electronics

MDPI

Review

Visible-Infrared Person Re-Identification: A Comprehensive Survey and a New Setting

Huantao Zheng [1], Xian Zhong [1,2,*], Wenxin Huang [3,*], Kui Jiang [4], Wenxuan Liu [1] and Zheng Wang [4]

[1] School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China; zhenghuantao@whut.edu.cn (H.Z.); lwxfight@whut.edu.cn (W.L.)
[2] School of Electronics Engineering and Computer Science, Peking University, Beijing 100091, China
[3] School of Computer Science and Information Engineering, Hubei University, Wuhan 430062, China
[4] School of Computer Science, Wuhan University, Wuhan 430072, China; kuijiang@whu.edu.cn (K.J.); wangzwhu@whu.edu.cn (Z.W.)
* Correspondence: zhongx@whut.edu.cn (X.Z.); wenxinhuang_wh@163.com (W.H.)

Abstract: Person re-identification (ReID) plays a crucial role in video surveillance with the aim to search a specific person across disjoint cameras, and it has progressed notably in recent years. However, visible cameras may not be able to record enough information about the pedestrian's appearance under the condition of low illumination. On the contrary, thermal infrared images can significantly mitigate this issue. To this end, combining visible images with infrared images is a natural trend, and are considerably heterogeneous modalities. Some attempts have recently been contributed to visible-infrared person re-identification (VI-ReID). This paper provides a complete overview of current VI-ReID approaches that employ deep learning algorithms. To align with the practical application scenarios, we first propose a new testing setting and systematically evaluate state-of-the-art methods based on our new setting. Then, we compare ReID with VI-ReID in three aspects, including data composition, challenges, and performance. According to the summary of previous work, we classify the existing methods into two categories. Additionally, we elaborate on frequently used datasets and metrics for performance evaluation. We give insights on the historical development and conclude the limitations of off-the-shelf methods. We finally discuss the future directions of VI-ReID that the community should further address.

Keywords: visible-infrared person re-identification; non-generative-based model; generative-based model; literature survey

Citation: Zheng, H.; Zhong, X.; Huang, W.; Jiang, K.; Liu, W.; Wang, Z. Visible-Infrared Person Re-Identification: A Comprehensive Survey and a New Setting. *Electronics* **2022**, *11*, 454. https://doi.org/10.3390/electronics11030454

Academic Editor: Stefanos Kollias

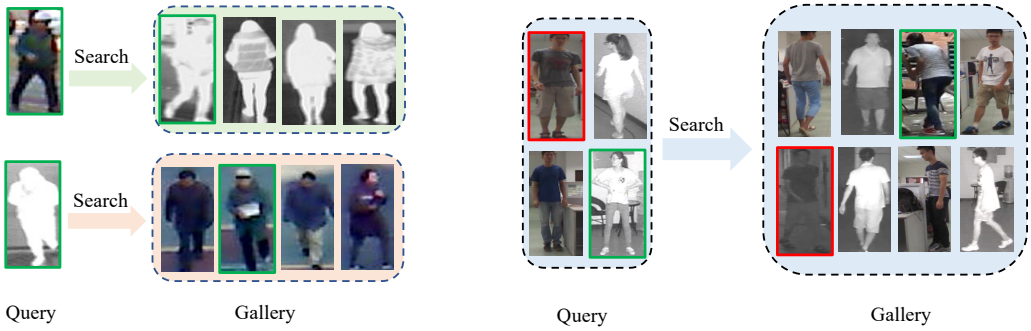
Received: 29 December 2021
Accepted: 29 January 2022
Published: 3 February 2022

1. Introduction

Person re-identification (ReID) is a fundamental building block in various tasks of computer vision, such as intelligent surveillance, video analysis [1], and criminal investigation [2]. With the advancement of intelligent monitoring and the enormous expansion of video data in recent years, conventional human power has been challenging and insufficient to deal with intricate surveillance scenarios. ReID aims at searching for a given individual across disjoint cameras. Numerous algorithms designed for ReID have been proposed with impressive results on some publicly available datasets, e.g., 98.1% and 94.5% Rank-1 accuracy on Market-1501 [3] and DukeMTMC-ReID [4] datasets, respectively [5]. However, the images captured by visible cameras may be unavailable in a dark environment. In such a case, infrared imaging equipment, which does not rely on visible light, should be applied. In 2017, Wu et al. [6] first introduced visible-infrared person re-identification (VI-ReID) and proposed a dataset named SYSU-MM01.

As shown in Figure 1a, for a certain pedestrian, the images of the corresponding identity (ID) should be matched from the other modality set. In addition to the common challenges, e.g., low-resolution, viewpoint change, pose variation, and occlusion, VI-ReID is an effortful problem that encounters additional modality discrepancy due to the significant

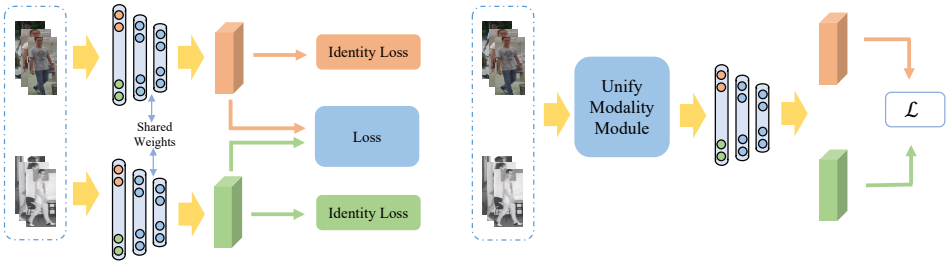
differences between the two modalities. The two modalities can be considered heterogeneous data, as visible images contain three rich color information channels. In contrast, infrared images only include one channel with near-infrared light intensity information. Additionally, from the aspect of the imaging principle, the two modalities have differences in terms of the wavelength range. Moreover, the datasets are relatively single and small in scale. Some works expanded the VI-ReID datasets, but these datasets cannot be disclosed because of privacy issues.



(a) Illustration of existing setting (b) Illustration of new setting

Figure 1. Comparison of two testing settings. Images with the same color of bounding boxes denote the same person identity. (**a**) The query and gallery only contain images from single modality. (**b**) Both the query and gallery contain images from two modalities.

To improve the practical application ability of VI-ReID, researchers previously achieved remarkable progress on VI-ReID. We divide existing methods into two categories—non-generative-based and generative-based—which were proposed in [7]. As shown in Figure 2a, the non-generative-based model mainly utilizes conventional methods, including feature representation learning and distance metric learning, to maximize the similarity between two images with the same ID and minimize the similarity between two images with different IDs [8–10]. In contrast, Figure 2b shows a generative-based model that unifies the modality on the data level, bridging the gap between two heterogeneous modalities [11,12].



(a) Non-generative-based Model (b) Generative-based Model

Figure 2. Illustration of two different pipelines of VI-ReID.

To the best of our knowledge, almost all VI-ReID systems evaluated their performance based on the setting as shown in Figure 1a. However, this may not be in line with the actual scene. Taking the visible image V as an example, it may be more similar with some negative visible samples than positive infrared samples. The existing testing setting removes all visible images in the gallery to avoid this challenge. In this paper, we propose a novel testing setting that is closer to the practical scene. As shown in Figure 1b, instead of containing images from only one modality in query and gallery, we simultaneously put visible and infrared images into the query and gallery. This setting makes VI-ReID more challenging. Existing works created various two-stream architectures to learn modality-specific information in order to alleviate the cross-modality discrepancy. However, this

kind of two-stream network may not extract effective features of visible and infrared images simultaneously in our new setting. Considering the realistic value of this setting, we believe that researchers should pay more attention to it.

In recent years, many excellent review papers have appeared in ReID. For example, Wang et al. [13] considered four different cross-modality application scenarios: low-resolution, infrared, sketch, text and then analyzed typical approaches. Ye et al. [8] categorized related works into closed-world ReID and open-world ReID, and proposed a strong baseline named AGW. Leng et al. [14] sorted out the papers in open-world ReID based on specific application scenarios. Inspired by them, we conduct a thorough overview for VI-ReID.

Our contributions are threefold:

- We propose a new testing setting which is closer to practical application scenarios and conduct preliminary experiments to verify the significant challenges of the new setting.
- We compare VI-ReID with ReID in detail and provide a thorough review of VI-ReID techniques, including datasets and performance metrics.
- We conclude the necessary components of networks and discuss possible future directions of VI-ReID.

2. Visible-Infrared Person Re-Identification

2.1. ReID vs. VI-ReID

Generally, there is just visible modality in ReID, while VI-ReID contains two modalities: visible and infrared. As all know, visible images have three channels containing rich color information, while infrared images contain intensity information with the red channel only. As shown in Figure 3, there is a noticeable modality gap between visible and infrared images [11]. As the inter-modality discrepancy is substantially greater than the intra-modality discrepancy, bridging the modality gap between the two heterogeneous modalities is a major aspect of VI-ReID research.

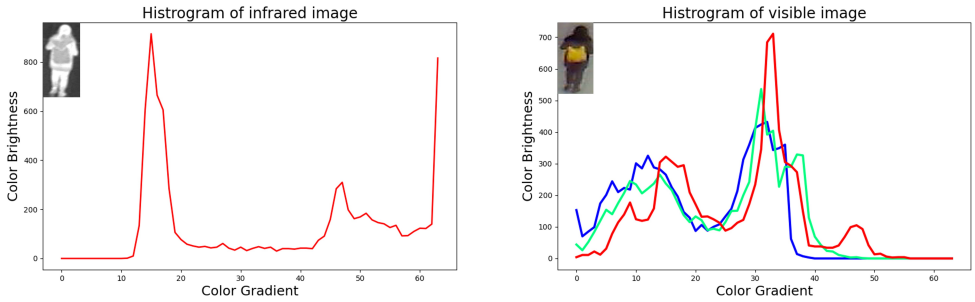


Figure 3. Comparison of visible three-channel brightness-gradient histograms of visible-infrared image pairs with the same ID on RegDB dataset.

For ReID, it only faces challenges of intra-modality, e.g., people's appearance change, viewpoint change, and occlusion. In contrast, VI-ReID confronts not only the difficulties that appear in ReID, but also the cross-modality discrepancy. The networks designed for ReID are not suitable for VI-ReID since the solutions to intra- and inter-modality discrepancies are completely different.

To our knowledge, the performance gap between VI-ReID and ReID is also large. Ref. [15], for example, achieved 95.7% rank-1 on Market-1501, while the rank-1 of [10] just reached 70.58% on the SYSU-MM01 dataset. The performance of VI-ReID is far lower than that of ReID. However, the VI-ReID is more valuable in practical application scenarios, and we should pay more attention to it.

2.2. A New Testing Setting

In a VI-ReID dataset, $V = \{X_v^i\}_{i=1}^{N_v}$ and $T = \{X_t^i\}_{i=1}^{N_t}$ represent the visible and infrared images, respectively, where N_v and N_t denote the number of samples in a single modality, respectively. Every image has a corresponding ID label $y \in \{Y_i\}_{i=1}^{N_p}$, where N_p denotes the number of IDs. Given a certain image as the query, the purpose of VI-ReID is to match images with the same label from the other modality according to the similarity.

However, this setting is not in line with practical scenarios. Just imagine that, given an image of a criminal who has been escaped for several days, we have to search for him via cross visible and infrared cameras. The off-the-shelf methods may not be useful in such a case. Instead of containing images from only one modality, as shown in Figure 1b, we set the probe $P = V_p \cup T_p$ and gallery $G = V_g \cup T_g$ to simultaneously contain visible and infrared images, where $\cup$ denotes union. V_p, V_g, and T_p, T_g are the mutually exclusive subsets of V and T, respectively. When researchers evaluate methods with our new setting, the images with the same ID and modality as the probe cannot appear in the gallery to avoid the impact of ReID in the same modality. When training with this setting, the mainstream dual-branch network structure may not extract effective features because of the effect of mixed modalities. The P and G first generate features $F_p = \{F_i^p\}_{i=1}^{N}$ and $F_g = \{F_i^g\}_{i=1}^{N}$ through the feature extraction module, and then they are matched by the feature matching module.

3. VI-ReID Methods

According to our investigation, there are no other types of articles published on mainstream conferences or journals except those that are deep-learning-based. Hence, only deep-learning-based approaches are included in this review. For the non-generative-based model, we subdivide the model into feature learning, metric learning, and training strategy. For the generative-based model, we subdivide the model into modality translation and extra modality. Besides, we introduce some methods using other technologies. We also summarize some algorithms intended for general ReID or other domains that perform well in VI-ReID. Some methods may be appropriate for multiple categories; however, we will take them to the most suitable position.

3.1. Milestones of Existing VI-ReID Studies

VI-ReID has achieved significant progress in a variety of areas thanks to the unwavering efforts of artificial intelligence researchers. We introduce these crucial milestones for VI-ReID following a timeline and present them in Figure 4. Note that the main basis of a paper selected as a milestone is its citations. We select the paper with highest citations among all papers in a category after dividing the papers into different categories.

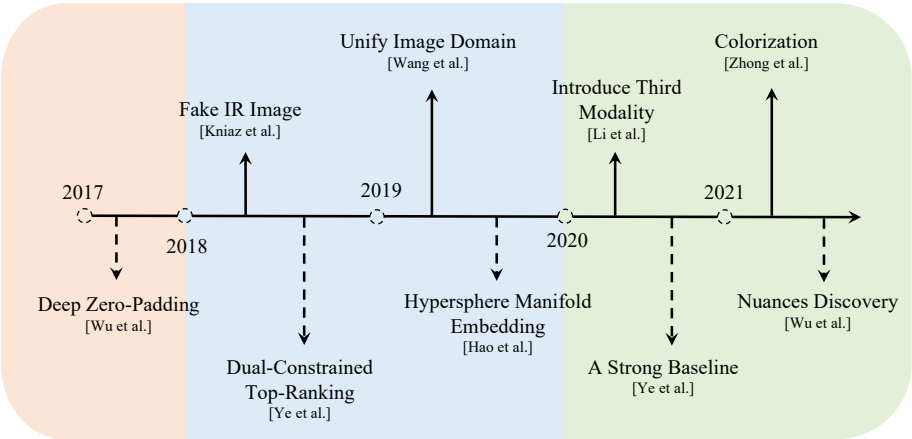


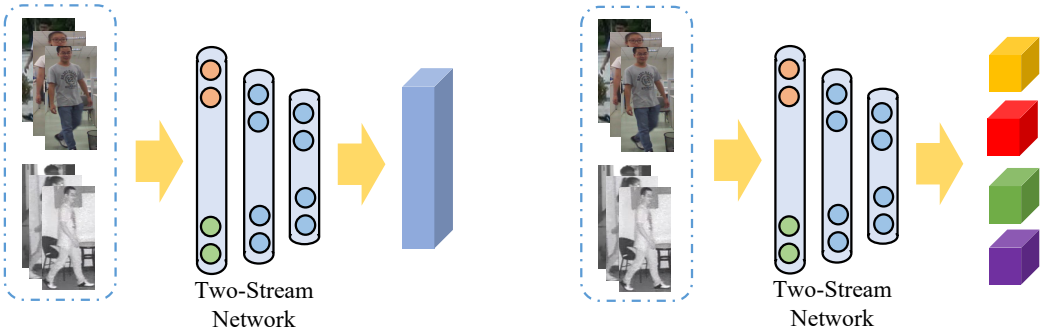
Figure 4. Milestones of existing methods of VI-ReID. Since the problem was proposed, researchers have proposed various methods to bridge the modality discrepancy. Note that the top row and bottom row denote a generative-based model and non-generative-based model, respectively.

3.2. Non-Generative-Based Model

3.2.1. Feature Representation Learning

It aims to extract robust and discriminative features to help the VI-ReID system correctly classify images into different fine-grained classes. We review three kinds of feature representation learning strategies.

Global Feature Representation Learning. As far as we know, most existing methods focus on extracting global features. An illustration is shown in Figure 5a. To obtain modality specific information, Feng et al. [16] established two individual branches for visible and infrared images, respectively. In [17], the authors thought only learning shared features means a massive loss of information, which reduces the difference of features. Therefore, they proposed a cross-modality shared-specific feature transfer algorithm. Ye et al. [18] pointed out that the consistency at the feature and classifier levels is essential when dealing with modality differences. To learn discriminative representations in each modality, Wei et al. [19] developed an attention-lifting mechanism. Wang et al. [20] excavated spatial and channel information of images to reduce the discrepancy between two heterogeneous modalities.



(a) Global Feature (b) Local Feature

Figure 5. Two alternative ways for learning feature representations. (**a**) Global feature, learning modality specific and -shared feature representations. (**b**) Local feature, learning part-aggregated local feature.

In addition, some works extract ID-invariant features by disentanglement to boost the performance. To achieve more robust retrieval for VI-ReID, Pu et al. [21] disentangled an ID-discriminable and an ID-ambiguous cross-modality feature subspace, respectively. In [22], the authors thought existing methods do not explicitly ignore spectrum information that is not related to VI-ReID. As a result, they disentangled the spectrum information in order to maximize invariant ID information while minimizing the influence of spectrum information. Zhao et al. [23] learned color-irrelevant features through color-irrelevant consistency learning and aligned the ID-level feature distributions by the ID-aware modality adaptation. Hao et al. [24] confused two modalities to learn modality irrelevant representation. In [10], the authors extracted modality irrelevant features by channel attention-guided instance normalization (IN).

Local Feature Representation Learning. As shown in Figure 5b, compared to the global feature, the local feature is more focused on the differences in details. Lin et al. proposed an attribute–person recognition network to make full use of the information contained in attributes [25]. Hao et al. [26] replaced global features with part-level features so that fine-grained camera-invariant information can be extracted. In [27], the authors proposed an adaptive body partition model for automatically detecting and distinguishing effective component representations. Liu et al. [28] presented a network that jointly learns global and local features to cope with viewpoint change and pose variation. Ye et al. [29] excavated contextual cues at the intra-modality components and cross-modality graph levels. Wang et al. [30] utilized global features and partial features to realize the

complement of global information and detailed information. To select useful features, Wei et al. [31] designed a flexible body partition module to distinguish part representations automatically. Zhang et al. concatenated the global feature and local feature to create a more powerful feature descriptor [32]. In [33], aiming to eliminate the interference of background information, the authors exploited the knowledge of human body parts to extract robust features. Wu et al. [10] utilized pattern alignment to discover nuances in different patterns. Zhang et al. [34] also made an attempt to discover semantic differences between contrastive features by cross correlation.

Auxiliary Feature Representation Learning. Ye et al. [35] exploited auxiliary information, including the distribution of cross-modality features and contextual information, to bridge the gap between heterogeneous modalities. In [36], the authors designed camera-based batch normalization (BN) to guarantee an invariant input distribution independent of all cameras.

3.2.2. Metric Learning

The purpose of metric learning is to guide feature representation learning. We will go through some prevalent loss functions and training strategies.

Loss Function Design. Generally, researchers design different loss functions to solve targeted problems based on the observed phenomenon. A large cross-modal discrepancy and intra-modal variations generated by varied camera angles, human postures, etc., impact the VI-ReID. As shown in Figure 6a, the function of identity loss is to classify a sample into a correct class in the training phase, which is widely used. For contrastive loss, as shown in Figure 6b, it mainly constrains the training of Siamese networks. For instance, Ye et al. [37] proposed a hierarchical cross-modal matching model, which jointly optimized the modality shared and -specific matrix, aiming at the problem of perspective changing when different cameras record a person. To minimize the difference between same modality and cross similarities, Wu et al. [38] guided the learning of cross-modality similarity by same-modality similarity.

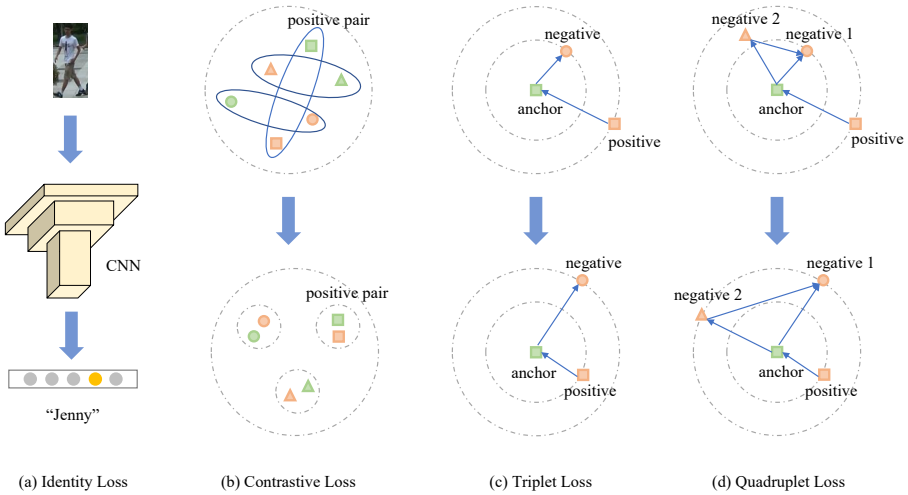


Figure 6. Four widely used loss functions. Different shapes denote different IDs, while the different color represents different modality. (**a**) Identity loss, (**b**) contrastive loss, (**c**) triplet loss, (**d**) quadruplet loss. Many works employ their combinations.

Figure 6c shows the triplet loss, which is contributed to pull the distance between positive sample pairs and push the distance between negative sample pairs. The samples with the same ID form clusters in feature space. The approach constrains the features by a set of triplets to obtain high performance [39]. Wang et al. [40] proposed an improved triplet loss to realize matching a video by an image. Ye et al. [9] proposed a bi-directional dual-constrained top-ranking loss to guide the feature learning objectives. Then, they improved

this work by replacing the similarity between two samples with similarity between sample and center [41]. To alleviate the strict constraint of classical triplet loss, Liu et al. [2] proposed an improved triplet loss with the mode of center to center instead of instance to instance. Zhang et al. mitigated the modality discrepancy by mapping the heterogeneous representations into a common space [42]. To learn an angularly separable common feature space, Ye et al. [1] constrained the angles between feature vectors. Cai et al. [43] proposed a dual-modality hard mining triplet-center loss (DTCL) which can reduce computational cost and mine hard triplet samples. In order to eliminate the effect of inconsistent feature distribution in different modalities, Zhang et al. [44] mapped the feature space to angular space and proposed several loss functions to conduct specific angular metric learning.

Figure 6d shows quadruplet loss, which is an improved version of triplet loss. It adds relative distance between the samples with different IDs. In [45], current approaches, according to the authors, primarily combine classification and metric learning to train models in order to generate discriminative and robust representations. However, these methods ignore the relationship between the classification and feature embedding subspaces. The authors presented a hyperspherical manifold-embedded network with classification and recognition constraints based on this information. Jia et al. [46] utilized the similarity transitivity to tackle the problem of mismatching hard positive samples.

Training Strategy. To incorporate different loss functions into an organic whole, researchers have proposed different training strategies. Dai et al. [47] proposed a generative adversarial training strategy to deal with the lack of discriminative information. Ye et al. [48] observed that existing VI-ReID learning strategies ignore the discriminative information of different modalities. Therefore, they presented a modality aware collaborative learning strategy to deal with the gap between two modalities in both the feature level and classifier level. Zhang et al. [49] proposed a mutual learning module that provides a bi-directional transfer between two modalities, aiming at excavating useful information from them. Ling et al. [50] thought most existing methods constrain the similarity of the instance or class level, which is inadequate to make full use of the hidden relationships in cross-modality data. Hence, they proposed multi-constraint similarity learning from instance to instance, instance to class center, and class center to class center. Gao et al. [51] proposed a learning strategy for joint optimization of a single modality and unified modality spaces.

3.3. Generative-Based Model

The generative-based model mainly utilizes generative adversarial network (GAN) or encoder–decoder module to realize the mutual translation between the two modalities. Then, the methods of ReID are used to constrain the appearance of discrepancy.

3.3.1. Modality Translation

In recent years, GAN-based modality translation has gradually become popular. As shown in Figure 7a, modality translation includes infrared to visible and visible to infrared. In contrast, some works disentangled ID-discriminative and ID-excluding factors, and then generated image pairs to extract highly discriminative features.

For infrared to visible, this kind of method can be regarded as image colorization that has been extensively used in various fields [52]. To our knowledge, there is little work in the literature using colorization. Zhong et al. [53] bridged the gap between the two modalities by fusing the features of original infrared images and generated fake visible images. After that, Zhong et al. [11] improved the performance by pixel-wise transformation, which can retain original structure information.

For visible to infrared, Kniaz et al. [54] matched the fake infrared images generated by GAN with the gallery images to mitigate the modality discrepancy. Wang et al. added a pixel alignment module based on feature alignment module [47] to further reduce the gap between the two modalities [12]. However, Liu et al. [55] thought that those methods employing GAN to generate fake images destroy the structure information of generated

images and introduce plenty of noise. Hence, they replaced fake images generated by GAN with grayscale images with three channels.

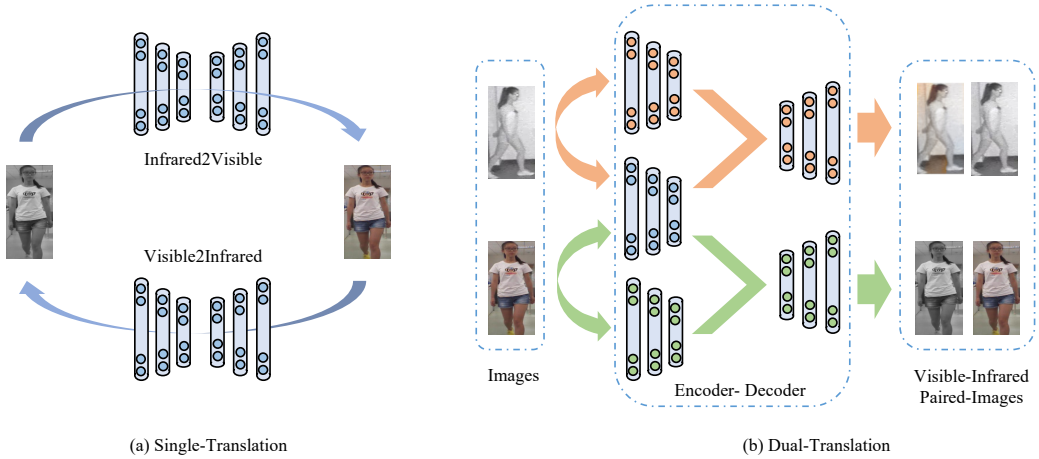


Figure 7. Three kinds of widely used methods in the literature. (**a**) Infrared to visible [11,53] and visible to infrared [12,54]; (**b**) more works generate visible-infrared image pairs, employing their combination [56–60].

For dual translation, as shown in Figure 7b, it encodes the visible and infrared modalities into a consistent space to eliminate the effect of modality style. It then generates fake cross-modality image pairs with the same ID. In 2019, Wang et al. [56] first generated visible-infrared image pairs by disentanglement and mapped them into a unified space. Analogous to [56], the idea of disentanglement is also indicated in [57–60]. Among them, Choi et al. [57] encoded the prototype and the attribute separately to generate fake images containing invariant features. Meanwhile, [58,59] acquired visible-infrared image pairs by feature disentanglement and [60] added unseen IDs to generate discriminative features based on [58]. In [61], the network extracted appearance invariant features by generating corresponding fake images.

3.3.2. Extra Modal

Aside from modality translation, some works alleviated the modality discrepancy by introducing an additional third modality. In 2020, Li et al. [62] first introduced an "X" modality as the middle modality to eliminate the cross-modality discrepancy. Subsequently, Huang et al. [63] learned the shared features of images from both modalities to guide the generation of extra images. Ye et al. [64] bridged the gap between the two modalities by generating 3-channel grayscale images. Miao et al. [65] introduced two novel relevant modalities to investigate modality invariant representations. In [66], the authors reduced the cross-modality discrepancy by fusing the two modalities. Wei et al. [67] combined information from visible and infrared images to generate syncretic modality, which can help the network extract modality invariant representations. Zhang et al. [68] projected the images from both modalities into a consolidated subspace to mitigate the modality discrepancy.

3.4. Other Methods

Besides the aforementioned methods, some works also alleviated the impact of a large modality discrepancy by introducing some other technologies. Almost all existing works bridge the gap between the two modalities by manually designing feature extraction modules. Such a manually designed routine usually requires plenty of domain knowledge and practical experience. Therefore, Fu et al. [69] and Chen et al. [70] proposed a cross-modality neural architecture search method and a neural feature search method, respectively, to automatically realize the process of feature extraction. Inspired by the information bottleneck

(IB), Tian et al. [71] designed a new strategy that can preserve sufficient label information while simultaneously getting rid of task-irrelevant details. Liang et al. [72] thought that the high cost of labeling person IDs in datasets greatly limits the development of supervised models. Hence, they proposed an unsupervised homogeneous–heterogeneous approach for the unsupervised visible-infrared problem. In [73], the authors used distance metrics instead of a fully connected layer to learn discriminative features. Ye et al. [74] decomposed three channels of visible images and excavated the relationship between each individual channel and infrared image.

In addition, as the tasks of ReID and VI-ReID are identical on the whole, some networks designed for ReID or other similar tasks are also valid on VI-ReID. For instance, Ye et al. [8] proposed a strong baseline for ReID, as it also shows excellent performance on VI-ReID. Jin et al. [75] combined the information removed by IN to achieve high performance. Methods aiming at solving problems of further related domains can also be applied to VI-ReID. For example, Yang et al. [76] proposed an unsupervised graph alignment method that aligns both data representations and distribution structures across the source and target domains, aiming at general cross-domain visual feature representations. The method [77] mitigates the negative effects of noise similarities in cross-modality retrieval by intra-modality distributions. These methods perform excellently on the corresponding tasks; therefore, we can learn from their ideas.

3.5. Summary

From the perspective of method categories, we make the following summaries:

- Different methods have different strengths. Non-generative-based model are dedicated to mitigating the modality gap on the feature-level (e.g., [9]), while generative-based models pay more attention to the pixel level (e.g., [11]). Compared to non-generative-based model, there is either an information loss or introducing noise in unifying the modality. However, a generative-based model can avoid the impact of color information. A more detailed summary about mainstream works is shown in Table 1.
- Combining with other techniques is a growing trend. To acquire more discriminative features, some researchers combined this task with some universal techniques (e.g., [70]), and there are also methods (e.g., [71]) that treat this issue from a fresh perspective.
- The existing setting is not in line with practical application scenarios. In some cases, the modality discrepancy is larger than the differences among IDs. However, the existing testing settings avoid this challenge by putting only single modal images into the gallery.

Table 1. A summary of non-generative-based and generative-based models.

Type	Strength	Weakness	Reference
Non-Generative			
Global Feature Learning	Modality specific and -shared feature	Miss some important nuances	[16,17,23]
Local Feature Learning	Well-aligned part features	Sensitive to noise	[26,29,30]
Contrastive Loss	Increase inter-class variance of classifier	Generally used for Siamese network	[37,38]
Triplet Loss	Form clusters	Lack of sufficient constraints to inter-class	[1,9,41]
Generative			
Single-Translation	Avoid color effects	Introduce noisy or miss some information	[11,12]
Dual-Translation	Neglect ID-irrelevant information	Introduce noise	[56,57]
Extra Modal	Pull the distance between two modalities	Miss some information	[62,64]
Others	Independent of manual design	Large time cost	[69,70]

4. Experimental Results

4.1. Datasets

We first review two prevalent VI-ReID datatsets (RegDB [78] and SYSU-MM01 [6]). Some pedestrian image samples from two datasets are shown in Figure 8.

Figure 8. Pedestrian image samples derived from RegDB [78] and SYSU-MM01 [6]. Each column denotes the same ID, and the top row and bottom row represent visible and infrared images, respectively.

RegDB [78] contains 412 different IDs, which are classified into 254 females and 158 males, and each ID corresponds to 10 visible images and 10 infrared images. From the samples shown in the first four columns of Figure 8, we can see clear differences between the images captured by two different cameras in terms of color and exposure. Generally, the dataset is randomly split into two halves for training and testing, respectively, according to the evaluation protocol in [37]. In the testing phase, the images from one modality are utilized as a query, while the gallery contains the images from the other modality. The final result is the average of 10 repeated operations.

SYSU-MM01 [6] is a public dataset for VI-ReID proposed in 2017. It contains four cameras for capturing visible images and two for capturing infrared images. Camera 1 and camera 2 are put in two bright rooms, and camera 4 and camera 5 are placed in bright outdoor scenes to capture visible images. Infrared cameras 3 and 6 are placed in a room and outdoor scene, respectively, to capture infrared images without light. There are, in total, 287,628 visible images and 15,792 infrared images of 491 different IDs in SYSU-MM01. As shown in Figure 8, the images in SYSU-MM01 are unpaired in terms of pose, viewpoint, etc.

4.2. Evaluation Metrics

Evaluation metrics play an important role when we want to test the pros and cons of a system. There are two widely used metrics for VI-ReID, named cumulative matching characteristics (CMC) [79] and mean average precision (mAP) [3].

CMC. Rank-*r* represents the probability that a correct match appears in the top-*r* search results ranked by confidence. For single shot, this is accurate. However, for multi-shot, CMC [79] cannot accurately represent a model's discriminability, as it only examines the first match of ranked result.

mAP. The other widely used metric, mAP [3], is a more comprehensive metric for measuring the performance of the VI-ReID algorithm. It reflects how forward all images with the same ID and the probe in the gallery are in the ranked sequence. Therefore, when we face the problem that two algorithms have equal performance in searching the first match, it can address it effectively. However, when a hard sample appears, mAP may still have difficulties evaluating a better one between two algorithms.

4.3. Analysis of the State of the Art with Existing Setting

The performance results of state-of-the-art methods on RegDB and SYSU-MM01 are shown in Tables 2 and 3, respectively. From the Table 2, we observe that [2] achieves superior performance rank1/mAP 91.05%/83.28% for visible to thermal query setting on RegDB. The main improvement comes from two aspects: replacing global-level features with part-level features and utilizing center-based triplet loss instead of instance-based triplet loss. As the images with the same ID but different modalities from RegDB are entirely aligned, the part-level features are more effective. In contrast, the images are not aligned well on SYSU-MM01. Hence, it is not as big a boost on SYSU-MM01. Moreover, the improvement in the loss function also plays a key role in performance enhancement. There are also some other works committed to this improvement, e.g., Ye et al. [9] adjusted the instance-to-instance-based triplet loss to the instance-to-class-center-based loss, and the performance was significantly improved on SYSU-MM01.

Table 2. Rank-*r* accuracy (%) and mAP (%) performance of state-of-the-art methods on RegDB. Bold numbers are the best results.

Approach	Venue	Visible to Infrared				Infrared to Visible			
		r = 1	*r* = 10	*r* = 20	mAP	*r* = 1	*r* = 10	*r* = 20	mAP
Zero-Padding [6]	ICCV'17	17.75	34.21	44.35	18.90	16.63	34.68	44.25	17.82
HCML [37]	AAAI'18	24.44	47.53	56.78	20.08	21.70	45.02	55.58	22.24
BDTR [9]	IJCAI'18	33.47	58.42	67.52	31.83	32.72	57.96	68.86	31.10
MAC [48]	ACM MM'19	36.43	62.36	71.63	37.03	36.20	61.68	70.99	36.63
D^2RL [56]	CVPR'19	43.40	66.10	76.30	44.10	-	-	-	-
HSME [45]	AAAI'19	50.85	73.36	81.66	47.00	50.15	72.40	81.07	46.16
AlignGAN [12]	ICCV'19	57.90	-	-	53.60	56.30	-	-	53.40
DFE [26]	ACM MM'19	70.13	86.32	91.96	69.14	67.99	85.56	91.41	66.70
eBDTR [41]	TIFS'20	34.62	58.96	68.72	33.46	34.21	58.74	68.64	32.49
MSR [16]	TIP'20	48.43	70.32	79.95	48.67	-	-	-	-
JSIA-ReID [58]	AAAI'20	48.50	-	-	49.30	48.10	-	-	48.90
EDFL [80]	Neurocomputing'20	52.58	72.10	81.47	52.98	51.89	72.09	81.04	52.13
XIV-ReID [62]	AAAI'20	62.21	83.13	91.72	60.18	-	-	-	-
FMSP [38]	IJCV'20	65.07	83.71	-	64.50	-	-	-	-
DDAG [29]	ECCV'20	69.34	86.19	91.49	63.46	68.06	85.15	90.31	61.80
Hi-CMD [57]	CVPR'20	70.93	86.39	-	66.04	-	-	-	-
cm-SSFT [17]	CVPR'20	72.30	-	-	72.90	71.00	-	-	71.70
MACE [18]	TIP'20	72.37	88.40	93.59	69.09	72.12	88.07	93.07	68.57
DG-VAE [21]	ACM MM'20	72.97	86.89	-	71.78	-	-	-	-
CoAL [19]	ACM MM'20	74.12	90.23	94.53	69.87	-	-	-	-
SIM [46]	IJCAI'20	74.47	-	-	75.29	75.24	-	-	78.30
CDP [81]	TIP'21	65.00	83.50	89.60	62.70	-	-	-	-
expAT [1]	TIP'21	66.48	-	-	67.31	67.45	-	-	66.51
CPN [44]	TIP'21	68.59	84.81	98.33	69.20	-	-	-	-
AGW [8]	TPAMI'21	70.05	-	-	66.37	-	-	-	-
HAT [64]	TIFS'21	71.83	87.16	92.16	67.56	70.02	86.45	91.61	66.30
FMI [71]	CVPR'21	73.20	-	-	71.60	71.80	-	-	70.10
MSO [51]	ACM MM'21	73.60	88.60	-	66.90	74.60	88.70	-	67.50
LbA [82]	ICCV'21	74.17	-	-	67.64	72.43	-	-	65.46
SFANet [55]	TNNLS'21	76.31	91.02	94.27	68.00	70.15	85.24	89.27	63.77
CICL [23]	AAAI'21	78.80	-	-	69.40	77.90	-	-	69.40
MCLNet [24]	ICCV'21	80.31	92.70	96.03	73.07	75.93	90.93	94.59	69.49
NFS [70]	CVPR'21	80.54	91.96	95.07	72.10	77.95	90.45	93.62	69.79
GECNet [11]	TCSVT'21	82.33	92.72	95.49	78.45	78.93	91.99	95.44	75.58
MPANet [10]	CVPR'21	83.70	-	-	80.90	82.80	-	-	80.70
SMCL [67]	ICCV'21	83.93	-	-	79.83	83.05	-	-	78.57
CM-NAS [69]	CVPR'21	84.54	95.18	97.85	80.32	82.57	94.51	97.37	78.31
CAJ [74]	ICCV'21	85.03	95.49	97.54	79.14	84.75	95.33	97.51	77.82
MPMN [30]	TMM'21	86.56	96.68	98.28	82.91	84.62	95.51	97.33	79.49
HCTL [2]	TMM'21	91.05	97.16	98.57	83.28	**89.30**	96.41	98.16	**81.46**
AMC-Net [20]	Neurocomputing'21	91.21	**98.16**	**99.22**	81.61	89.03	**97.62**	**99.27**	79.85
MMN [68]	ACM MM'21	**91.60**	97.70	98.90	**84.10**	87.50	96.00	98.10	80.50

Table 3. Rank-*r* accuracy (%) and mAP (%) performances of state-of-the-art methods on SYSU-MM01. Bold numbers are the best results.

Approach	All Search								Indoor Search							
	Single-Shot				Multi-Shot				Single-Shot				Multi-Shot			
	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
Zero-Padding [6]	14.8	54.1	71.3	16.0	19.1	61.4	78.4	10.9	20.6	68.4	85.8	26.9	24.4	75.9	91.3	18.6
HCML [37]	14.3	53.2	69.2	16.2	-	-	-	-	24.5	73.3	86.7	30.1	-	-	-	-
BDTR [9]	17.0	55.4	72.0	19.7	-	-	-	-	-	-	-	-	-	-	-	-
cmGAN [47]	26.9	67.5	80.6	27.8	31.5	72.7	85.0	22.3	31.6	77.2	89.2	42.2	37.0	80.9	92.1	32.8
TCMDL [42]	16.9	58.8	76.6	19.3	-	-	-	-	21.6	71.4	87.9	32.3	-	-	-	-
HSME [45]	20.7	62.7	78.0	23.1	-	-	-	-	-	-	-	-	-	-	-	-
D^2RL [56]	28.9	70.6	82.4	29.2	-	-	-	-	-	-	-	-	-	-	-	-
SDL [22]	28.1	70.2	83.7	29.0	-	-	-	-	32.6	80.5	90.7	39.6	-	-	-	-
MAC [48]	33.2	79.0	90.1	36.2	-	-	-	-	33.4	82.5	93.7	45.0	-	-	-	-
AlignGAN [12]	42.4	85.0	93.7	40.7	51.5	89.4	95.7	33.9	45.9	87.6	94.4	54.3	57.1	92.7	97.4	45.3
DFE [26]	48.7	88.9	95.3	48.6	54.6	91.6	96.8	42.1	52.3	89.9	95.9	59.7	59.6	94.5	98.1	50.6
eBDTR [41]	27.8	67.3	81.3	28.4	-	-	-	-	32.5	77.4	89.6	42.5	-	-	-	-
Hi-CMD [57]	34.9	77.6	-	35.9	-	-	-	-	-	-	-	-	-	-	-	-
MSR [16]	37.3	83.4	93.3	38.1	43.9	86.9	95.7	30.5	39.6	89.3	97.7	50.9	46.6	93.6	98.8	40.1
JSIA-ReID [58]	38.1	80.7	89.9	36.9	45.1	85.7	93.8	29.5	43.8	86.2	94.2	52.9	52.7	91.1	96.4	42.7
XIV-ReID [62]	49.9	89.8	96.0	50.7	-	-	-	-	-	-	-	-	-	-	-	-
MACE [18]	51.6	87.3	94.4	50.1	-	-	-	-	57.4	93.0	97.5	64.8	-	-	-	-
CML [66]	51.8	92.7	97.7	51.2	56.3	94.1	98.1	43.4	55.0	94.4	99.4	63.7	60.4	96.9	99.5	53.5
DDAG [29]	54.8	90.4	95.8	53.0	-	-	-	-	61.0	94.1	98.4	68.0	-	-	-	-
SIM [46]	56.9	-	-	60.9	-	-	-	-	-	-	-	-	-	-	-	-
HC [83]	57.0	91.5	96.8	55.0	-	-	-	-	59.7	92.1	96.2	64.9	-	-	-	-
DG-VAE [21]	59.5	93.8	-	58.5	-	-	-	-	-	-	-	-	-	-	-	-
cm-SSFT [17]	61.6	89.2	93.9	63.2	63.4	91.2	95.7	62.0	70.5	94.9	97.7	72.6	73.0	96.3	99.1	72.4
CDP [81]	38.0	82.3	91.7	38.4	-	-	-	-	-	-	-	-	-	-	-	-
expAT [1]	38.6	76.6	86.4	38.6	-	-	-	-	-	-	-	-	-	-	-	-
AGW [8]	47.5	-	-	47.7	-	-	-	-	54.2	-	-	63.0	-	-	-	-
GECNet [11]	53.4	89.9	95.7	51.8	-	-	-	-	60.6	94.3	98.1	62.9	-	-	-	-
HAT [64]	55.3	92.1	97.4	53.9	-	-	-	-	62.1	95.8	99.2	69.4	-	-	-	-
LbA [82]	55.4	-	-	54.1	-	-	-	-	58.5	-	-	66.3	-	-	-	-
NFS [70]	56.9	91.3	96.5	55.5	63.5	94.4	97.8	48.6	62.8	96.5	99.1	69.8	70.0	97.7	99.5	61.5
CICL [23]	57.2	94.3	98.4	59.3	60.7	95.2	98.6	52.6	66.6	**98.8**	**99.7**	74.7	73.8	99.4	**99.9**	68.3
CPN [44]	57.3	92.6	97.1	56.9	63.1	93.9	97.4	50.7	59.3	94.5	98.4	66.7	66.3	97.4	99.8	58.5
MSO [51]	58.7	92.1	-	56.4	65.9	94.4	-	49.6	63.1	96.6	-	70.3	72.1	97.8	-	61.7
FMI [71]	60.0	94.2	98.1	58.8	-	-	-	-	66.1	96.6	99.4	73.0	-	-	-	-
HCTL [2]	61.7	93.1	97.2	57.5	-	-	-	-	63.4	91.7	95.3	68.2	-	-	-	-
CM-NAS [69]	62.0	92.9	97.3	60.0	68.7	94.9	98.4	53.5	67.0	97.0	99.3	73.0	76.5	98.7	**99.9**	65.1
MCLNet [24]	65.4	93.3	97.1	62.0	-	-	-	-	72.6	96.7	99.2	76.6	-	-	-	-
SFANet [55]	65.7	93.0	97.0	60.8	-	-	-	-	71.6	96.6	99.5	80.0	-	-	-	-
SMCL [67]	67.4	92.9	96.8	61.8	72.2	90.7	94.3	54.9	68.8	96.6	98.8	75.6	79.6	95.3	98.0	66.6
CAJ [74]	69.9	95.7	98.5	66.9	-	-	-	-	76.3	97.9	99.5	80.4	-	-	-	-
MMN [68]	**70.6**	**96.2**	**99.0**	66.9	-	-	-	-	76.2	97.2	99.3	79.6	-	-	-	-
MPANet [10]	**70.6**	**96.2**	98.8	**68.2**	**75.6**	**97.9**	**99.4**	**62.9**	**76.7**	98.2	99.6	**81.0**	**84.2**	**99.7**	**99.9**	**75.1**

As shown in Table 3, MPANet [10] performs best on SYSU-MM01 [10]. As the infrared modality contains limited information, the difference among the infrared IDs is extremely inconspicuous. Most existing methods deal with the cross-modality discrepancy by proposing novel loss functions or introducing other modalities. In addition to addressing the modality discrepancy, MPANet exploits the nuances among different infrared images to extract more discriminative features.

4.4. Results of the State-of-the-Arts with New Setting

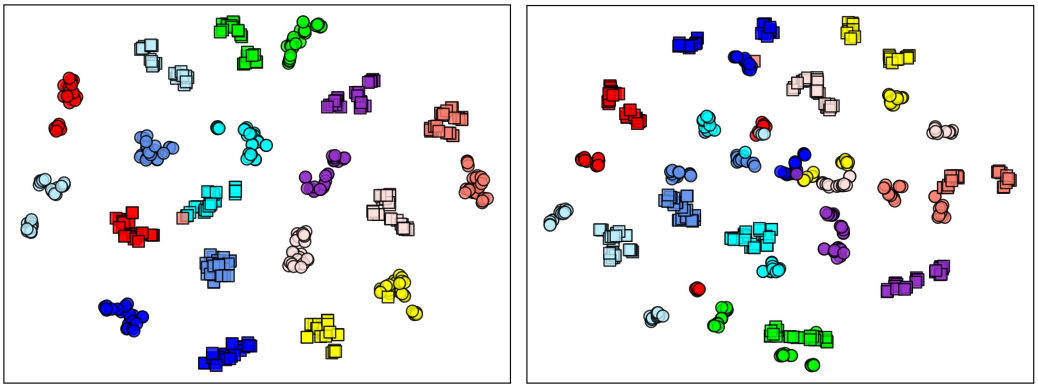
To evaluate the new proposed testing setting, we propose new testing datasets based on RegDB and SYSU-MM01, named RegDB_Mix and SYSU-MM01_Mix, respectively. The reconstructed datasets have the same number of identities and images with original datasets. Rather than putting images of the two modalities into the query and gallery respectively, we mix visible and infrared images, and remove the images in the gallery which have the same

modality and identity as the images in the query. We train and test these approaches on a single NVIDIA Tesla P100 GPU. The other settings are consistent with those in the original paper. The multi-shot results on the two datasets are presented in Table 4. Compared to Tables 2 and 3, we observe that the Rank-1 and mAP both have a great degree of decline. Specially, on RegDB_Mix, the CM-NAS [69] achieves the Rank-1 accuracy of 42.50% and mAP of 41.73%, approximately only half the value of Rank-1 and mAP on RegDB with the visible to infrared mode. On SYSU-MM01_Mix, the CM-NAS achieves the Rank-1 accuracy of 36.48% and mAP of 29.51%, significantly dropping the Rank-1 accuracy by 25.51% and mAP by 30.51% on SYSU-MM01 with all-search and single-shot modes.

Table 4. Rank-1 accuracy (%) and mAP (%) performance of state-of-the-art methods on RegDB_Mix and SYSU-MM01_Mix. Bold numbers are the best results. All approaches listed in this table are reproduced by us.

Approach	Venue	RegDB_Mix				SYSU-MM01_Mix			
		$r = 1$	$r = 10$	$r = 20$	mAP	$r = 1$	$r = 10$	$r = 20$	mAP
DDAG [29]	ECCV'20	36.28	52.05	59.41	33.85	32.16	55.05	66.27	25.44
LbA [82]	ICCV'21	38.57	55.07	62.38	35.98	34.80	**65.38**	**77.27**	27.46
GECNet [11]	TCSVT'21	41.76	57.94	64.66	37.49	22.29	60.71	75.58	14.19
AGW [8]	TPAMI'21	42.02	57.03	63.95	38.11	30.87	64.81	77.12	23.70
CM-NAS [69]	CVPR'21	42.50	51.37	56.04	41.73	**36.48**	54.08	62.70	**29.51**
HCTL [2]	TMM'21	**58.41**	**70.85**	**76.21**	**52.59**	33.62	52.24	61.47	26.97

To better present the challenges posed by the new setting, we randomly select 10 IDs from the testing set to visualize the distributions of learned features by t-SNE [84]. Here, we choose AGW [8] to extract features of the selected images. As shown in Figure 9, most the features extracted by AGW can be clustered well. However, compared to Figure 9a, many infrared images with different IDs (e.g., blue, red, and yellow) are gathered in Figure 9b. This means that pedestrian images with a certain ID are more likely to be influenced by the images with other different IDs. In fact, this is also more in line with reality, as infrared images contain less information. Hence, they are more difficult to discern by VI-ReID systems.



(a) Testing data with existing setting (b) Testing data with new setting

Figure 9. Visualization of the feature extracted with AGW [8] distributions. A total of 10 IDs are randomly selected from the testing set of SYSU-MM01. Here, samples with the same color indicate they are of the same person. The markers "circle" and "square" represent the images from infrared and visible modalities, respectively.

5. Conclusions and Future Directions

With the increase in functional application requirements, VI-ReID has attracted some researchers' attention. This paper presents a comprehensive survey of VI-ReID. We first compare it with ReID in detail to show the different challenges of VI-ReID. With powerful

deep learning techniques, VI-ReID has achieved remarkable progress, and we divide the existing methods into two categories: non-generative-based and generative-based methods. For the non-generative-based model, we analyze the method in terms of feature learning, metric learning, and training strategy. In contrast, the generative-based model applies modality translation to bridge the modality gap. Finally, we describe standard datasets in detail, evaluation metrics, and performance of the state-of-the-art methods on two datasets.

From Tables 2 and 3, we observe that the performance of VI-ReID on two public datasets has improved a lot in recent years. Meanwhile, the complexity of networks architecture has also increased. Among the existing network architectures, feature learning and metric learning are the essential modules. The primary function of feature learning is to extract modality specific and -shared features. Recently, some works aiming to extract effective features have become more popular, including global-local features fusion. Here, the distance between two features with the same ID would be pulled, while the distance between two features with different IDs would be pushed by distance metric learning.

From the experimental results, we observe the following directions in VI-ReID:

- One-stream network architecture. In terms of testing baseline, we believe that the new testing baseline with a more practical setting is more valuable to research than the existing setting. Considering that existing two-stream network architectures cannot validly solve the challenges of the new setting, a one-stream network that can extract more robust and effective features of two heterogeneous modalities may be a trend.
- Weakly supervised or self-supervised. Considering the difficulties of obtaining a sufficient amount of high-confidence data, we should concentrate on those data with no labels or low label confidence. The approaches, such as those of [85,86] of leveraging this kind of data to address related issues is highly advanced in ReID. We believe that numerous works about weakly supervised or self-supervised data will appear in VI-ReID in the future.
- Transfer learning. As the number of neural networks grows, the structures become more and more complicated, we expect that the neural network can draw on some current resources when facing comparable tasks. Further research on transfer learning, which has been widely used in ReID [87–89], may be a great direction in VI-ReID.

Limitations. First of all, this review draws on the authors' summary of the literature analysis. Although we aim to be objective in the analysis process, we still cannot avoid a robust subjective tone. Thus, all descriptions are built on personal opinions. Moreover, this survey classifies the networks according to the criteria in [8]. In contrast, we focus on VI-ReID, which accounts for a small percentage of [8]. Finally, this review only covers the research results published in mainstream conferences or journals in this field. The main reason is perhaps that these articles sufficiently represent the research methods and research trends in the area.

Author Contributions: Conceptualization, X.Z. and Z.W.; methodology, H.Z. and X.Z.; software, H.Z.; resources, X.Z.; writing—original draft preparation, H.Z.; writing—review and editing, X.Z., W.H., K.J., W.L. and Z.W.; funding acquisition, X.Z. and W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Department of Science and Technology, Hubei Provincial People's Government under grants 2021CFB513 and 2021CFB281.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Not applicable

Acknowledgments: The authors would like to thank the anonymous reviewers and the editor for their careful reviews and constructive suggestions to help us improve the quality of this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ye, H.; Liu, H.; Meng, F.; Li, X. Bi-Directional Exponential Angular Triplet Loss for visible-Infrared Person Re-Identification. *IEEE Trans. Image Process.* **2021**, *30*, 1583–1595.
2. Liu, H.; Tan, X.; Zhou, X. Parameters Sharing Exploration and Hetero-Center based Triplet Loss for Visible-Thermal Person Re-Identification. *IEEE Trans. Multimed.* **2020**, *23*, 4414–4425.
3. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable Person Re-identification: A Benchmark. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1116–1124.
4. Ristani, E.; Solera, F.; Zou, R.S.; Cucchiara, R.; Tomasi, C. Performance Measures and a Data Set for Multi-target, Multi-camera Tracking. In Proceedings of the 2016 Springer European Conference on Computer Vision Workshops, Amsterdam, The Netherlands, 8–16 October 2016; pp. 17–35.
5. Wang, G.; Lai, J.; Huang, P.; Xie, X. Spatial-Temporal Person Re-Identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 8933–8940.
6. Wu, A.; Zheng, W.; Yu, H.; Gong, S.; Lai, J. RGB-Infrared Cross-Modality Person Re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5390–5399.
7. Baltrusaitis, T.; Ahuja, C.; Morency, L. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443.
8. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, https://doi.org/10.1109/TPAMI.2021.3054775.
9. Ye, M.; Wang, Z.; Lan, X.; Yuen, P.C. Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, 13–19 July 2018; pp. 1092–1099.
10. Wu, Q.; Dai, P.; Chen, J.; Lin, C.; Wu, Y.; Huang, F.; Ji, R. Discover Cross-Modality Nuances for Visible-Infrared Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 4330–4339.
11. Zhong, X.; Lu, T.; Huang, W.; Ye, M.; Jia, X.; Lin, C. Grayscale Enhancement Colorization Network for Visible-infrared Person Re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2021**. https://doi.org/10.1109/TCSVT.2021.3072171.
12. 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 IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 3622–3631.
13. Wang, Z.; Wang, Z.; Zheng, Y.; Wu, Y.; Zeng, W.; Satoh, S. Beyond Intra-modality: A Survey of Heterogeneous Person Re-identification. In Proceedings of the 2020 29th IJCAI International Joint Conference Artificial Intelligence, Yokohama, Japan, 7-15 January 2021; pp. 4973–4980.
14. Leng, Q.; Ye, M.; Tian, Q. A Survey of Open-World Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 1092–1108.
15. Chen, X.; Fu, C.; Zhao, Y.; Zheng, F.; Song, J.; Ji, R.; Yang, Y. Salience-Guided Cascaded Suppression Network for Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3297–3307.
16. Feng, Z.; Lai, J.; Xie, X. Learning Modality-Specific Representations for Visible-Infrared Person Re-Identification. *IEEE Trans. Image Process.* **2020**, *29*, 579–590.
17. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13376–13386.
18. 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.
19. Wei, X.; Li, D.; Hong, X.; Ke, W.; Gong, Y. Co-Attentive Lifting for Infrared-Visible Person Re-Identification. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1028–1037.
20. Wang, H.; Zhao, J.; Zhou, Y.; Yao, R.; Chen, Y.; Chen, S. AMC-Net: Attentive modality-consistent network for visible-infrared person re-identification. *Neurocomputing* **2021**, *463*, 226–236.
21. Pu, N.; Chen, W.; Liu, Y.; Bakker, E.M.; Lew, M.S. Dual Gaussian-based Variational Subspace Disentanglement for Visible-Infrared Person Re-Identification. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2149–2158.
22. 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.
23. Zhao, Z.; Liu, B.; Chu, Q., Lu, Y.; Yu, N. Joint Color-irrelevant Consistency Learning and Identity-aware Modality Adaptation for Visible-infrared Cross Modality Person Re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; pp. 3520–3528.
24. Hao, X.; Zhao, S.; Ye, M.; Shen, J. Cross-Modality Person Re-Identification via Modality Confusion and Center Aggregation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11-17 October 2021; pp. 16403–16412.

25. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; Yang, Y. Improving person re-identification by attribute and identity learning. *Pattern Recognit.* **2019**, *95*, 151–161.
26. Hao, Y.; Wang, N.; Gao, X.; Li, J.; Wang, X. Dual-alignment Feature Embedding for Cross-modality Person Re-identification. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 57–65.
27. Wei, Z.; Yang, X.; Wang, N.; Song, B.; Gao, X. ABP: Adaptive Body Partition Model For Visible Infrared Person Re-Identification. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME 2020), London, UK, 6–10 July 2020; pp. 1–6.
28. Liu, S.; Zhang, J. Local Alignment Deep Network for Infrared-Visible Cross-Modal Person Re-identification in 6G-Enabled Internet of Things. *IEEE Internet Things J.* **2020**, *8*, 15170–15179.
29. 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 2020 16th European Conference Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 229–247.
30. Wang, P.; Zhao, Z.; Su, F.; Zhao, Y.; Wang, H.; Yang, L.; Li, Y. Deep Multi-Patch Matching Network for Visible Thermal Person Re-Identification. *IEEE Trans. Multimed.* **2021**, *23*, 1474–1488.
31. 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.* **2021**. https://doi.org/10.1109/TNNLS.2021.3059713.
32. Zhang, L.; Du, G.; Liu, F.; Tu, H.; Shu, X. Global-Local Multiple Granularity Learning for Cross-Modality Visible-Infrared Person Reidentification. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**. https://doi.org/10.1109/TNNLS.2021.3085978.
33. Dai, H.; Xie, Q.; Li, J.; Ma, Y.; Li, L.; Liu, Y. Visible-infrared Person Re-identification with Human Body Parts Assistance. In Proceedings of the 2021 International Conference on Multimedia Retrieval, Taipei, Taiwan, 21–24 August 2021; pp. 631–637.
34. Zhang, S.; Yang, Y.; Wang, P.; Liang, G.; Zhang, X.; Zhang, Y. Attend to the Difference: Cross-Modality Person Re-Identification via Contrastive Correlation. *IEEE Trans. Image Process.* **2021**, *30*, 8861–8872.
35. Ye, M.; Cheng, Y.; Lan, X.; Zhu, H. Improving Night-Time Pedestrian Retrieval With Distribution Alignment and Contextual Distance. *IEEE Trans. Ind. Inform.* **2020**, *16*, 615–624.
36. Zhuang, Z.; Wei, L.; Xie, L.; Ai, H.; Tian, Q. Camera-based Batch Normalization: An Effective Distribution Alignment Method for Person Re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 374–387.
37. Ye, M.; Lan, X.; Li, J.; Yuen, P.C. Hierarchical Discriminative Learning for Visible Thermal Person Re-Identification. In Proceedings of the 2018 32th AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 7501–7508.
38. Wu, A.; Zheng, W.; Gong, S.; Lai, J. RGB-IR Person Re-identification by Cross-Modality Similarity Preservation. *Int. J. Comput. Vis.* **2020**, *128*, 1765–1785.
39. Ding, S.; Lin, L.; Wang, G.; Chao, H. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognit.* **2015**, *48*, 2993–3003.
40. Wang, G.; Lai, J.; Xie, X. P2SNet: Can an Image Match a Video for Person Re-Identification in an End-to-End Way? *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 2777–2787.
41. 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.
42. Zhang, P.; Xu, J.; Wu, Q.; Huang, Y.; Zhang, J. Top-Push Constrained Modality-Adaptive Dictionary Learning for Cross-Modality Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 4554–4566.
43. Cai, X.; Liu, L.; Zhu, L.; Zhang, H. Dual-modality hard mining triplet-center loss for visible infrared person re-identification. *Knowl. Based Syst.* **2021**, *215*, 106772.
44. Zhang, Q.; Lai, J.; Xie, X. Learning Modal-Invariant Angular Metric by Cyclic Projection Network for VIS-NIR Person Re-Identification. *IEEE Trans. Image Process.* **2021**, *30*, 8019–8033.
45. Hao, Y.; Wang, N.; Li, J.; Gao, X. HSME: Hypersphere Manifold Embedding for Visible Thermal Person Re-Identification. In Proceedings of the 2019 33th AAAI Conference on Artificial Intelligence, New Orleans, Honolulu, HI, USA, 27 January–1 February 2019; pp. 8385–8392.
46. Jia, M.; Zhai, Y.; Lu, S.; Ma, S.; Zhang, J. A Similarity Inference Metric for RGB-Infrared Cross-Modality Person Re-identification. In Proceedings of the 2020 29th IJCAI International Joint Conference Artificial Intelligence, Yokohama, Japan, 7-15 January 2021; pp. 1026–1032.
47. Dai, P.; Ji, R.; Wang, H.; Wu, Q.; Huang, Y. Cross-Modality Person Re-Identification with Generative Adversarial Training. In Proceedings of the 2018 27th IJCAI International Joint Conference Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 677–683.
48. Ye, M.; Lan, X.; Leng, Q. Modality-aware Collaborative Learning for Visible Thermal Person Re-Identification. In Proceedings of the 2019 27th ACM International Conference Multimedia, Nice, France, 21–25 October 2019; pp. 347–355.
49. Zhang, Z.; Wang, S. Visible Thermal Person Re-identification via Mutual Learning Convolutional Neural Network in 6G-Enabled Visual Internet of Things. *IEEE Internet Things J.* **2020**, *8*, 15259–15266.
50. Ling, Y.; Luo, Z.; Lin, Y.; Li, S. A Multi-Constraint Similarity Learning with Adaptive Weighting for Visible-Thermal Person Re-Identification. In Proceedings of the 2021 30th IJCAI International Joint Conference Artificial Intelligence, Montreal, QC, Canada, 19–27 August 2021; pp. 845–851.

51. Gao, Y.; Liang, T.; Jin, Y.; Gu, X.; Liu, W.; Li, Y.; Lang, C. MSO: Multi-Feature Space Joint Optimization Network for RGB-Infrared Person Re-Identification. In Proceedings of the ACM International Conference Multimedia, ChengDu, China, 20–24 October 2021; pp. 5257–5265.
52. Anwar, S.; Tahir, M.; Li, C.; Mian, A.; Khan, F.S.; Muzaffar, A.W. Image Colorization: A Survey and Dataset. *arXiv* **2020** arXiv:2008.10774.
53. Zhong, X.; Lu, T.; Huang, W.; Yuan, J.; Liu, W.; Lin, C. Visible-infrared Person Re-identification via Colorization-based Siamese Generative Adversarial Network. In Proceedings of the 2020 ACM International Conference Multimedia Retrieval, Dublin, Ireland, 8–11 June 2020; pp. 421–427.
54. Kniaz, V.V.; Knyaz, V.A.; Hladuvka, J.; Kropatsch, W.G.; Mizginov, V. ThermalGAN: Multimodal Color-to-Thermal Image Translation for Person Re-identification in Multispectral Dataset. In Proceedings of the 2018 European Conference Computer Vision Workshops, Munich, Germany, 8–14 September 2018; pp. 606–624.
55. Liu, H.; Ma, S.; Xia, D.; Li, S. SFANet: A Spectrum-Aware Feature Augmentation Network for Visible-Infrared Person Re-Identification.*IEEE Trans. Neural Netw. Learn. Sys.* **2021**. https://doi.org/10.1109/TNNLS.2021.3105702.
56. Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.; Satoh, S. Learning to Reduce Dual-Level Discrepancy for Infrared-Visible Person Re-Identification. In Proceedings of the 2019 IEEE/CVF Conference Computer Vision Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 618–626.
57. 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 the2020 IEEE/CVF Conference Computer Vision Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10254–10263.
58. 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. In Proceedings of the 2020 34th AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12144–12151.
59. Hu, B.; Liu, J.; Zha, Z.J. Adversarial Disentanglement and Correlation Network for RGB-Infrared Person Re-Identification. In Proceedings of the 2021 IEEE International Conference on Multimedia Expo, Shenzhen, China, 5–9 July 2021.
60. Wang, G.; Yang, Y.; Zhang, T.; Cheng, J.; Hou, Z.; Tiwari, P.; Pandey, H.M. Cross-modality Paired-images Generation and Augmentation for Visible-infrared Person Re-identification. *Neural Netw.* **2020**, *128*, 294–304.
61. Chen, Y.; Zhang, S.; Qi, Z. MAENet: Boosting Feature Representation for Cross-Modal Person Re-Identification with Pairwise Supervision. In Proceedings of the2020 ACM International Conference Multimedia Retrieval, Dublin, Ireland, 8–11 June 2020; pp. 442–449.
62. Li, D.; Wei, X.; Hong, X; Gong, Y. Infrared-Visible Cross-Modal Person Re-Identification with an X Modality. In Proceedings of the 2020 34th AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 4610–4617.
63. 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.* **2021**. https://doi.org/10.1109/TMM.2021.3067760.
64. Ye, M.; Shen, J.; Shao, L. Visible-Infrared Person Re-Identification via Homogeneous Augmented Tri-Modal Learning. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 728–739.
65. Miao, Z.; Liu, H.; Shi, W.; Xu, W.; Ye, H. Modality-aware Style Adaptation for RGB-Infrared Person Re-Identification. In Proceedings of the 30th IJCAI International Joint Conference Artificial Intelligence, Montreal, QC, Canada, 19–27 August 2021; pp. 916–922.
66. Ling, Y.; Zhong, Z.; Luo, Z.; Rota, P.; Li, S.; Sebe, N. Class-Aware Modality Mix and Center-Guided Metric Learning for Visible-Thermal Person Re-Identification. In Proceedings of the 28th ACM International Conference Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 889–897.
67. 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 Computer Vision, Virtual, 11-17 October 2021.
68. Zhang, Y.; Yan, Y.; Lu, Y.; Wang, H. Towards a Unified Middle Modality Learning for Visible-Infrared Person Re-Identification. In Proceedings of the 2021 ACM International Conference Multimedia, ChengDu, China, 20–24 October 2021; pp. 788–796.
69. 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 Computer Vision, Virtual, 11-17 October 2021.
70. Chen, Y.; Wan, L.; Li, Z.; Jing, Q.; Sun, Z. Neural Feature Search for visible-Infrared Person Re-Identification. In Proceedings of the 2021 IEEE/CVF Conference Computer Vision Pattern Recognition, Virtual, 19-25 June 2021.
71. Tian, X.; Zhang, Z.; Lin, S.; Qu, Y.; Xie, Y.; Ma, L. Farewell to Mutual Information: Variational Distillation for Cross-Modal Person Re-Identification. In Proceedings of the 2021 IEEE/CVF Conference Computer Vision Pattern Recognition, Virtual, Virtual, 19-25 June 2021.
72. Liang, W.; Wang, G.; Lai, J.; Xie, X. Homogeneous-to-Heterogeneous: Unsupervised Learning for RGB-Infrared Person Re-Identification. *IEEE Trans. Image Process.* **2021**, *30*, 6392–6407.
73. Tekeli, N.; Can, A.B. Distance Based Training for Cross-Modality Person Re-Identification. In Proceedings of the 2019 IEEE/CVF International Conference Computer Vision Workshops, Seoul, Korea, 27–28 October 2019; pp. 4540–4549.
74. Ye, M.; Ruan, W.; Du, B.; Mike, Z. Channel Augmented Joint Learning for Visible-Infrared Recognition. In Proceedings of the 2021 IEEE/CVF International Conference Computer Vision, Virtual, 11-17 October 2021.

75. Jin, X.; Lan, C.; Zeng, W.; Chen, Z.; Zhang, L. Style Normalization and Restitution for Generalizable Person Re-Identification. In Proceedings of the2020 IEEE/CVF Conference Computer Vision Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3140–3149.
76. Yang, B.; Yuen, P.C. Cross-Domain Visual Representations via Unsupervised Graph Alignment. In Proceedings of the 2019 33th AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 5613–5620.
77. Yang, F.; Wang, Z.; Xiao, J.; Satoh, S. Mining on Heterogeneous Manifolds for Zero-Shot Cross-Modal Image Retrieval. In Proceedings of the 2020 34th AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 12589–12596.
78. 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.
79. Wang, X.; Doretto, G.; Sebastian, T.; Rittscher, J.; Tu, P.H. Shape and Appearance Context Modeling. In Proceedings of the 2007 IEEE/CVF International Conference Computer Vision, Rio de Janeiro, Brazil, 14–20 October 2007; pp. 1–8.
80. Liu, H.; Cheng, J.; Wang, W.; Su, Y.; Bai H. Enhancing the discriminative feature learning for visible-thermal cross-modality person re-identification. *Neurocomputing* **2020**, *398*, 11–19.
81. Fan, X.; Luo, H.; Zhang, C.; Jiang, W. Cross-Spectrum Dual-Subspace Pairing for RGB-infrared Cross-Modality Person Re-Identification. arXiv **2020**, arXiv:2003.00213
82. Park, H.; Lee, S.; Lee, J.; Ham, B. Learning by Aligning: Visible-Infrared Person Re-identification using Cross-Modal Correspondences. *arXiv* **2021**, arXiv:2108.07422.
83. Zhu, Y.; Yang, Z.; Wang, L.; Zhao, S.; Hu, X.; Tao, D. Hetero-Center loss for cross-modality person Re-identification. *Neurocomputing* **2020**, *386*, 97–109.
84. Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
85. Wang, G.; Wang, G.; Zhang, X.; Lai, J.; Yu, Z.; Lin, L. Weakly Supervised Person Re-ID: Differentiable Graphical Learning and a New Benchmark. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2142–2156.
86. Wang, G.; Wang, K.; Wang, G.; Torr, P.; Lin, L. Solving Inefficiency of Self-supervised Representation Learning. In Proceedings of the 2021 IEEE/CVF International Conference Computer Vision, Virtual, 11-17 October 2021.
87. Lin, Y.; Wu, Y.; Yan, C.; Xu, M.; Yang, Y. Unsupervised Person Re-identification via Cross-Camera Similarity Exploration. *IEEE Trans. Image Process.* **2020**, *29*, 5481–5490.
88. Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; Bian, W.; Yang, Y. Progressive Learning for Person Re-Identification with One Example. *IEEE Trans. Image Process.* **2019**, *28*, 2872–2881.
89. Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; OuYang, W.; Yang, Y. Exploit the Unknown Gradually: One-Shot Video-Based Person Re-Identification by Stepwise Learning. In Proceedings of the 2018 IEEE/CVF Conference Computer Vision Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5177–5186.