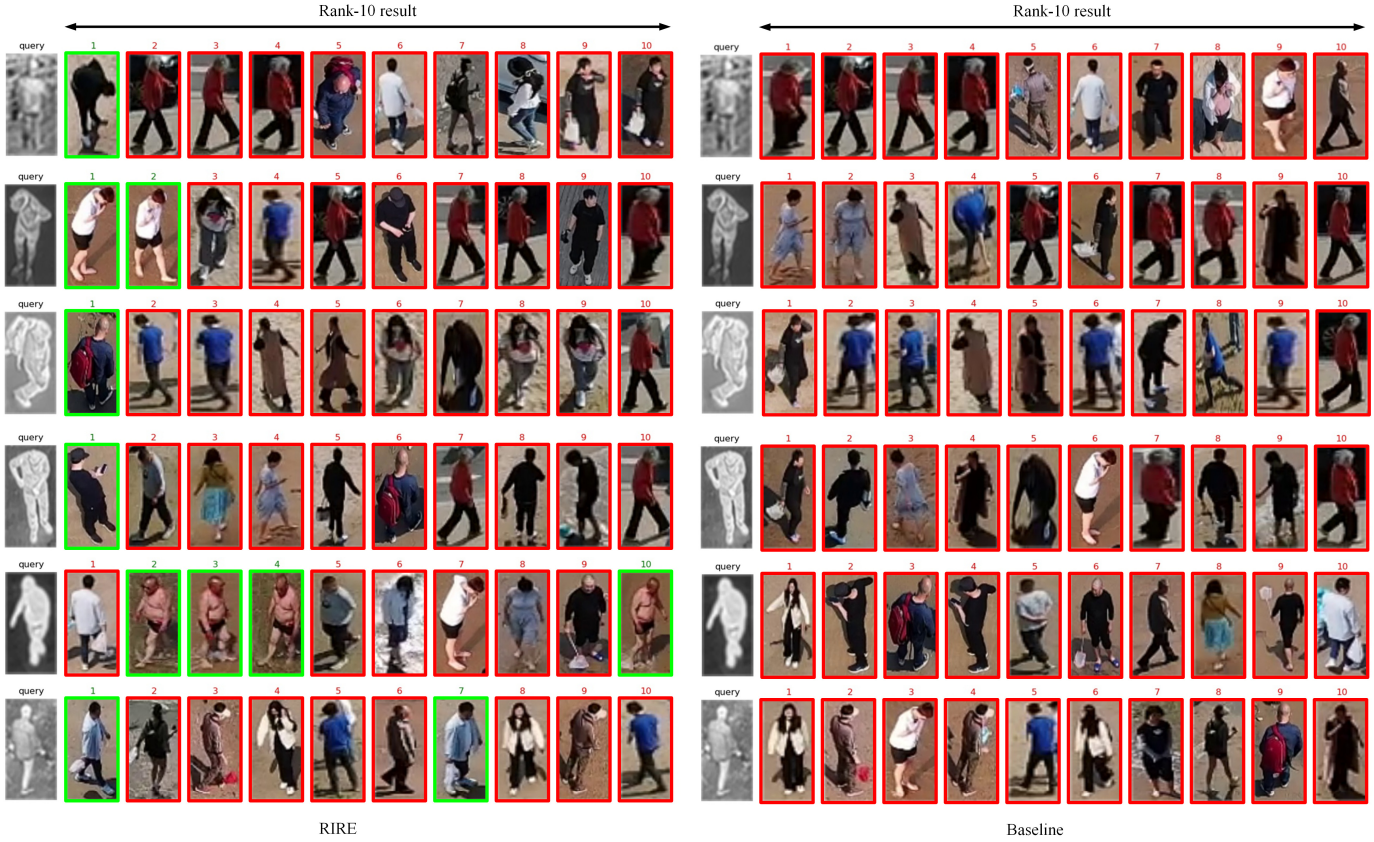

Figure 5. The Rank-10 retrieval results obtained by the proposed RIRE and baseline on the BA-VIP dataset. The red box represents a search error, and the green box represents a correct search.

Specifically, in the visualized results of the person dataset presented in Figure 5, RIRE successfully retrieves person images with perspective variations that baseline methods failed to match, demonstrating its adaptability to perspective differences. However, in the 5th-row retrieval results of Figure 5, although RIRE correctly matches targets in the 2nd to 4th ranks, the top-ranked retrieval and subsequent results contain distractors with similar perspective and body postures. This reveals the model's insufficient cross-modality generalization capability for fine-grained appearance features (e.g., clothing), particularly given that texture information loss in thermal-infrared images may lead to misidentification of posture changes as different identities. Similarly, the 2nd-row vessel retrieval results in Figure 6 show that RIRE correctly associates side-view thermal-infrared images unmatched by baselines, but exhibits false detection of similar images at Rank-5. This might stem from random erasing data augmentation obscuring the mast in visible images, while thermal images fail to display distinct thermal signatures of masts due to material conductivity differences, resulting in information loss across discriminative key regions. These cases indicate that although RIRE alleviates large perspective variations through global-local feature decoupling, modality-specific interference under extreme conditions (e.g., thermal radiation noise, occlusion of critical components) remains a core challenge limiting high-precision retrieval.

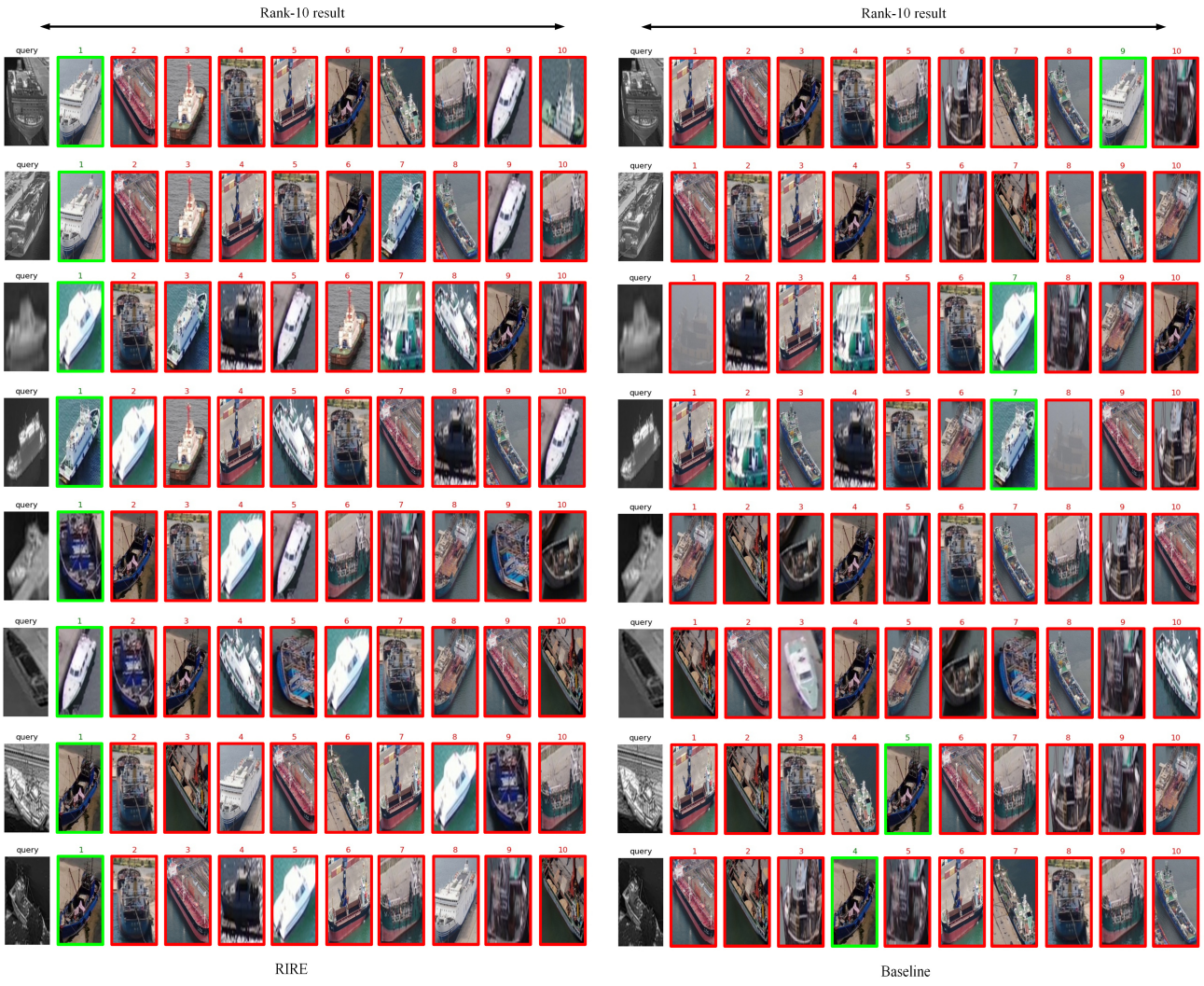

Figure 6. The Rank-10 retrieval results obtained by the proposed RIRE and baseline on the AVC-ReID dataset. The red box represents a search error, and the green box represents a correct search.

5.4.2. Feature Distribution

To further demonstrate the effectiveness of RIRE, we visualized the intra-class and inter-class distances of features extracted by our RIRE on the BA-VIP dataset. We used t-SNE [64] and UMAP [65] algorithms to map high-dimensional features into 2D and 3D spaces, and visualized them using the Matplotlib library, which the version is 3.3.4. The feature distributions mapped on the BA-VIP dataset shown in Figure 7 clearly illustrate that the features of visible light images and infrared images are essentially segregated in the feature space of the original feature distribution map due to significant modality differences among various modal image data. When the DSGFA is included, the feature space distance between features of multi-modality datasets belonging to the same individual decreases. This effect becomes more pronounced when the GRM operates on the network. Finally, when the two modules are both used in the framework, the features of multiple images of different modalities belonging to the same individual have converged in the feature space, further proving the positive role played by the DSGFA and GRM modules in alleviating observation perspective differences and modality differences.

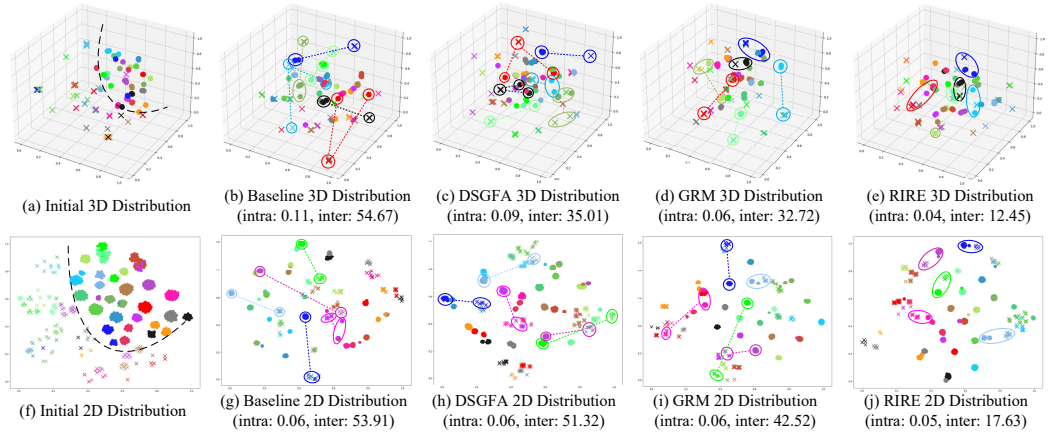

Figure 7. (**a**–**j**) illustrate the feature distribution of cross-modality images on the BA-VIP dataset. We select 20 IDs from the test set, of which 10 visible images and 10 infrared images are selected for each ID to calculate the feature distance. The "dots" and "crosses" represent visible light images and infrared images, respectively. The same color indicates features from the same ID, and the distance between the same-colored "dot" and "cross" represents the intra-class distance, while the distance between different-colored "dots" and "crosses" represents the inter-class distance. The quantitative indicators of intra-class variance and inter-class distance are shown in parentheses.

5.5. Experiments on Public Datasets

While classical VI-ReID datasets such as SYSU-MM01 and RegDB inherently encapsulate perspective-related challenges, their controlled acquisition environments fail to fully replicate the large-scale perspective disparities (particularly elevation angle variations) observed in UAV-captured data. However, given our method's theoretical capability to address severe perspective discrepancies, we hypothesize its potential adaptability across these canonical benchmarks. To empirically validate this hypothesis, we conduct supplementary cross-dataset evaluations on SYSU-MM01 and RegDB, both widely recognized as standard testbeds in VI-ReID research. SYSU-MM01 is a larger VI-ReID dataset that includes more identities and more complex scene changes, while the RegDB dataset is relatively small with fewer identities.

The experimental results are summarized in Table 6, where the accuracy metrics of comparative methods are cited from [23]. The red bold in the table represents the optimal result, and the blue represents the suboptimal result. On both SYSU-MM01 and RegDB datasets, RIRE demonstrates remarkable superiority in VI-ReID, comprehensively outperforming state-of-the-art methods. Despite the inherent limitations of classical benchmarks in fully simulating extreme perspective disparities (e.g., drastic elevation angle variations) typical of UAV-captured scenarios due to constrained acquisition conditions, RIRE exhibits exceptional generalization capability even within these controlled environments.

Specifically, under the all-search mode of SYSU-MM01, RIRE achieves 80.9% R-1 and 79.1% mAP, while in indoor-search scenarios, its performance further improves to 86.2% R-1 and 88.9% mAP, underscoring its robustness in complex environments. For cross-modality tasks on RegDB, RIRE attains leading R-1 scores of 92.6% (VIS-to-IR) and 89.8% (IR-to-VIS), yet falls slightly behind DEEN [47] in mAP. This discrepancy may stem from RIRE's design focus: its strong robustness and feature generalization capability in large-scale, complex scenarios enhance R-1 accuracy, while its limited sensitivity to fine-grained discrepancies in smaller-scale datasets like RegDB results in marginally lower mAP. This trade-off highlights the need for future refinements in cross-modality ranking consistency without compromising perspectives invariance.

Table 6. VI-ReID accuracy between RIRE and the other state-of-the-art methods on SYSU-MM01 and RegDB.

Methods	SYSU-MM01								RegDB							
	All Search				Indoor Search				VIS to IR				IR to VIS			
	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP	R-1	R-10	R-20	mAP
BDTR [49]	17.0	55.4	72.0	19.7	-	-	-	-	33.6	58.6	67.4	32.8	32.9	58.5	68.4	32.0
D^2RL [35]	28.9	70.6	82.4	29.2	-	-	-	-	43.4	66.1	76.3	44.1	-	-	-	-
Hi-CMD [43]	34.9	77.6	-	35.9	-	-	-	-	70.9	86.4	-	66.0	-	-	-	-
JSIA-ReID [59]	38.1	80.7	89.9	36.9	43.8	86.2	94.2	52.9	48.1	-	-	48.9	48.5	-	-	49.3
AlignGAN [34]	42.4	85.0	93.7	40.7	45.9	87.6	94.4	54.3	57.9	-	-	53.6	56.3	-	-	53.4
X-Modality [39]	49.9	89.8	96.0	50.7	-	-	-	-	62.2	83.1	91.7	60.2	-	-	-	-
DDAG [42]	54.8	90.4	95.8	53.0	61.0	94.1	98.4	68.0	69.3	86.2	91.5	63.5	68.1	85.2	90.3	61.8
CM-NAS [62]	60.8	92.1	96.8	58.9	68.0	94.8	97.9	52.4	82.8	95.1	97.7	79.3	81.7	94.1	96.9	77.6
MCLNet [36]	65.4	93.3	97.1	62.0	72.6	97.0	99.2	76.6	80.3	92.7	96.0	73.1	75.9	90.9	94.6	69.5
SMCL [37]	67.4	92.9	96.8	61.8	68.8	96.6	98.8	75.6	83.9	-	-	79.8	83.1	-	-	78.6
CAJ [38]	69.9	95.7	98.5	66.9	76.3	97.9	99.5	80.4	85.0	95.5	97.5	79.1	84.8	95.3	97.5	77.8
MPANet [45]	70.6	96.2	98.8	68.2	76.7	98.2	99.6	81.0	82.8	-	-	80.7	83.7	-	-	80.9
MMN [40]	70.6	96.2	99.0	66.9	76.2	97.2	99.3	79.6	91.6	97.7	98.9	84.1	87.5	96.0	98.1	80.5
DEEN [47]	74.7	97.6	99.2	71.8	80.3	99.0	99.8	83.3	91.1	97.8	98.9	85.1	89.5	96.8	98.4	83.4
PartMix [41]	77.8	-	-	74.6	81.5	-	-	84.4	84.9	-	-	82.5	85.7	-	-	82.3
RIRE (ours)	80.9	98.1	99.8	79.1	86.2	99.1	100.0	88.9	92.6	98.5	99.2	84.3	89.8	97.7	98.9	83.2

6. Discussion

6.1. Analysis of the Effectiveness of Robust Identity Representation Extraction Framework

The proposed RIRE framework demonstrates significant advancements in addressing the challenges of VTI-ReID for aerial images with large elevation angle variations. By decoupling global and local features through global representation decomposition and local aggregation, RIRE effectively aligns cross-modality features across diverse perspectives, achieving state-of-the-art performance on both the BA-VIP (person) and AVC-ReID (vessel) datasets. Specifically, the integration of the DSGFA and GRM modules enhances the adaptability of identity-related features to perspective differences, as evidenced by improved Rank-1 accuracy and mAP metrics compared to existing methods. Visualization of feature distributions further validates that RIRE successfully reduces intra-class distances between cross-modality samples, confirming its robustness in handling perspective variations.

6.2. Limitations Analysis of Robust Identity Representation Extraction Framework

However, the framework exhibits limitations in fine-grained feature generalization under extreme conditions. For instance, in the BA-VIP dataset's retrieval results (Figure 5), while RIRE matches targets with perspective variations, the top-ranked results still include distractors with similar postures or clothing, indicating insufficient sensitivity to texture details lost in thermal-infrared modalities. Similarly, for vessels (Figure 6), false detections at Rank-5 arise from thermal signature ambiguities (e.g., mast occlusion in visible images due to random erasing). These cases highlight that modality-specific interference—such as thermal radiation noise or critical component occlusion—remains a core challenge. Additionally, RIRE's computational efficiency poses a practical limitation: the framework's complexity (76.88M parameters and high FLOPs) may hinder real-time deployment on resource-constrained UAV platforms, where lightweight architectures are often prioritized to balance accuracy and operational costs. The dataset's inherent limitations, including restricted scene diversity (e.g., beach/sidewalk-only backgrounds in BA-VIP) and sample imbalance (uneven modality distributions), may further constrain the model's ability to generalize to complex real-world scenarios.

6.3. Future Research Trends

Future work should prioritize enhancing fine-grained feature discrimination, particularly for texture and structural attributes susceptible to modality-specific information loss. Developing robust data augmentation strategies to mitigate occlusion and thermal

noise, while preserving discriminative regions (e.g., masts or clothing patterns), could further improve retrieval precision. Secondly, expanding dataset diversity to include urban environments, nighttime conditions, and balanced identity distributions would better reflect real-world UAV surveillance challenges. Additionally, lightweight model designs could address the computational overhead of RIRE, facilitating deployment on resource-constrained aerial platforms. By addressing these limitations, future frameworks could achieve higher precision in cross-modality retrieval while maintaining robustness against extreme perspective and modality variations.

7. Conclusions

In this paper, a robust identity representation extraction framework for VTI-ReID of aerial images, named RIRE, is proposed. The framework adopts a mapping method based on global representation decomposition and local representation aggregation. Identity-related representations in aerial images are captured, and the global representations of targets from different perspectives are mapped to the same identity space, thereby enhancing adaptability to target observation perspective differences in aerial images, particularly elevation angle variations.

To validate the proposed framework, we construct a group of cross-modality aerial image Re-ID datasets, including two typical elongated categories: humans and vessels. Extensive testing data demonstrate that the framework achieves the best accuracy for different types of elongated target VTI-ReID under aerial perspectives. Visual testing results and feature analysis further confirm the superior adaptability of the framework to observation perspective differences.

Although superior adaptability to observation perspective differences of humans and vessels in UAV images for VTI-ReID tasks is demonstrated by our method, the complexity of the model makes it more sensitive to intra-class distances caused by perspective differences, thereby severely impeding the further improvement of mAP and mINP. This will be a key focus of future research.

Author Contributions: Conceptualization and funding acquisition, C.Z., Y.Y. and B.G.; methodology and writing the original draft W.W. and Y.Y.; the processing and analysis of cross-dimensional data, visualization, and editing of the manuscript, W.H. and F.G.; ablation experiment and polishing and modification of manuscripts, B.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 62371153, No. 62071136, No. 62271159), Heilongjiang Provincial Natural Science Foundation of China Grant (JJ2024LH2397), Fundamental Research Funds for the Central Universities Grant (3072024XX0801 and 3072024XX0805), and the Key Laboratory of Target Cognition and Application Technology (2023-CXPT-LC-005).

Data Availability Statement: The proposed dataset are available on request from the corresponding author due to the privacy.

Conflicts of Interest: Authors Yiming Yan, Baoyu Ge and Wei Hou were employed by the company Harbin Space Star Data System Science and Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Liu, Y.; Zhou, W.; Liu, J.; Qi, G.-J.; Tian, Q.; Li, H. An End-to-End Foreground-Aware Network for Person Re-Identification. *IEEE Trans. Image Process.* **2021**, *30*, 2060–2071. [CrossRef] [PubMed]

2. Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; Yang, Y. Learning to Adapt Invariance in Memory for Person Re-Identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2723–2738. [CrossRef] [PubMed]
3. Wang, K.; Wang, P.; Ding, C.; Tao, D. Batch Coherence-Driven Network for Part-Aware Person Re-Identification. *IEEE Trans. Image Process.* **2021**, *30*, 3405–3418. [CrossRef] [PubMed]
4. Wang, X.; Liu, M.; Wang, F.; Dai, J.; Liu, A.-A.; Wang, Y. Relation-Preserving Feature Embedding for Unsupervised Person Re-Identification. *IEEE Trans. Multimed.* **2024**, *26*, 714–723. [CrossRef]
5. Chen, G.; Lu, J.; Yang, M.; Zhou, J. Spatial-Temporal Attention-Aware Learning for Video-Based Person Re-Identification. *IEEE Trans. Image Process.* **2019**, *28*, 4192–4205. [CrossRef]
6. Li, J.; Zhang, S.; Tian, Q.; Wang, M.; Gao, W. Pose-Guided Representation Learning for Person Re-Identification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 622–635. [CrossRef]
7. Ferdous, S.N.; Li, X.; Lyu, S. Uncertainty Aware Multitask Pyramid Vision Transformer for UAV-Based Object Re-Identification. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 2381–2385. [CrossRef]
8. Yao, A.; Qi, J.; Zhong, P. Self-Aligned Spatial Feature Extraction Network for UAV Vehicle Reidentification. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 6002305. [CrossRef]
9. Xiong, M.; Yang, X.; Chen, H.; Aly, W.H.; AlTameem, A.; Saudagar, A.K.; Mumtaz, S.; Muhammad, K. Cloth-Changing Person Re-Identification with Invariant Feature Parsing for UAVs Applications. *IEEE Trans. Veh. Technol.* **2024**, *73*, 12448–12457. [CrossRef]
10. He, B.; Wang, F.; Wang, X.; Li, H.; Sun, F.; Zhou, H. Temporal Context and Environment-Aware Correlation Filter for UAV Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5630915. [CrossRef]
11. Fang, H.; Wu, C.; Wang, X.; Zhou, F.; Chang, Y.; Yan, L. Online Infrared UAV Target Tracking with Enhanced Context-Awareness and Pixel-Wise Attention Modulation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5005417. [CrossRef]
12. Zheng, L.; Huang, Y.; Lu, H.; Yang, Y. Pose-Invariant Embedding for Deep Person Re-Identification. *IEEE Trans. Image Process.* **2019**, *28*, 4500–4509. [CrossRef] [PubMed]
13. Yang, F.; Li, W.; Liang, B.; Zhang, J. Spatiotemporal Interaction Transformer Network for Video-Based Person Reidentification in Internet of Things. *IEEE Internet Things J.* **2023**, *10*, 12537–12547. [CrossRef]
14. Wei, J.; Pan, C.; He, S.; Wang, G.; Yang, Y.; Shen, H.T. Towards Robust Person Re-Identification by Adversarial Training with Dynamic Attack Strategy. *IEEE Trans. Multimed.* **2024**, *26*, 10367–10380. [CrossRef]
15. Fu, H.; Cui, K.; Wang, C.; Qi, M.; Ma, H. Mutual Distillation Learning for Person Re-Identification. *IEEE Trans. Multimed.* **2024**, *26*, 8981–8995. [CrossRef]
16. Cheng, Y.; Liu, Y. Person Reidentification Based on Automotive Radar Point Clouds. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5101913. [CrossRef]
17. Zhang, M.; Zhu, M.; Wei, X.; Wang, X.; Zhu, J.; Cheng, J.; Yang, Y. Deep learning-based person re-identification methods: A survey and outlook of recent works. *Image Vis. Comput.* **2022**, *119*, 104394. [CrossRef]
18. Duan, R.; Chen, L.; Li, Z.; Chen, Z.; Wu, B. A Scene Graph Encoding and Matching Network for UAV Visual Localization. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 9890–9902. [CrossRef]
19. Maboudi, M.; Homaei, M.; Song, S.; Malihi, S.; Saadatseresht, M.; Gerke, M. A Review on Viewpoints and Path Planning for UAV-Based 3-D Reconstruction. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 5026–5048. [CrossRef]
20. Liu, W.; Quijano, K.; Crawford, M.M. YOLOv5-Tassel: Detecting Tassels in RGB UAV Imagery with Improved YOLOv5 Based on Transfer Learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 8085–8094. [CrossRef]
21. Liu, X.; Qi, J.; Chen, C.; Bin, K.; Zhong, P. Relation-Aware Weight Sharing in Decoupling Feature Learning Network for UAV RGB-Infrared Vehicle Re-Identification. *IEEE Trans. Multimed.* **2024**, *26*, 9839–9853. [CrossRef]
22. Zhang, H.; Cao, H.; Yang, X.; Deng, C.; Tao, D. Self-Training With Progressive Representation Enhancement for Unsupervised Cross-Domain Person Re-Identification. *IEEE Trans. Image Process.* **2021**, *30*, 5287–5298. [CrossRef] [PubMed]
23. Yang, X.; Tian, M.; Li, M.; Wei, Z.; Yuan, L.; Wang, N.; Gao, X. SSRR: Structural Semantic Representation Reconstruction for Visible-Infrared Person Re-Identification. *IEEE Trans. Multimed.* **2024**, *26*, 6273–6284. [CrossRef]
24. Kansal, K.; Subramanyam, A.V.; Wang, Z.; Satoh, S. SDL: Spectrum-Disentangled Representation Learning for Visible-Infrared Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 3422–3432. [CrossRef]
25. Ye, H.; Liu, H.; Meng, F.; Li, X. Bi-Directional Exponential Angular Triplet Loss for RGB-Infrared Person Re-Identification. *IEEE Trans. Image Process.* **2021**, *30*, 1583–1595. [CrossRef]
26. Huang, Y.; Wu, Q.; Xu, J.; Zhong, Y.; Zhang, P.; Zhang, Z. Alleviating Modality Bias Training for Infrared-Visible Person Re-Identification. *IEEE Trans. Multimed.* **2022**, *24*, 1570–1582. [CrossRef]
27. Ye, M.; Lan, X.; Leng, Q.; Shen, J. Cross-Modality Person Re-Identification via Modality-Aware Collaborative Ensemble Learning. *IEEE Trans. Image Process.* **2020**, *29*, 9387–9399. [CrossRef]

28. Tan, X.; Chai, Y.; Chen, F.; Liu, H. A Fourier-Based Semantic Augmentation for Visible-Thermal Person Re-Identification. *IEEE Signal Process. Lett.* **2022**, *29*, 1684–1688. [CrossRef]
29. Wei, Z.; Yang, X.; Wang, N.; Gao, X. Flexible Body Partition-Based Adversarial Learning for Visible Infrared Person Re-Identification. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 4676–4687. [CrossRef]
30. Yang, X.; Dong, W.; Li, M.; Wei, Z.; Wang, N.; Gao, X. Cooperative Separation of Modality Shared-Specific Features for Visible-Infrared Person Re-Identification. *IEEE Trans. Multimed.* **2024**, *26*, 8172–8183. [CrossRef]
31. Feng, Y.; Chen, F.; Yu, J.; Ji, Y.; Wu, F.; Liu, T.; Liu, S.; Jing, X.Y.; Luo, J. Cross-Modality Spatial-Temporal Transformer for Video-Based Visible-Infrared Person Re-Identification. *IEEE Trans. Multimed.* **2024**, *26*, 6582–6594. [CrossRef]
32. Zhang, Q.; Yan, Y.; Gao, L.; Xu, C.; Su, N.; Feng, S. A Third-Modality Collaborative Learning Approach for Visible-Infrared Vessel Reidentification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 19035–19047. [CrossRef]
33. Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; Lai, J. RGB-Infrared Cross-Modality Person Re-identification. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 5390–5399. [CrossRef]
34. Wang, G.; Zhang, T.; Cheng, J.; Liu, S.; Yang, Y.; Hou, Z. RGB-Infrared Cross-Modality Person Re-Identification via Joint Pixel and Feature Alignment. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3622–3631. [CrossRef]
35. Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.-Y.; Satoh, S. Learning to Reduce Dual-Level Discrepancy for Infrared-Visible Person Re-Identification. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 618–626. [CrossRef]
36. Hao, X.; Zhao, S.; Ye, M.; Shen, J. Cross-Modality Person Re-Identification via Modality Confusion and Center Aggregation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 16383–16392. [CrossRef]
37. Wei, Z.; Yang, X.; Wang, N.; Gao, X. Syncretic Modality Collaborative Learning for Visible Infrared Person Re-Identification. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 225–234. [CrossRef]
38. Ye, M.; Ruan, W.; Du, B.; Shou, M.Z. Channel Augmented Joint Learning for Visible-Infrared Recognition. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 13547–13556. [CrossRef]
39. Li, D.; Wei, X.; Hong, X.; Gong, Y. Infrared-visible cross-modal person re-identification with an x modality. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 4610–4617. [CrossRef]
40. Zhang, Y.; Yan, Y.; Lu, Y.; Wang, H. Towards a unified middle modality learning for visible-infrared person re-identification. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 788–796. [CrossRef]
41. Kim, M.; Kim, S.; Park, J.; Park, S.; Sohn, K. PartMix: Regularization Strategy to Learn Part Discovery for Visible-Infrared Person Re-Identification. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 18621–18632. [CrossRef]
42. Ye, M.; Shen, J.; Crandall, D.J.; Shao, L.; Luo, J. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XVII 16; Springer International Publishing: Cham, Switzerland, 2020; pp. 229–247. [CrossRef]
43. Choi, S.; Lee, S.; Kim, Y.; Kim, T.; Kim, C. Hi-CMD: Hierarchical Cross-Modality Disentanglement for Visible-Infrared Person Re-Identification. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 10254–10263. [CrossRef]
44. Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; Yu, N. Cross-Modality Person Re-Identification with Shared-Specific Feature Transfer. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 13376–13386. [CrossRef]
45. Wu, Q.; Dai, P.; Chen, J.; Lin, C.W.; Wu, Y.; Huang, F.; Zhong, B.; Ji, R. Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 4328–4337. [CrossRef]
46. Liu, J.; Sun, Y.; Zhu, F.; Pei, H.; Yang, Y.; Li, W. Learning Memory-Augmented Unidirectional Metrics for Cross-modality Person Re-identification. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 19344–19353. [CrossRef]
47. Zhang, Y.; Wang, H. Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-identification. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 2153–2162. [CrossRef]

48. Feng, J.; Wu, A.; Zheng, W.-S. Shape-Erased Feature Learning for Visible-Infrared Person Re-Identification. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 22752–22761. [CrossRef]
49. Ye, M.; Wang, Z.; Lan, X.; Yuen, P. Visible thermal person re-identification via dual-constrained top-ranking. *IJCAI* **2018**, *1*, 2. [CrossRef]
50. Hao, Y.; Wang, N.; Li, J.; Gao, X. HSME: Hypersphere manifold embedding for visible thermal person re-identification. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 8385–8392. [CrossRef]
51. Feng, Z.; Lai, J.; Xie, X. Learning Modality-Specific Representations for Visible-Infrared Person Re-Identification. *IEEE Trans. Image Process.* **2020**, *29*, 579–590. [CrossRef]
52. Ye, M.; Lan, X.; Wang, Z.; Yuen, P.C. Bi-Directional Center-Constrained Top-Ranking for Visible Thermal Person Re-Identification. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 407–419. [CrossRef]
53. Tan, L.; Zhang, Y.; Shen, S.; Wang, Y.; Dai, P.; Lin, X.; Wu, Y.; Ji, R. Exploring invariant representation for visible-infrared person re-identification. *arXiv* **2023**, arXiv:2302.00884. [CrossRef]
54. Zhou, X.; Zhong, Y.; Cheng, Z.; Liang, F.; Ma, L. Adaptive Sparse Pairwise Loss for Object Re-Identification. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 19691–19701. [CrossRef]
55. Nguyen, D.T.; Hong, H.G.; Kim, K.W.; Park, K.R. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* **2017**, *17*, 605. [CrossRef]
56. Available online: https://enterprise.dji.com/cn/matrice-300 (accessed on 1 November 2023).
57. Available online: https://enterprise.dji.com/cn/zenmuse-h20-series (accessed on 1 November 2023).
58. Luo, H.; Gu, Y.; Liao, X.; Lai, S.; Jiang, W. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–20 June 2019; pp. 1487–1495. [CrossRef]
59. Wang, G.; Zhang, T.; Yang, Y.; Cheng, J.; Chang, J.; Liang, X.; Hou, Z. Cross-modality paired-images generation for RGB-infrared person re-identification. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12144–12151. [CrossRef]
60. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 13001–13008. [CrossRef]
61. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C.H. Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2872–2893. [CrossRef] [PubMed]
62. Fu, C.; Hu, Y.; Wu, X.; Shi, H.; Mei, T.; He, R. CM-NAS: Cross-Modality Neural Architecture Search for Visible-Infrared Person Re-Identification. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 11803–11812. [CrossRef]
63. Yang, M.; Huang, Z.; Hu, P.; Li, T.; Lv, J.; Peng, X. Learning with Twin Noisy Labels for Visible-Infrared Person Re-Identification. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 14288–14297. [CrossRef]
64. Van der Maaten, L.; Hinton, G.E. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
65. Xuan, H.; Stylianou, A.; Liu, X.; Pless, R. Hard negative examples are hard, but useful. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XIV 16; Springer International Publishing: Cham, Switzerland, 2020; pp. 126–142. [CrossRef]