mathematics

MDPI

Review

Survey of Cross-Modal Person Re-Identification from a Mathematical Perspective

Minghui Liu, Yafei Zhang * and Huafeng Li

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China

* Correspondence: zyfeimail@163.com

Abstract: Person re-identification (Re-ID) aims to retrieve a particular pedestrian's identification from a surveillance system consisting of non-overlapping cameras. In recent years, researchers have begun to focus on open-world person Re-ID tasks based on non-ideal situations. One of the most representative of these is cross-modal person Re-ID, which aims to match probe data with target data from different modalities. According to the modalities of probe and target data, we divided cross-modal person Re-ID into visible–infrared, visible–depth, visible–sketch, and visible–text person Re-ID. In cross-modal person Re-ID, the most challenging problem is the modal gap. According to the different methods of narrowing the modal gap, we classified the existing works into picture-based style conversion methods, feature-based modality-invariant embedding mapping methods, and modality-unrelated auxiliary information mining methods. In addition, by generalizing the aforementioned works, we find that although deep-learning-based models perform well, the black-box-like learning process makes these models less interpretable and generalized. Therefore, we attempted to interpret different cross-modal person Re-ID models from a mathematical perspective. Through the above work, we attempt to compensate for the lack of mathematical interpretation of models in previous person Re-ID reviews and hope that our work will bring new inspiration to researchers.

Keywords: cross-modal person re-identification; review; mathematical perspective

MSC: 68T99

Citation: Liu, M.; Zhang, Y.; Li, H. Survey of Cross-Modal Person Re-Identification from a Mathematical Perspective. *Mathematics* **2023**, *11*, 654. https://doi.org/10.3390/math11030654

Academic Editor: Jianping Gou

Received: 30 December 2022
Revised: 20 January 2023
Accepted: 25 January 2023
Published: 28 January 2023

1. Introduction

Person re-identification (Re-ID) [1] originates from multi-target multi-camera tracking [2], the purpose of which is to retrieve pedestrians' identities from non-overlapping cameras, as shown in Figure 1. Recently, most Re-ID models [3,4] have been trained on visible image datasets. However, single-modal person Re-ID models are not applicable to complex situations in real scenarios. For example, suspects often conduct criminal activities at night, and visible-light imaging equipment has difficulty capturing the appearance information of pedestrians under low-light conditions. By contrast, infrared imaging equipment is not affected by light conditions and can collect clearer pedestrian images at night. Therefore, for person Re-ID across daytime and nighttime scenes, cross-modal retrieval between infrared and visible images can improve the performance of a person Re-ID model; such a model is called an infrared–visible cross-modal person Re-ID model [5]. Further, if a surveillance camera is damaged or is not installed in the place in which a suspect conducts criminal activities, it is necessary to match witnesses' textual descriptions of the suspect's characteristics with pedestrian images captured on other occasions; this is called text–image cross-modal person Re-ID [6]. In addition, a sketch of a suspect can also be drawn according to a description from a witness, and the sketch can be matched with pedestrian images captured on other occasions; this is called sketch person Re-ID [7]. Moreover, to overcome the difficulty of person Re-ID caused by changes in pedestrians' clothing, a depth camera

can be used to obtain the structural information of pedestrians, such as their body shapes and motion signatures; this is called visible–depth person Re-ID [8]. Cross-modal person Re-ID expands the application scope of the person Re-ID task.

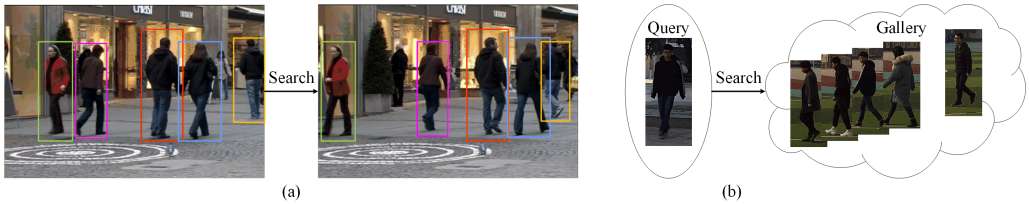

Figure 1. (**a**) Multi-target multi-camera tracking. (**b**) Person Re-ID.

However, cross-modal person Re-ID suffers from modal differences. Owing to the large differences in pedestrian features in different modalities, it may be that the differences in the modal features of the same pedestrian are larger than those in the identity features of different pedestrians. This will lead to pedestrians with different identities but the same modal being classified into one category and pedestrians with the same identity but different modalities being classified into different categories. Therefore, the key problem to be solved in cross-modal pedestrian Re-ID is the elimination of modal differences and the extraction of modality-invariant features. Taking infrared–visible cross-modal pedestrian Re-ID as an example, the main methods for reducing modal differences are generation-based and feature-alignment-based methods. Methods based on generation use a generative model to transform infrared and visible images into the same style to reduce the modal differences between them. Methods based on feature alignment are used to map images of different modalities to the same feature space and narrow the feature distribution of pedestrians with the same identity at the feature level to extract modality-invariant features.

Although cross-modal person Re-ID models have made some progress, their performance is still inferior to that of single-modal Re-ID models, which means that the research prospects of cross-modal person Re-ID are very promising. However, we find that the current reviews primarily focus on single-modal person Re-ID, and systematic and profound reviews are lacking for cross-modal person Re-ID. Moreover, current mainstream cross-modal person Re-ID methods are based on deep networks, which resemble a black box. Therefore, we attempted to analyze the existing cross-modal person Re-ID studies from a mathematical perspective, hoping to bring new inspiration to researchers. The contributions of this study are as follows.

- A detailed review of cross-modal person Re-ID studies, which fills the gaps in previous works, is performed. Based on the modalities of data, we classify cross-modal person Re-ID into visible–infrared, visible–depth, visible–sketch, and visible–text person Re-ID. Moreover, from the perspective of mitigating modal differences, we classify cross-modal person Re-ID methods into style-transformation-based methods, shared-feature-mapping-based methods, and auxiliary-information-based methods.
- An analysis of cross-modal person Re-ID methods from a mathematical perspective highlights the importance of mathematical constraints. Constrants from mathematical theory can free deep learning networks from uncontrollable defects and make black-box-like neural networks more explanatory.

In the remainder of this paper, we describe our work according to the following organization: Section 2 reviews the history of cross-modal person Re-ID, previous survey works, and datasets; Section 3 reviews current mainstream deep-learning-based methods; Section 4 presents the conclusions.

2. Literature Overview

In 2006, person Re-ID [9–13] was formally proposed as a separate task for the first time [14], instead of only as a subtask of multi-camera tracking [15–17]. However, in the early stages of development, both low-level features (e.g., color and texture) and high-level

features (e.g., semantic and structural) were extracted by handcrafted methods based on mathematical theories, such as weighted color histograms (e.g., lab color histograms) [18], edgel histograms [19], texture histograms (e.g., local binary pattern texture histograms) [20], maximally stable color regions [21], and recurrent high-structured patches [22]. These feature extraction methods completely rely on manual design and can lead to stable and reliable experimental results. However, their computational process is extremely complicated. With the rapid development of machine learning, end-to-end deep learning networks have shown excellent performance and have greatly reduced the difficulty of feature extraction.

In 2014, Li et al. [23] completed person Re-ID by using a convolutional neural network (CNN) for the first time and achieved a leap from manual feature extraction [24] to deep-learning-network-based feature extraction [25,26] in this field. However, until 2017, most of the studies in the person Re-ID field were conducted based on single-modal visible image datasets. Therefore, person Re-ID models could not realize cross-modal retrieval between visible images captured under normal light conditions and infrared images captured under low-light conditions, which limited the application scenarios of the person Re-ID task. Therefore, Wu et al. [27] proposed a multimodal dataset called *SYSU-MM*01, which contains a large number of visible and infrared images. Since then, an increasing number of researchers have turned to infrared–visible cross-modal person Re-ID and extended a series of cross-modal person Re-ID tasks, such as visible–depth, visible–sketch, and visible–text person Re-ID tasks, to meet the actual needs. By counting the number of person Re-ID papers published by the three top conferences in the computer vision field, we obtained the results shown in Table 1.

Table 1. The number of pedestrian Re-ID articles published by the top three conferences.

	ICCV		ECCV		CVPR	
	Single-Modal	**Multi-Modal**	**Single-Modal**	**Multi-Modal**	**Single-Modal**	**Multi-Modal**
2017	12	1		-	11	-
2018		-	13	1	28	-
2019	29	1		-	19	2
2020		-	21	1	22	2
2021	14	4		-	21	4
2022		-		-	14	4

The proportion of cross-modal person Re-ID studies has been increasing since 2018, but related survey work is absent. To fill the gap in previous surveys of person Re-ID works, we primarily analyze cross-modal person Re-ID works and elaborate on the problems caused by completely relying on deep learning networks. Moreover, we highlight the indispensability of mathematical theories.

2.1. Analysis of Previous Reviews

To enable a quick understanding of person Re-ID methods, many surveys have provided a careful and systematic overview of previous representative works. We provide a brief overview of these works in Table 2. In 2019, Wu et al. [28] provided a detailed overview of six types of deep models in the field of person Re-ID (i.e., identification, verification, distance metric-based, part-based, video-based, and data-augmentation-based deep models) and concluded that more high-quality training samples should be generated in the future to improve the performance of person Re-ID models while focusing on the relationship between different granularity features. In addition, this survey indicated that integrating target detection with Re-ID and using multimodal data are developmental trends for the future. However, the description of the datasets and classification of related works in this survey were not accurate. In the same year, Almasawa et al. [29] classified person Re-ID into three categories based on the training data: image-based, video-based, and

image-to-video-based. In addition, image-based works were further divided into single-modal and multimodal person Re-ID. This work demonstrated person Re-ID datasets well, which enabled researchers to gain a more intuitive understanding of the application scenarios of person Re-ID. The above-mentioned studies did not consider that person Re-ID datasets can be further divided into open-world and closed-world datasets. Then, Leng et al. [30] conducted an in-depth analysis of open-world works, which remedied the shortcomings of the previous studies in this regard. Unlike previous surveys, which focused on closed-world studies, their survey provided a comprehensive review of open-world studies for the first time; this contributed to the application of person Re-ID in the real world. However, this survey did not fully describe the most typical cross-modal person Re-ID in an open world. Similarly, Wang et al. [31] categorized person Re-ID works according to feature extraction methods and metric methods and summarized some of the mathematical theories. However, they also did not focus on cross-modal studies. Mathur et al. [32] reported that CNNs have improved the performance and efficiency of person Re-ID, but that they have also produced some inevitable problems (e.g., difficulty designing optimal features, over-reliance on labels, requiring adequate hardware resources, and overfitting to the training data). They then classified person Re-ID studies according to their methods to avoid the above problems. Ye at el. [33] published an in-depth and systematic survey that illustrated that open-world person Re-ID would be the future direction. They also proposed more effective baseline and metric losses. However, due to the lack of a mathematical foundation, newcomers cannot gain a deeper understanding based on these surveys and will ignore the use of mathematical theories in future work. Therefore, we summarize and analyze the cross-modal person Re-ID from a mathematical perspective.

Table 2. A summary of recent reviews.

Year	Survey	Contribution
2022	Deep Learning for Person Re-Identification: A Survey and Outlook [33]	A comprehensive and in-depth analysis of the pedestrian Re-ID work. Meanwhile, a baseline named AGW and a new assessment method that was introduced to complement mAP were proposed and achieved excellent performance in four types of Re-ID work.
2020	A Brief Survey of Deep Learning Techniques for Person Re-Identification [32]	Summary of the architectures used in pedestrian re-identification work and classification thereof into five categories according to the differences in architectures: a CNN, RNN, deep belief neural network, deep stacking network, and autoencoder.
2020	A Comprehensive Overview of Person Re-Identification Approaches [31]	Summary of the datasets and metric methods used for pedestrian re-identification tasks in detail and classification of the feature extraction means according to the purpose of the task.
2020	A Survey of Open-World Person Re-Identification [30]	Analysis of the open-world pedestrian re-identification work for the first time and clarification of the definition, realization, and development of the former by comparing open-world works with closed-world works. At the same time, the three development directions of generalization, scalability, and tolerance of clothing changes were summarized for open-world work.
2019	A Survey on Deep-Learning-Based Person Re-Identification Systems [29]	The work on deep-learning-based pedestrian re-identification was classified in terms of the differences in the data used and was divided into three categories: picture-based, video-based, and picture-to-video-based.
2019	Deep-learning-based methods for person re-identification: A comprehensive review [28]	Six types of models in the field of pedestrian re-recognition were analyzed in detail: depth-based recognition models, depth-based verification models, distance-metric-based models, part-based models, video-based models, and data-augmentation-based models.

2.2. Comparison of Datasets

To provide researchers with a more comprehensive understanding of person Re-ID datasets, we classified them from different perspectives. Based on the type of data, we classified them into image-based and video-based datasets. According to the number of modalities contained in the datasets, we classified them into single-modal and multimodal datasets. According to the type of modality, we further classified the multimodal datasets into RGB–infrared (IR), RGB–depth, RGB–text, and RGB–sketch datasets. To the best of our knowledge, this is the most complete compendium of multimodal datasets to date.

2.2.1. Single-Modal Datasets

Single-modal datasets contain only RGB data captured by visible-light cameras. As shown in Table 3, limited by the conditions of the hardware, the initial person Re-ID datasets [34–41] usually contained a small amount of nonconsecutive image-based data to eliminate the need for high-performance equipment. With the improvement of hardware performance, training person Re-ID models with video datasets [23,42–47] is gradually becoming a trend. Video-based datasets for person Re-ID have the following advantages: (1) less workload—the original unfiltered video sequence is used directly without data filtering, which can reduce manual operations; (2) richer information—compared with image-based datasets, video-based datasets with more information allow the network to be fully trained; (3) more temporal information—video-based datasets can completely preserve gait information, which makes it possible to mine temporal information. Nowadays, the deep-learning-based person Re-ID methods outperform methods that use handcrafted features in terms of accuracy. However, it has been found that if person Re-ID models are trained on closed-world datasets, the obtained models tend to have poor generalization when applied to open-world datasets. Taking TCLNet [48] as an example, its rank-1 accuracy could reach 96.2% on the video-based RGB dataset Duke-Video, but only 52.13% on the video-based RGB–IR cross-modal dataset VCM. Considering that cross-modal person Re-ID has wider application possibilities, an increasing number of researchers have begun to explore how to train models with multimodal datasets and extract modality-invariant features.

Table 3. Single-modal datasets.

Dataset	Year	Res	Identities/Images(Tracks)/Cameras	Types
VIPeR	2007	fixed	632/1264/2	image
iLIDS	2009	vary	119/476/2	image
GRID	2009	vary	250/1275/8	image
PRID2011	2011	fixed	200/1134/2	image
PRID-2011	2011	fixed	200/400/2	video
CUHK01	2012	fixed	971/3884/2	image
CUHK02	2013	fixed	1816/7264/10	image
CUHK03	2014	vary	1467/13,164/2	image
iLIDS-VID	2014	vary	300/600/2	video
Market-1501	2015	fixed	1501/32,668/6	image
MARS	2016	fixed	1261/20,715/6	video
DukeMTMC	2017	fixed	1404/36,411/8	image
Airport	2017	fixed	9651/39,902/6	image
MSMT17	2018	vary	4101/126,441/16	image
Duke-Video	2018	fixed	1812/4832/8	video
Duke-Tracklet	2018	fixed	1788/12,647/8	video
LPW	2018	fixed	2731/7694/4	video
LS-VID	2019	fixed	3772/14,943/15	video

2.2.2. Multimodal Datasets

Multimodal person Re-ID datasets include RGB–IR [27,49,50], RGB–depth [51–54], RGB–text [55], and RGB–sketch [56] datasets. Unlike single-modal person Re-ID, cross-modal Re-ID requires cross-retrieval among data from different modalities. For example, for an infrared–visible person Re-ID, visible images are needed to retrieve pedestrians photographed in infrared conditions, and IR images are needed to retrieve pedestrians captured by visible cameras. In addition, to address the lack of pedestrian visual data, text–image Re-ID and sketch Re-ID utilize text and sketches, respectively, to retrieve pedestrians

captured under visible-light conditions. To overcome the difficulty in person Re-ID caused by changes in pedestrians' clothing, a depth camera can be used to obtain the structural information of pedestrians, such as bones. Therefore, in RGB–depth person Re-ID, it is necessary to match the visible and depth images. Parts of multimodal datasets are shown in Figure 2. As can be seen in Figure 2, there are significant modal differences between IR, sketch, depth, text data, and visible images. Therefore, it is a challenging problem to reduce the modal differences in cross-modal person Re-ID.

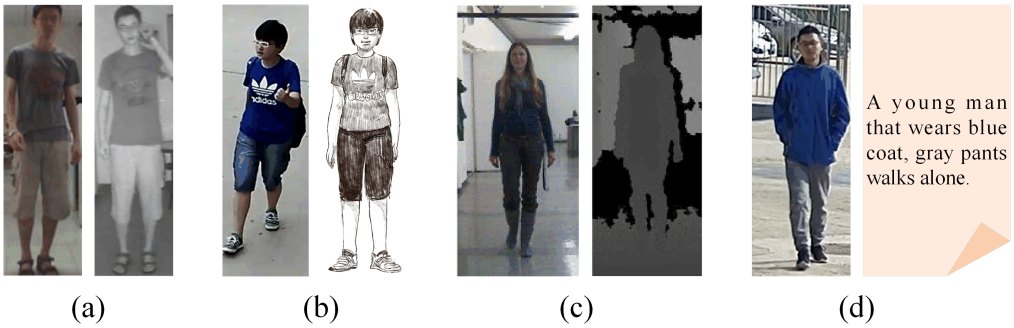

Figure 2. Parts of multi-modality datasets. (**a**) RGB–IR data. (**b**) RGB–sketch data. (**c**) RGB–depth data. (**d**) RGB–text data.

Cross-modal datasets play a crucial role in the task of cross-modal person Re-ID, and cross-modal person Re-ID datasets are summarized in this paper, as shown in Table 4. It can be seen in Table 4 that the number and scale of the cross-modal datasets are relatively small compared with those of the single-modal datasets. For example, the Sketch Re-ID dataset contains only 200 people, each with one sketch and two photos. This limits the performance of cross-modal person Re-ID models. Therefore, expanding the scale of cross-modal datasets is also a basic task that needs to be performed in the field of cross-modal person Re-ID in the future.

Table 4. Multimodal datasets.

Dataset	Year	Boxes	Identities/Data/Cameras	Types
RegDB	2017	128 × 64	412/8240/2	image-based RGB–IR
SYSU-MM01	2017	vary	491/303,420/6	image-based RGB–IR
VCM	2022	vary	927/485,122/12	video-based RGB–IR
PAVIS	2012	-	79/79/2	image-based RGB–depth
BIWI	2014	1280 × 960	50/50/2	image-based RGB–depth
IAS-Lab	2014	-	11/11/2	image-based RGB–depth
DPI-T	2016	-	12/300/5	image-based RGB–depth
CUHK-pedes	2016	-	13003/120,618/2	image-based RGB–text
ICFG-PEDES	2021	-	4102/54,522/37.2	image-based RGB–text
SKETCH	2018	-	200/600/3	image-based RGB–sketch

3. Summary of Deep-Learning-Based Cross-Modal Person Re-ID Methods

The process of person Re-ID can be roughly divided into three parts: data collection [57–59], feature extraction [60–62], and pedestrian matching [63–65]. In deep-learning-based methods, off-the-shelf target detection tools are typically used for personal data collection. CNNs, recurrent neural networks, deep belief networks, deep stacking networks, and autoencoders have been used for discriminative feature extraction. Representational and metric learning are used for pedestrian matching. In cross-modal person Re-ID, feature extraction and pedestrian matching must be studied according to the characteristics of different modalities.

3.1. Types of Cross-Modal Person Re-ID

3.1.1. Visible–Infrared Re-ID

At present, single-modal visible-light cameras with poor imaging effects at night are gradually being replaced by dual-modal cameras with an adaptively switching imaging

mechanism. Dual-modal cameras can use a visible-light camera to capture color images when the light is sufficient and use an IR camera to capture black-and-white images when the light is poor. Although a dual-modal camera can compensate for the limitation that the visible-light camera cannot work properly when the light is insufficient, visible images captured by visible-light cameras and IR images captured by infrared cameras have considerable modal differences, which introduce challenges for the visible–IR person Re-ID. To improve the accuracy of cross-modal person Re-ID, visible–IR person Re-ID should not only solve the problems faced by traditional person Re-ID (such as posture changes, angle changes, light changes, and local occlusion), but also strive to overcome imbalances in color information between the different modalities.

3.1.2. Visible–Depth Re-ID

A depth image is also called a range image, the pixel value of which is the distance from the camera to each point in the scene. A depth image directly reflects the geometric shape of a target object. Therefore, it can be used to easily solve many problems related to 3D objects and is primarily used for information acquisition in the field of motion capture. Similarly, depth images are widely used in person Re-ID to capture the body shapes and sketch information of pedestrians. In this manner, depth and visible images are fused to overcome changes in illumination and color. Ren et al. [66] proposed a multimodal uniform deep-learning method to accomplish person Re-ID. This method combined anthropometric measures and body shape information contained in depth images with appearance information contained in visible images through a multimodal fusion layer to obtain pedestrian features that were robust to noise. To further adapt to different application scenarios, Imani et al. [67] used RGB, depth, and skeleton modules to realize person Re-ID and attempted to find a modal combination that was suitable for different datasets. To overcome the effect of viewpoint variation on depth devices, Wu et al. [68] designed a depth voxel covariance descriptor that was insensitive to rotation and noise. However, the above works only considered how to use depth features to complement traditional visual features. Because environmental and clothing changes have no impact on the distance information, a depth camera can compensate for the poor performance of a visible-light camera in a low-light environment, which also provides a new approach for long-term retrieval tasks. However, it also creates a problem: that of matching depth images with visible images. Hafner et al. [8] proposed a cross-modal distillation method to learn the representation space shared by visible images and depth images and implemented visible–depth person Re-ID for the first time. This is the only visible–depth Re-ID model available.

3.1.3. Visible–Text Re-ID

In a case in which a visual image of pedestrians is missing but a textual description can be obtained, it is necessary to construct a network with the ability to perform text–image cross-modal retrieval. Because there is a large modal gap between textual and image data, text–image Re-ID studies primarily focus on enhancing the connection between the two types of data at different levels [69]. Ding et al. [70] captured the relationship between different parts of the body through a multi-view nonlocal network for better correspondence between body parts and noun phrases. Yu et al. [6] extended the re-ranking commonly used in single-modal retrieval to cross-modal retrieval and enhanced the accuracy by constructing a bidirectional coarse-to-fine framework. Inspired by the interactions between images and sentences, Wang et al. [71] proposed cross-modal adaptive message passing to adaptively control the cross-modal information flow. Niu et al. [72] divided features according to granularity and realized cross-modal fine-grained retrieval through global–global, global–local, and local–local alignments. Kansal et al. [73] proposed an end-to-end hierarchical attention image–text alignment network to solve the problems of alignment uncertainty and text complexity. Additionally, this network determined the potential relationship between image content and textual information at different levels by using a hierarchical attention alignment network. Farooq et al. [74] enhanced the

correlation between the two modalities by performing a canonical correlation analysis. In addition, attributes are often used in visible–text person Re-ID. Zha et al. [75] proposed an adversarial attribute–text-embedding network to learn discriminative modality-invariant visual and text features. Simultaneously, a visual attribute graph convolutional network was proposed to learn the visual attributes of pedestrians with better descriptiveness, interpretability, and robustness than those of appearance features. In addition to the above studies, some researchers have tried to reduce the heavier computational requirements of visible–text Re-ID work by building weakly supervised models [76] to eliminate reliance on identity annotation.

3.1.4. Visible–Sketch Re-ID

Some researchers have attempted to convert textual descriptions into sketch images [7] to alleviate the modal gap between visual and textual data. For example, the police always sketch portraits based on descriptions from eyewitnesses and match the sketches with visible images from surveillance systems. In 2018, Lu et al. [56] first proposed the visible–sketch Re-ID task and provided the first sketch Re-ID dataset. Similarly to most cross-modal retrieval tasks, they used generative adversarial networks to extract discriminative modality-invariant features in sketches and visible images. Gui et al. [77] extracted domain-invariant features by using a gradient reverse layer. Thus, it is possible to model photos and sketches in a common embedding space to solve the domain gap problem. To bridge the domain gap, unlike in the above-mentioned studies, Yang et al. [78] proposed instance-level heterogeneous domain adaptation to transfer the instance-level knowledge in an inductive transfer manner. A sketch contains only contour information and does not contain color and background information; therefore, it is particularly important to process redundant information of visible images in visible–sketch cross-modal retrieval, which makes sketch Re-ID more challenging than traditional person Re-ID.

3.2. Common Solutions for Cross-Modal Person Re-ID

In cross-modal person Re-ID, the challenge is to narrow the gap between different modal features. To achieve the cross-modal person Re-ID task, researchers have proposed a series of methods for solving the modal difference problem. These methods can be roughly classified into the following three categories. The first transforms the style of input images before feature extraction [5,79] to create consistency in the image styles under different modalities and transform heterogeneous data into traditional homogeneous data. The second embeds the features of different modalities into the same mapping space [80–82] to achieve consistency. The third extracts modality-independent discriminative features, such as gait features, to assist in cross-modal person Re-ID [83,84].

3.2.1. Style-Transformation-Based Method

Owing to modal differences, it is difficult to extract modality-invariant features directly from the original data. Therefore, it is particularly important to perform image-level style transformation in advance to bridge the modal gap.

Translation between Original Modalities

Initially, a generative adversarial network (GAN), which was constructed in 2014 by Goodfellow et al. [85], was most commonly used for style transformation. In 2017, Zheng et al. [39] applied GANs to person Re-ID for the first time and expanded their training set by generating unlabeled samples that did not belong to a known class. By using real and generated data simultaneously, semi-supervised training of the network was realized, which effectively avoided the overfitting of network parameters and further improved the network performance. In contrast to single-modal person Re-ID, cross-modal person Re-ID often uses GANs to generate heterogeneous data to enhance the relationship between different modal data or convert the cross-modal task into a single-modal one. In 2020, Wang et al. [86] first applied a GAN to a cross-modal person Re-ID and

proposed a cross-modal image-generation method for person Re-ID. The method entangled modality-invariant features by generating cross-modal image pairs so that the set-level and instance-level modality variations were simultaneously reduced.

Owing to the inconsistency between labels in training and test sets, independent pixel-level and feature-level generative adversarial training causes the two to be mismatched in testing, which severely damages a network's performance. Therefore, Wang et al. [87] used a GAN to generate sufficiently realistic fake IR images to achieve pixel-level alignment and alleviate modal differences. They then performed feature-level alignment to resolve the intra-modal differences. Finally, desirable fake IR images could be obtained by jointly using the above two alignments.

In general, GANs use generators to generate fake images, which are input into the discriminator along with real images to verify the authenticity of the images. Through adversarial learning between the generator and discriminator, the performance of the generator is continuously improved, and it can generate real enough fakes to complete cross-modal work. In visible–IR cross-modal person Re-ID, a GAN was used to convert visible images into the style of IR images, which also meant that the color information in the visible images could not be fully utilized. Unlike the previous one-way fake IR image generation, Zhong et al. [88] aimed to completely preserve the rich discriminative information in visible images. Therefore, they colored single-channel IR images to learn the correspondence between single-channel IR and three-channel visible images.

However, it is impossible to generate fake images that are the same as real images. If distorted fake images are used for network training, network performance is adversely affected. Simultaneously, unsupervised generation destroys the local structure of a fake image and introduces a large amount of noise. Therefore, some studies have attempted to corrupt color information by using a homogeneous generation method. For example, instead of using a traditional generator, Zhao et al. [89] used a color jitter operation for clothing color transformation, which could learn discriminative color-independent features and mitigate the effect of the generated noise. In addition, inconsistency in the convergence direction of the parameters of the generator and discriminator could lead to difficulties in network optimization. Thus, the Kullback–Leibler divergence constraint [90,91] can be used to replace the discriminator, which can implicitly promote the gradual convergence of the distributions of the generated and real data.

Translation to a New Intermediate Modality

To eliminate the limitations of GANs, such as noise introduction and convergence difficulties, images of different modalities can be transferred to a common modality. Li et al. [92] built an X modality with multimodal information by transferring knowledge from the visible and IR modalities. This self-supervised generation method was more lightweight and efficient than GAN-based models. In addition, Liu et al. [93] generated intermediate modal data by using a homogeneous generation strategy and resolved the problem of the distortion of heterogeneous generation to a certain extent.

Another method generated an intermediate modality by converting visible images into grayscale images. For example, Ye et al. [94] transformed visible data into homogeneous grayscale data to assist cross-modal feature extraction. In this way, intermediate modality generation did not require an additional training process, and the structural information of the visible images could be completely retained. Because the IR domain was not completely consistent with the gray domain, Liu et al. [95] converted both visible and IR data into grayscale data simultaneously, which eliminated the luminance differences among the IR data during conversion.

However, an unconstrained generation process cannot be controlled. If the generated intermediate data are too similar to the original data, modal differences cannot be mitigated. In general, when using only a deep learning method to automatically learn the parameters of the nonlinear layer without the constraints of the underlying mathematical theory, the network often fails to achieve predictability in practical applications, which is a disadvan-

tage of simply improving the network structure. Unlike previous studies that ignored the underlying mathematical theory, Huang et al. [96] did not consider intermediate modality generation as a general step, but treated it as a Markov decision process [97], which allowed the negative impact of random modality mixing to be successfully avoided.

Mathematical Summary

In general, cross-modal person Re-ID works often use GANs containing both a generator (G) and discriminator (D) to generate images that satisfy the training requirements. For example, the data from modality M_1 are used to generate the data from modality M_2 through the generator module, or the data from different modalities are fused to generate intermediate modality data. In GANs, the generator is used to generate realistic images to deceive the discriminator and continuously optimize the network to make the discriminative loss as large as possible. At the same time, it is desired to simultaneously improve the discriminator so that it will not be affected by the generator and continuously optimize the network to make the discriminative loss as small as possible. In the above generative process, the performance of the generator and discriminator can be simultaneously improved, and finally, the cross-modal task can be converted into a single-modal task by the generator. The optimization goal of GANs is as follows:

$$\min_G \max_D V(D,G) = \mathrm{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathrm{E}_{z\sim p_z(z)}[(1 - D(G(z)))] \tag{1}$$

where $p_z(z)$ is the distribution of the input noise variable, and $p_{data}(x)$ is the distribution of the original data. In addition, the discriminator is updated by ascending its stochastic gradient, while the generator is updated by descending its stochastic gradient.

However, it is harder to optimize GANs due to the inconsistent gradient convergence direction of the discriminator and generator. Therefore, some works [91,98] used the Kullback–Leibler (KL) divergence constraint to replace the discriminator module and implicitly promote the gradual convergence of the distributions of the modal data M_1 and modal data M_2 by reducing the scatter value.

$$L_{KL}^{M_1M_2} = E_i \sum_{c=0}^{C-1} [p(c|f_i^{M_1}) \log \frac{p(c|f_i^{M_1})}{p(c|f_i^{M_2})} + p(c|f_i^{M_2}) \log \frac{p(c|f_i^{M_2})}{p(c|f_i^{M_1})}] \tag{2}$$

where C is the number of categories in the classification task. p is the probability distribution of the data from different modes. E is the expectation of different category scatter values. ${f_i}^{M_1}$ and ${f_i}^{M_2}$ are features of a person with identity i, which come from modal data M_1 and M_2, respectively.

Furthermore, directly adding feature maps can also allow the data from different modalities to be fused:

$$x^{mix} = m f^{M_1} + (1 - m) f^{M_2} \tag{3}$$

where m is the mixing ratio, and f^{M_1} and f^{M_2} are feature maps under the modalities M_1 and M_2, respectively. In order to solve the problem of the fusion process being too random, the calculation process of the mixing ratio can be converted into a Markov decision process:

$$\begin{aligned} &\text{State:} && [f^{M_1}; f^{M_2}] \\ &\text{Action:} && m = \text{Sigmoid AvgPool}(\text{Conv}([f^{M_1}; f^{M_2}])) \\ &\text{Reward:} && L_A = -Q(A([f^{M_1}; f^{M_2}]), [f^{M_1}; f^{M_2}]) \end{aligned} \tag{4}$$

where State, Action, and Reward are the basic steps in the Markov decision process; $[f^{M_1}, f^{M_2}]$ is the concatenation of the intermediate feature maps. Meanwhile, the network is updated to obtain a smaller L_A and bigger Q, which implies higher rank-k accuracy and mAP for person Re-ID works. It is worth noting that we want to get a value of m that can produce the smallest L_A.

3.2.2. Shared-Feature-Mapping-Based Methods

To mitigate the modal gap and achieve cross-modal person Re-ID, many researchers have attempted to design an ingenious deep-based network to map pedestrian features in different modalities into a common distribution space. Currently, the most commonly used network frameworks for cross-modal person Re-ID are single-stream networks, dual-stream networks, and GANs, as shown in Figure 3.

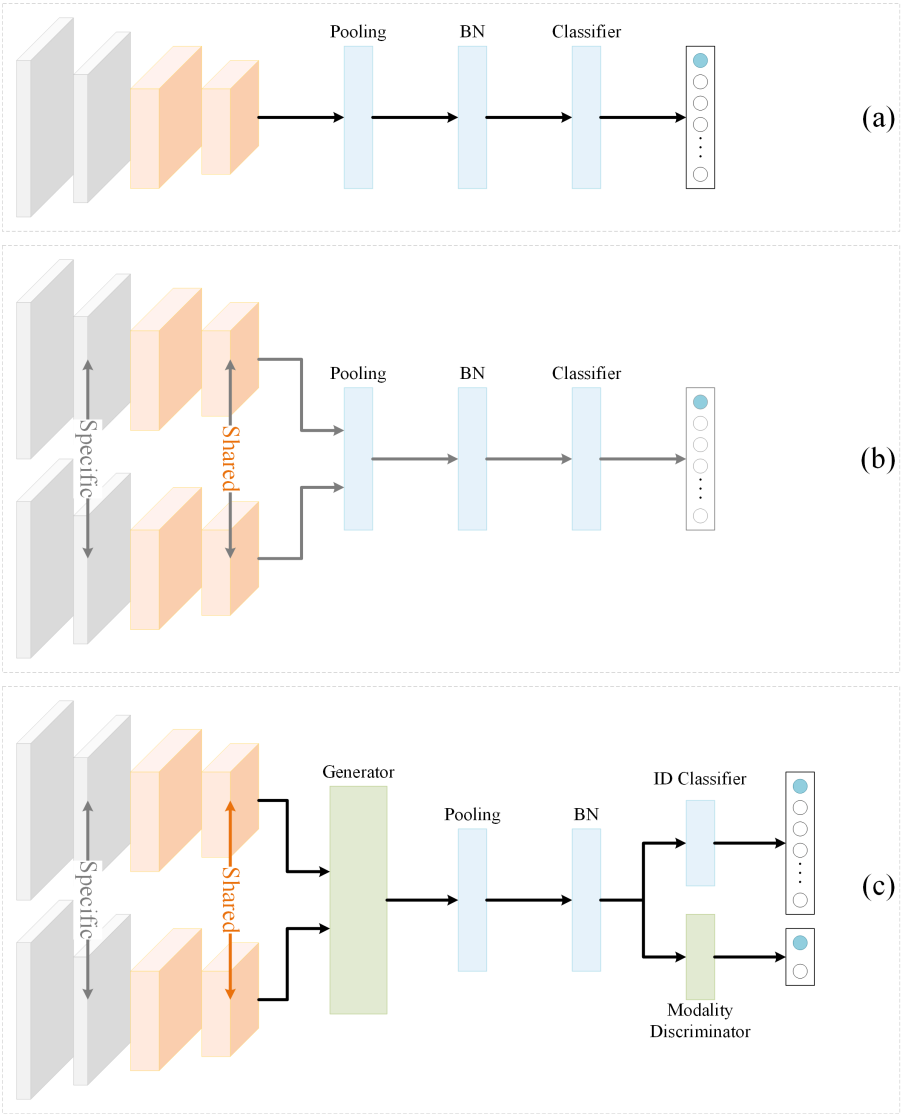

Figure 3. Different networks for cross-modal person Re-ID: (**a**) single-stream networks; (**b**) dual-stream networks; (**c**) GAN-based networks.

Single-Stream Networks

AlexNet [99], VGG [100], GoogleNet [101], and ResNet [102] are representative models of traditional single-stream networks, which have only one input and one output. Unlike in a traditional single-stream network, the network inputs in cross-modal Re-ID [27] are pedestrian images with different modalities. A single-stream network often has two different branches, but the branches share parameters. Pedestrian images with different modalities are mapped to the same feature space after processing by the same network.

Dual-Stream Networks

Owing to the large differences between different modal features, single-stream networks destroy the discriminative features of different modalities. Therefore, some studies

have begun to use dual-stream networks with partially shared parameters to extract features [103–105]. In cross-modal person Re-ID, dual-stream networks always have two branches, with independent parameters in shallower layers and shared parameters in deeper layers. In 2018, Ye et al. [106] first applied dual-stream networks to the field of visible–thermal cross-modal person Re-ID. The proposed end-to-end dual-path feature and metric learning framework not only achieved excellent performance, but also provided a good research basis for future work. However, cross-modal person Re-ID not only needs to address the inter-modal differences, but also needs to deal with the intra-modal perspective differences that exist in traditional person Re-ID. Thus, Ye et al. [107] used hierarchical cross-modal metric learning to constrain the dual-stream network to reduce the impact of view changes on feature extraction. Since then, several cross-modal person Re-ID methods based on this architecture have emerged. These methods extract features from different modal data through the unique convolution layer of each modality and make the feature distribution of different modal outputs gradually consistent under the constraint of a shared convolution layer.

In contrast to the aforementioned learning-based representation methods that focus on designing complex networks to extract ideal features, metric-learning-based methods do not directly treat person Re-ID as a classification problem, but only consider the similarities between pedestrian images. For example, Zhang et al. [108] improved the basic dual-stream network from the perspective of metric learning by mapping features to a hypersphere manifold, thereby eliminating the inconsistency of the norm distributions of different modal features. Ye et al. [109] performed a detailed analysis of triplet loss and proposed a more suitable central triplet loss for metric learning in cross-modal person Re-ID. Unlike in the above studies, which only performed angle transformation in metric learning, Hao et al. [110] noted that it is necessary to strengthen the relationship between the classification subspace and the feature embedding subspace in the spherical space. In their study, a spherical softmax was designed to map the original features, and a two-stage training scheme was used to obtain de-correlated features for classification. Tian et al. [111] found that an information bottleneck, which could preserve useful information and reduce redundant information, achieved good performance in computer vision. However, this method relied heavily on the accurate estimation of mutual information. Therefore, they attempted to provide analytical solutions to fit the mutual information without estimating it and successfully alleviated the above difficulties. Similarly, Meshky et al. [112] noticed the important role of mutual information in mitigating modal effects and extracted intermediate features to retain more mutual information. Furthermore, the entropy value was used to weigh the features of different layers by analyzing the information reflected by it (i.e., the higher the entropy is, the lower discrimination of the features is).

GAN-Based Networks

Inspired by GANs, Dai et al. [113] designed an adversarial network to discriminate the originating modalities of the output features and achieved further alignment of different modal feature distributions to maximize the discriminative loss. Since then, an increasing number of studies have begun to use adversarial ideas for feature-level alignment rather than image-level style conversion [114]. In contrast to previous studies that used a simple discriminator for the network, Hu et al. [115] set up an exclusive identity classifier and modality discriminator for modality-related and identity-related branches separately to disentangle modality-related features. Moreover, to mitigate the randomness of a deep-learning-based network, this method set the representational orthogonal decorrelation at the feature and parameter levels and enforced the process of feature extraction more reliably by using mathematical constraints. To simultaneously address the modal and domain gaps, Fu et al. [116] attempted to reduce the domain gap by using an adversarial method. They calculated the moments for the features in the target domain and achieved cross-modal distribution alignment by closing the moments of the same ID.

Mathematical Summary

Theoretically, a Siamese network can map features of different modalities into the same distribution space, but in practical applications, this capability is usually limited. To obtain features with more consistent distributions, researchers often set up auxiliary networks based on mathematical theories to process the features of different modalities so that they can force them to have similar distributions. For example, features can be further mapped to the corner space [108] in which modal differences are smaller:

$$\begin{aligned}\Theta(f^*;W^*) &= f^*W^* = [f^*w_1^*, f^*w_2^*, \cdots, f^*w_c^*] \\ &= [\|f^*\| \times \|w_1^*\| \times \cos\theta_1, \cdots, \|f^*\| \times \|w_c^*\| \times \cos\theta_c] \\ &= [\cos\theta_1, \cos\theta_2, \cdots, \cos\theta_c]\end{aligned} \tag{5}$$

where the angle transformation matrix W^* projects feature f^* from the feature space to the angular space, and c represents the number of classes in the training data. $W^* = [w_1^*, w_2^*, \cdots, w_c^*] = [\frac{w_1}{\|w_1\|_2}, \frac{w_2}{\|w_2\|_2}, \cdots, \frac{w_c}{\|w_c\|_2}]$, $f^* = \frac{f}{\|f\|_2}$.

In addition to transformation at the coordinate level, it is also a feasible method to calculate entropy or moment or to orthogonalize the features at the feature level. For example, by reducing the differences in moments, the consistency of samples that have the same IDs and different modalities can be strengthened [116]. In this way, the modality gap can be implicitly alleviated:

$$\begin{aligned}\sigma_c^{M_1} &= \frac{1}{N_c^{M_1}} \sum_{i=1}^{N_c^{M_1}} \left(f_i^{M_1} - \mu^{M_1}\right)^2 \\ \sigma_c^{M_2} &= \frac{1}{N_c^{M_2}} \sum_{i=1}^{N_c^{M_2}} \left(f_i^{M_2} - \mu^{M_2}\right)^2 \\ d_c^{M_1M_2} &= D(\sigma_c^{M_1}, \sigma_c^{M_2})\end{aligned} \tag{6}$$

where $N_c^{M_1}$ and $N_c^{M_2}$ represent the numbers of samples of the c-th identity in modality M_1 and modality M_2, respectively. c denotes the identity. $\sigma_c^{M_1}$ and $\sigma_c^{M_2}$ denote the second-order moments (variance). μ^{M_1} and μ^{M_2} represent the prototype (the mean within one class). D denotes the distance function.

For a certain class of features, the moment can reflect the stability of the features in terms of distribution, while the entropy can reflect the accuracy of the features in terms of discriminability. Therefore, we can give a weight w to different features according to their entropy. Thus, the modality-irrelevant features can be enhanced and the modality-relevant features can be weakened:

$$\begin{aligned}H(f) &= -\sum_{j=1}^{C} P(j|f) \log[P(j|f)] \\ w &= \frac{1 + e^{-H(f)}}{\sum_{k=1}^{M} 1 + e^{-H(f^k)}}\end{aligned} \tag{7}$$

where $H(f)$ denotes the entropy in the class distribution for the input images with feature f. M denotes the batch size. C is the total number of identities. $P(j|f)$ represents the probability of identity j given descriptor f.

In order to further separate the modality-irrelevant features f_c^I from the modality-relevant features f_c^R, they can be orthogonalized:

$$(f_c^I)^T f_c^R = 0 \tag{8}$$

where c denotes the identity.

Obviously, all of the above operations are done for the features themselves. Otherwise, a more reasonable metric learning method can be designed for the features. For example, a network can be constrained by a more effective loss function, such as the bi-directional center-constrained top-ranking loss:

$$\begin{aligned} L_{bicenter} = & \sum_{c=1}^{K} \max[\rho_1 + D(f_c^{M_1}, \mu_c) - \min_{\forall c \neq k} D(f_c^{M_1}, \mu_k), 0] \\ & + \sum_{c=1}^{K} \max[\rho_1 + D(f_c^{M_2}, \mu_c) - \min_{\forall c \neq k} D(f_c^{M_2}, \mu_k), 0] \end{aligned} \tag{9}$$

where K is the batch size. μ_c and μ_k are d-dimensional vectors representing the center of class c. f^{M_1} and f^{M_2} are features of modalities M_1 and M_1, respectively.

3.2.3. Auxiliary-Information-Based Methods

Visual features change significantly due to modal influence. It is impossible to achieve the desired accuracy by relying only on a small amount of modality-invariant discriminative visual information for the identity metric. Therefore, some studies have attempted to explore higher-order relations and temporal information, which are robust to modality changes, to assist identification. Thus, the role of visual information is weakened, and the effect of the modality is mitigated.

Higher-Order Relationship Assistance

Yin et al. [117] built an affinity modeling module to further mine the relationships between the global features of different pedestrians and enhance the ability of network cross-modality matching. However, this method focused more on the relationships between different pedestrians, which essentially enhanced the compactness in the same category and the separability between different categories, making it difficult to constrain the network to the extraction of structural features unrelated to appearance features. To address this issue, Huang et al. [118] used 3DConv to extract discriminative modality-invariant relationships between different parts of the body. With the complement of the relationship features between different parts of the body, the ability of the network to discriminate difficult samples with similar appearances but obvious differences in body shape was further improved. In addition to 3DConv, a transformer is also a common tool for extracting structural information of the human body [119]. To mine more accurate structural information, Gao et al. [120] used key points to assist in feature extraction. This method constructed a graph structure based on the actual structure of the human body and mined the intra-modality and inter-modality structural information of the human body by using cross-modal graph matching.

Temporal Information Assistance

Video-based methods are more advantageous than image-based cross-modal person Re-ID methods. This is because more comprehensive appearance features and modality-independent temporal features can be extracted from video, which is conducive to solving the network performance degradation caused by the modal gap. To fully extract the temporal information in videos, researchers have attempted to use optical flow [121,122], recurrent neural networks [123,124], temporal pooling [44,125–127], and 3DConv [128–130] to aggregate sequence-level features. For example, Aich et al. [131] used 3DConv with different pooling sizes to exploit temporal information. Hou et al. [132] designed a temporal kernel selection module to adaptively capture short- and long-term temporal relationships. Although the current temporal information extraction methods are relatively mature, few cross-modal person Re-ID methods have used modality-invariant temporal information. MITML [50] is the only method that combines temporal information extraction with cross-modal person Re-ID.

Mathematical Summary

Since visual features are highly susceptible to modal changes, some methods have tried to explore the relationships (e.g., the relationships between different features in a spatial distribution, the relationships between different parts of the same person, the movement rules of individuals, etc.) that are robust to modality changes to assist identity discrimination. Usually, the relationship between features f_i and f_j can be better represented by the similarity matrix A, and the value of A at (i, j) can be calculated with the following equation:

$$A_{ij} = \exp^{-dist(f_i, f_j)} \tag{10}$$

where *dist* denotes the distance measure function.

Unlike in some works that independently performed the calculation of the relationship matrix before the feature output, a graph convolutional network (GCN) integrates the relationship calculation into the process of feature extraction. Each node of a GCN can not only extract local features as ordinary convolutional networks do, but can also update its own state while receiving messages from neighboring nodes so that each local node can contain more comprehensive global information. Therefore, graph convolution is very suitable for extracting the relationships between local blocks of the human body and outputting the high-order relationship characteristics of the human body. Graph convolution can be expressed as:

$$H^{l+1} = \sigma(AH^l W^l) \tag{11}$$

where H^l and H^{l+1} are features of layer l and layer $l+1$, respectively. A is an adjacency matrix, which is used to identify whether certain edges in the graph exist (for example, A_{ij} represents the edge between node i and node j, and if the edge exists, $A_{ij} = 1$; otherwise, $A_{ij} = 0$). W_l is a weight matrix for layer l, which can represent the relationship between nodes. σ is a nonlinear activation function.

Similarly to graph convolutional networks, recurrent neural networks (RNNs) can selectively pass information from the previous layer to the next layer, so they can also act as information aggregators. RNNs can be represented as:

$$\begin{aligned} h_t &= f(Uh_{t-1} + Wx_t + b) \\ y &= Vh_t \end{aligned} \tag{12}$$

where h is the hidden state, $f()$ is the nonlinear activation function, and U, W, V and b are the network parameters.

3.3. Loss Functions

In the training stage, the quality of the features is evaluated at the end of the network. In order to classify output features to an exact category, a clear demarcation between different identities is expected. At the same time, it is expected that the distance between the same identity features is closer than that between different identity features. To achieve this goal, suitable loss functions are needed to optimize the network, as shown in Figure 4. According to the types of loss functions, optimization methods can be divided into metric learning, which is represented by triplet loss (L_{Tri}), and representation learning, which is represented by identity loss (L_{ID}).

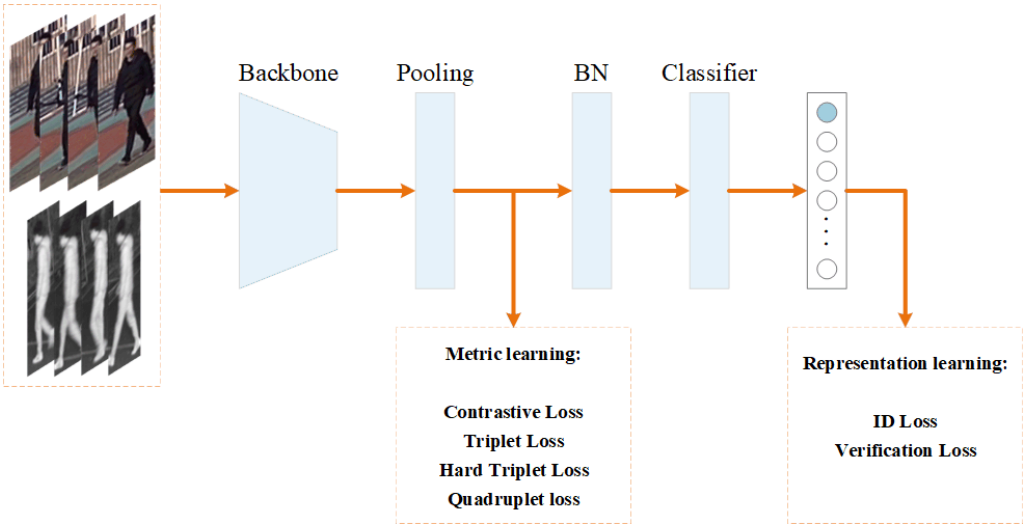

Figure 4. Loss function settings.

Moreover, it is worth noting that deep-learning-based methods usually update the network parameters through gradient descent:

$$\theta_{t+1} = \theta_t - \alpha \frac{1}{N} \sum_{n=1}^{N} \frac{\partial L\left(y^{(n)}, f\left(x^{(n)}; \theta\right)\right)}{\partial \theta} \tag{13}$$

where θ denotes the network's parameters. t is the number of optimizations. α is the learning rate. x denotes the training samples. y denotes the label. L is the loss function.

3.3.1. Distance Metric

In order to clearly introduce metric learning and representation learning, we briefly introduce their mathematical models. To measure the distance of features, Re-ID work often uses the Euclidean, Manhattan, Chebyshev, and cosine distance, as shown in Figure 5.

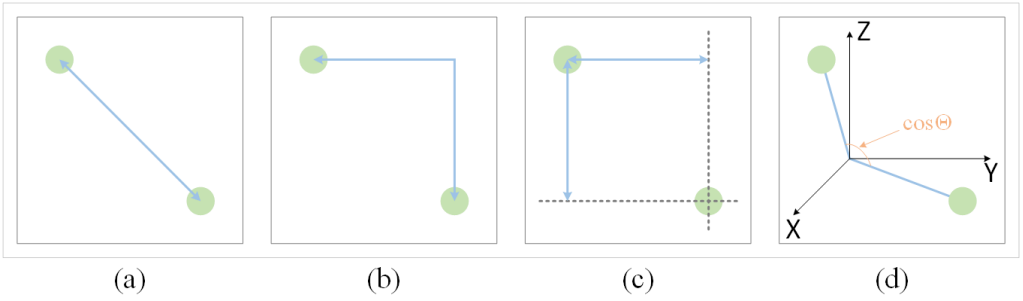

Figure 5. Distance metric: (**a**) Euclidean distance; (**b**) Manhattan distance; (**c**) Chebyshev distance; (**d**) cosine distance.

Euclidean Distance (L2 Distance)

The Euclidean distance [133] represents the true distance between two points in n-dimensional space, which can be regarded as the linear distance:

$$D_e(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{14}$$

where $X = [x_1, x_2, \cdots, x_n]$ and $Y = [y_1, y_2, \cdots, y_n]$.

Manhattan Distance (L1 Distance)

The Manhattan distance [134] represents the sum of the axial distances between two points X and Y in each dimension, which can be regarded as the rectilinear distance:

$$D_m(X,Y) = \sum_{i=1}^{n} |x_i - y_i| \tag{15}$$

Chebyshev Distance

The Chebyshev distance [135] represents the maximum of the distance in different axes between two points. The Euclidean distance [133] represents the true distance between two points in the n-th dimension, which can be regarded as the linear distance:

$$\begin{aligned} D_c(X,Y) &= \max(|x_1 - y_1|, |x_2 - y_2|, \cdots, |x_n - y_n|) \\ &= \lim_{p\to\infty} \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \end{aligned} \tag{16}$$

Obviously, the Manhattan, Euclidean, and Chebyshev distances can be expressed by a unified formula, i.e., the Minkoff distance formula:

$$D = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{17}$$

where $p = 1$ is the Manhattan distance, $p = 2$ is the Euclidean distance, and $p = \infty$ is the Chebyshev distance.

Cosine Distance

Cosine similarity [136] measures the differences between individuals by calculating the cosine of the angle between two vectors. Compared to the above methods, the cosine distance focuses more on the difference between two vectors in terms of direction rather than distance or length:

$$\begin{aligned} \cos(\theta) &= \frac{X \cdot Y}{\|X\|\|Y\|} \\ &= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} {x_i}^2} \sqrt{\sum_{i=1}^{n} {y_i}^2}} \end{aligned} \tag{18}$$

3.3.2. Metric Learning

The purpose of metric learning [137] is to find an appropriate metric that can make the distance of different identities larger than that of the same identity. In this way, features of each identity can cluster in metric space. In person Re-ID, Euclidean distance and its variants are often used to construct metric learning loss, which can promote feature clustering. The commonly used metric learning losses are as follows.

Contrastive Loss

In contrastive loss [138], if the inputs X_i and X_j are a pair of samples of the same identity, $\delta_{ij} = 1$; conversely, $\delta_{ij} = 0$. Therefore, the value of the loss will increase when the distance of samples that come from the same identity is larger or the distance of features that come from different identities is smaller:

$$L_{con} = (1 - \delta_{ij}) \{ \max(0, \rho - d_{ij}) \}^2 + \delta_{ij} {d_{ij}}^2 \tag{19}$$

where d_{ij} denotes the distance of samples X_i and X_j. ρ is the threshold value to be set.

Contrastive loss requires a fixed margin, so a very strong assumption that the distribution of each class of samples is the same is implied here. However, this strong assumption is difficult to hold in general. At the same time, it is difficult to ensure that the distance of features from the same ID is smaller than that of those from different IDs if only the distance between a pair of samples is considered.

Triplet Loss

Unlike contrastive loss, the triplet loss [139] usually contains an anchor sample a, a positive sample p with the same identity as that of a, and a negative sample n with a different identity from that of a. Obviously, the value of the triplet loss will descend when the anchor is closer to the positive sample and distant from the negative sample.

$$L_{Tri} = \max(0, \rho + d_{ap} - d_{an}) \tag{20}$$

where ρ is the threshold. d_{ap} denotes the distance between the anchor sample and positive sample, and d_{ap} denotes the distance between the anchor sample and the negative sample. However, due to the characteristics of metric learning itself, triplet loss is easily overfitted to the current samples, which makes it difficult for the network to mine discriminative features from unseen samples.

Hard-Sample-Based Triplet Loss

To improve the generalization of the network and force the network to extract more discriminative features rather than overfit to simple samples, hard-sample-based triplet loss is calculated with a hard negative sample and hard positive sample in a batch instead of all samples:

$$L_{H-Tri} = \max(0, \rho + \max d_{ap} - \min d_{an}) \tag{21}$$

However, not only hard-sample-based triplet loss, but also traditional triplet loss cannot get rid of calculating the loss based entirely on a relative distance. Therefore, it is impossible to avoid both d_{ap} and d_{an} decreasing or increasing at the same time during optimization, which will have a negative impact on the network performance.

Quadruplet Loss

Compared with triplet loss, quadruplet loss [140] adds another negative sample to calculate the distance between two negative samples:

$$L_Q = \max(\rho_1 + d_{ap} - d_{an_1}, 0) + \max(\rho_2 + d_{ap} - d_{n_1 n_2}, 0) \tag{22}$$

Obviously, in the second term, there is no shared identity between d_{ap} and $d_{n_1 n_2}$. Because the quadruplet loss uses the absolute distance between positive and negative samples, it can solve the problem of simultaneously increasing or decreasing.

3.3.3. Representation Learning

Representation learning [141] does not directly compare the similarities between different features as metric learning does, but treats the cross-modal person Re-ID task as a classification problem (multi-classification) or verification problem (binary classification). At the same time, for representation learning, the network needs to connect the feature output layer with a fully connected (FC) layer to make the number of output feature dimensions equal to the number of categories through the FC layer.

Classification loss is used to judge the accuracy of classification results. The higher the accuracy is, the smaller the classification loss is. On the contrary, the classification

loss is greater. Cross-entropy loss is usually used as the classification loss for cross-modal person Re-ID.

$$L_{ID} = -\sum_{i=1}^{N} q_i \log(p(f_i)) \tag{23}$$

where $N$is the number of samples. q_i represents the one-hot vector of the i-th sample. $p(f_i)$ represents the prediction probability of f_i.

Obviously, due to the FC layers, when there are too many parameters in the network, the computation of representation learning will be complex, and it is difficult to obtain a converged network. However, representation learning will have a more stable result compared to that of metric learning because it takes all categories into account.

4. Conclusions

We provided a comprehensive and systematic analysis of cross-modal person Re-ID works and classified them according to the means of alleviating the modality gap. These methods were classified into picture-level style transform-based, shared-embedding-mapping-based, and modality-invariant auxiliary-feature-based methods. By generalizing and comparing different methods, we found that deep-learning-based networks with mathematical constraints tend to show more robust performance and stronger interpretation. Through the above analysis, we have summarized the current problems and future developmental directions of cross-modal person Re-ID as follows:

- Relying exclusively on off-the-shelf networks leads to researchers' lack of understanding of the computational processes. Thus, most researchers are unfamiliar with the underlying logic of the networks and are unable to judge whether the overall framework of a network is reasonable. Therefore, an understanding of the mathematical principles related to deep-learning-based networks is required, and the characteristics of the network should be considered.
- In the case of large modal differences, it is difficult to obtain accurate retrieval results based only on a simple deep-learning-based network. Therefore, additional theoretical constraints are required to improve the performance. The existing cross-modal person Re-ID methods primarily rely on appearance features with great modal differences to determine identity. Therefore, more detailed analysis and processing of data are needed to extract more discriminative features in the future.
- When the training data are insufficient, deep-learning-based networks tend to overfit to the visible modality, producing ineffective performance in realistic scenarios. Fortunately, handcrafted feature-extraction-based methods are not troubled by this problem. Therefore, combining the mathematical tools in traditional methods and frameworks in deep-learning-based methods will be very important in the future.
- Although the technology of person re-identification has been rapidly developed, downstream tasks based on person re-identification have still not received attention. For example, in medical health, the behavior of patients with the same identity can be identified and analyzed with different cameras according to the person re-identification technology. The pedestrian behavior of the same patient with a mental illness can be monitored with different cameras on the basis of this technology.

Author Contributions: Investigation, M.L.; data curation, M.L.; writing—original draft preparation, M.L. and Y.Z.; writing—review and editing, Y.Z. and H.L.; supervision, H.L.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Nos. 62276120 and 61966021).

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

References

1. Coifman, B.; Cassidy, M. Vehicle reidentification and travel time measurement on congested freeways. *Transp. Res. Part A Policy Pract.* **2002**, *36*, 899–917. [CrossRef]
2. An, L.; Bhanu, B.; Yang, S. Face recognition in multi-camera surveillance videos. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 2885–2888.
3. Yan, S.; Zhang, Y.; Xie, M.; Zhang, D.; Yu, Z. Cross-domain person re-identification with pose-invariant feature decomposition and hypergraph structure alignment. *Neurocomputing* **2022**, *467*, 229–241. [CrossRef]
4. Li, S.; Li, F.; Wang, K.; Qi, G.; Li, H. Mutual prediction learning and mixed viewpoints for unsupervised-domain adaptation person re-identification on blockchain. *Simul. Model. Pract. Theory* **2022**, *119*, 102568. [CrossRef]
5. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 618–626.
6. Yu, X.; Chen, T.; Yang, Y.; Mugo, M.; Wang, Z. Cross-modal person search: A coarse-to-fine framework using bi-directional text-image matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27 October–2 November 2019.
7. Zhang, Y.; Wang, Y.; Li, H.; Li, S. Cross-Compatible Embedding and Semantic Consistent Feature Construction for Sketch Re-identification. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 3347–3355.
8. Hafner, F.M.; Bhuyian, A.; Kooij, J.F.; Granger, E. Cross-modal distillation for RGB-depth person re-identification. *Comput. Vis. Image Underst.* **2022**, *216*, 103352. [CrossRef]
9. Li, H.; Dong, N.; Yu, Z.; Tao, D.; Qi, G. Triple Adversarial Learning and Multi-View Imaginative Reasoning for Unsupervised Domain Adaptation Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 2814–2830. [CrossRef]
10. Li, H.; Chen, Y.; Tao, D.; Yu, Z.; Qi, G. Attribute-Aligned Domain-Invariant Feature Learning for Unsupervised Domain Adaptation Person Re-Identification. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 1480–1494. [CrossRef]
11. Li, H.; Xu, J.; Yu, Z.; Luo, J. Jointly Learning Commonality and Specificity Dictionaries for Person Re-Identification. *IEEE Trans. Image Process.* **2020**, *29*, 7345–7358. [CrossRef]
12. 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, Montreal, QC, Canada, 10–17 October 2021; pp. 16403–16412.
13. Li, H.; Xu, K.; Li, J.; Yu, Z. Dual-stream Reciprocal Disentanglement Learning for domain adaptation person re-identification. *Knowl.-Based Syst.* **2022**, *251*, 109315. [CrossRef]
14. Gheissari, N.; Sebastian, T.B.; Hartley, R. Person reidentification using spatiotemporal appearance. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 1528–1535.
15. Chen, C.H.; Chen, T.Y.; Lin, J.C.; Wang, D.J. People Tracking in the Multi-camera Surveillance System. In Proceedings of the 2011 Second International Conference on Innovations in Bio-Inspired Computing and Applications, Shenzhen, China, 16–18 December 2011; Number 47, pp. 1–4. [CrossRef]
16. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision—ECCV 2016 Workshops, Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–10 and 15–16 October 2016*; Springer: Cham, Switzerland, 2016; pp. 17–35.
17. Krumm, J.; Harris, S.; Meyers, B.; Brumitt, B.; Hale, M.; Shafer, S. Multi-camera multi-person tracking for easyliving. In Proceedings of the Third IEEE International Workshop on Visual Surveillance, Dublin, Ireland, 1 July 2000; pp. 3–10.
18. Ciobanu, A.; Luca, M.; Păvăloi, I.; Barbu, T. *Iris Identification Based on Optimized Lab Histograms Applied to Iris Partitions*. Buletinul Institutului Politehnic, Iaşi, Tomul LX (LXIV), Fasc. 1, 2014.
19. Liao, S.; Law, M.W.; Chung, A.C. Dominant local binary patterns for texture classification. *IEEE Trans. Image Process.* **2009**, *18*, 1107–1118. [CrossRef]
20. Forssén, P.E. Maximally stable colour regions for recognition and matching. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
21. Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; Cristani, M. Person re-identification by symmetry-driven accumulation of local features. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2360–2367.
22. Nosaka, R.; Ohkawa, Y.; Fukui, K. Feature extraction based on co-occurrence of adjacent local binary patterns. In *Proceedings of the Pacific-Rim Symposium on Image and Video Technology*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 82–91.
23. Li, W.; Zhao, R.; Xiao, T.; Wang, X. Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 152–159.
24. Antipov, G.; Berrani, S.A.; Ruchaud, N.; Dugelay, J.L. Learned vs. hand-crafted features for pedestrian gender recognition. In Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia, 26–30 October 2015; pp. 1263–1266.
25. Dara, S.; Tumma, P. Feature extraction by using deep learning: A survey. In Proceedings of the 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 29–31 March 2018; pp. 1795–1801.

26. Liang, H.; Sun, X.; Sun, Y.; Gao, Y. Text feature extraction based on deep learning: A review. *EURASIP J. Wirel. Commun. Netw.* **2017**, *2017*, 1–12. [CrossRef]
27. Wu, A.; Zheng, W.S.; Yu, H.X.; 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. 5380–5389.
28. Wu, D.; Zheng, S.J.; Zhang, X.P.; Yuan, C.A.; Cheng, F.; Zhao, Y.; Lin, Y.J.; Zhao, Z.Q.; Jiang, Y.L.; Huang, D.S. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing* **2019**, *337*, 354–371. [CrossRef]
29. Almasawa, M.O.; Elrefaei, L.A.; Moria, K. A Survey on Deep Learning-Based Person Re-Identification Systems. *IEEE Access* **2019**, *7*, 175228–175247. [CrossRef]
30. Leng, Q.; Ye, M.; Tian, Q. A Survey of Open-World Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 1092–1108. [CrossRef]
31. Wang, H.; Du, H.; Zhao, Y.; Yan, J. A Comprehensive Overview of Person Re-Identification Approaches. *IEEE Access* **2020**, *8*, 45556–45583. [CrossRef]
32. Mathur, N.; Mathur, S.; Mathur, D.; Dadheech, P. A Brief Survey of Deep Learning Techniques for Person Re-identification. In Proceedings of the 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), Jaipur, India, 7–8 February 2020; Number 10, pp. 129–138. [CrossRef]
33. 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]
34. Gray, D.; Tao, H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In Proceedings of the European Conference on Computer Vision, Marseille France, 12–18 October 2008; Springer: Berlin/Heidelberg, Germany, 2008; Number 100, pp. 262–275.
35. Zheng, W.S.; Gong, S.; Xiang, T. Associating groups of people. In Proceedings of the BMVC, London, UK, 7–10 September 2009; Volume 2, pp. 1–11.
36. Loy, C.C.; Liu, C.; Gong, S. Person re-identification by manifold ranking. In Proceedings of the 2013 IEEE International Conference on Image Processing, Melbourne, Australia, 15–18 September 2013; Number 102, pp. 3567–3571.
37. Hirzer, M.; Beleznai, C.; Roth, P.M.; Bischof, H. Person re-identification by descriptive and discriminative classification. In *Image Analysis, Proceedings of the Scandinavian Conference on Image Analysis, Ystad, Sweden, 9 May 2011*; Springer: Berlin/Heidelberg, Germany, 2011; Number 103, pp. 91–102.
38. 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, Santiago, Chile, 7–13 December 2015; Number 106, pp. 1116–1124.
39. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; Number 16, pp. 3754–3762.
40. Karanam, S.; Gou, M.; Wu, Z.; Rates-Borras, A.; Camps, O.; Radke, R.J. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 523–536.
41. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; Number 110, pp. 79–88.
42. Wang, T.; Gong, S.; Zhu, X.; Wang, S. Person re-identification by video ranking. In *Computer Vision—ECCV 2014, Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014*; Springer: Cham, Switzerland, 2014; Number 105, pp. 688–703.
43. Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q. Mars: A video benchmark for large-scale person re-identification. In *Computer Vision—ECCV 2016, Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Cham, Switzerland, 2016; Number 107, pp. 868–884.
44. 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 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5177–5186.
45. Li, M.; Zhu, X.; Gong, S. Unsupervised person re-identification by deep learning tracklet association. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Number 112, pp. 737–753.
46. Song, G.; Leng, B.; Liu, Y.; Hetang, C.; Cai, S. Region-based quality estimation network for large-scale person re-identification. *AAAI Conf. Artif. Intell.* **2018**, *32*. [CrossRef]
47. Li, J.; Wang, J.; Tian, Q.; Gao, W.; Zhang, S. Global-local temporal representations for video person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; Number 114, pp. 3958–3967.
48. Hou, R.; Chang, H.; Ma, B.; Shan, S.; Chen, X. Temporal Complementary Learning for Video Person Re-identification. In *Computer Vision—ECCV 2020, Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Number 14, pp. 388–405.
49. 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]

50. Lin, X.; Li, J.; Ma, Z.; Li, H.; Li, S.; Xu, K.; Lu, G.; Zhang, D. Learning Modal-Invariant and Temporal-Memory for Video-Based Visible-Infrared Person Re-Identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20973–20982.
51. Barbosa, I.B.; Cristani, M.; Bue, A.D.; Bazzani, L.; Murino, V. Re-identification with rgb-d sensors. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2012; pp. 433–442.
52. Munaro, M.; Basso, A.; Fossati, A.; Van Gool, L.; Menegatti, E. 3D reconstruction of freely moving persons for re-identification with a depth sensor. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 4512–4519.
53. Munaro, M.; Ghidoni, S.; Dizmen, D.T.; Menegatti, E. A feature-based approach to people re-identification using skeleton keypoints. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 5644–5651.
54. Haque, A.; Alahi, A.; Fei-Fei, L. Recurrent attention models for depth-based person identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1229–1238.
55. Li, S.; Xiao, T.; Li, H.; Zhou, B.; Yue, D.; Wang, X. Person search with natural language description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1970–1979.
56. Pang, L.; Wang, Y.; Song, Y.Z.; Huang, T.; Tian, Y. Cross-domain adversarial feature learning for sketch re-identification. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 609–617.
57. Thys, S.; Van Ranst, W.; Goedemé, T. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
58. Zhao, Z.Q.; Zheng, P.; Xu, S.t.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef]
59. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
60. Humeau-Heurtier, A. Texture feature extraction methods: A survey. *IEEE Access* **2019**, *7*, 8975–9000. [CrossRef]
61. Latif, A.; Rasheed, A.; Sajid, U.; Ahmed, J.; Ali, N.; Ratyal, N.I.; Zafar, B.; Dar, S.H.; Sajid, M.; Khalil, T. Content-based image retrieval and feature extraction: A comprehensive review. *Math. Probl. Eng.* **2019**, *2019*. [CrossRef]
62. Salau, A.O.; Jain, S. Feature extraction: A survey of the types, techniques, applications. In Proceedings of the 2019 International Conference on Signal Processing and Communication (ICSC), Noida, India, 7–9 March 2019; pp. 158–164.
63. Kaya, M.; Bilge, H.Ş. Deep metric learning: A survey. *Symmetry* **2019**, *11*, 1066. [CrossRef]
64. Musgrave, K.; Belongie, S.; Lim, S.N. A metric learning reality check. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 681–699.
65. Ge, W. Deep metric learning with hierarchical triplet loss. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 269–285.
66. Ren, L.; Lu, J.; Feng, J.; Zhou, J. Multi-modal uniform deep learning for RGB-D person re-identification. *Pattern Recognit.* **2017**, *72*, 446–457. [CrossRef]
67. Imani, Z.; Soltanizadeh, H.; Orouji, A.A. Short-term person re-identification using rgb, depth and skeleton information of rgb-d sensors. *Iran. J. Sci. Technol. Trans. Electr. Eng.* **2020**, *44*, 669–681. [CrossRef]
68. Wu, A.; Zheng, W.S.; Lai, J.H. Robust depth-based person re-identification. *IEEE Trans. Image Process.* **2017**, *26*, 2588–2603. [CrossRef] [PubMed]
69. Xu, R.; Shen, F.; Wu, H.; Zhu, J.; Zeng, H. Dual Modal Meta Metric Learning for Attribute-Image Person Re-identification. In Proceedings of the 2021 IEEE International Conference on Networking, Sensing and Control (ICNSC), Xiamen, China, 3–5 December 2021; Volume 1, pp. 1–6.
70. Ding, Z.; Ding, C.; Shao, Z.; Tao, D. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv* **2021**, arXiv:2107.12666.
71. Wang, Z.; Liu, X.; Li, H.; Sheng, L.; Yan, J.; Wang, X.; Shao, J. Camp: Cross-modal adaptive message passing for text-image retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5764–5773.
72. Niu, K.; Huang, Y.; Ouyang, W.; Wang, L. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Trans. Image Process.* **2020**, *29*, 5542–5556. [CrossRef]
73. Kansal, K.; Subramanyam, A.; Wang, Z.; Satoh, S. Hierarchical Attention Image-Text Alignment Network For Person Re-Identification. In Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shenzhen, China, 5–9 July 2021; pp. 1–6.
74. Farooq, A.; Awais, M.; Kittler, J.; Akbari, A.; Khalid, S.S. Cross modal person re-identification with visual-textual queries. In Proceedings of the 2020 IEEE International Joint Conference on Biometrics (IJCB), Houston, TX, USA, 28 September–1 October 2020; pp. 1–8.
75. Zha, Z.J.; Liu, J.; Chen, D.; Wu, F. Adversarial attribute-text embedding for person search with natural language query. *IEEE Trans. Multimed.* **2020**, *22*, 1836–1846. [CrossRef]

76. Zhao, S.; Gao, C.; Shao, Y.; Zheng, W.S.; Sang, N. Weakly Supervised Text-based Person Re-Identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 11395–11404.
77. Gui, S.; Zhu, Y.; Qin, X.; Ling, X. Learning multi-level domain invariant features for sketch re-identification. *Neurocomputing* **2020**, *403*, 294–303. [CrossRef]
78. Yang, F.; Wu, Y.; Wang, Z.; Li, X.; Sakti, S.; Nakamura, S. Instance-level heterogeneous domain adaptation for limited-labeled sketch-to-photo retrieval. *IEEE Trans. Multimed.* **2020**, *23*, 2347–2360. [CrossRef]
79. 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**, *24*, 1570–1582. [CrossRef]
80. Zhou, H.; Huang, C.; Cheng, H. A relation network design for visible thermal person re-identification. In Proceedings of the 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 9–11 April 2021; pp. 511–515.
81. Cheng, D.; Li, X.; Qi, M.; Liu, X.; Chen, C.; Niu, D. Exploring cross-modality commonalities via dual-stream multi-branch network for infrared-visible person re-identification. *IEEE Access* **2020**, *8*, 12824–12834. [CrossRef]
82. Zhang, C.; Liu, H.; Guo, W.; Ye, M. Multi-scale cascading network with compact feature learning for RGB-infrared person re-identification. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 8679–8686.
83. Wang, G.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Yu, G.; Zhou, E.; Sun, J. High-order information matters: Learning relation and topology for occluded person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6449–6458.
84. Shen, Y.; Li, H.; Yi, S.; Chen, D.; Wang, X. Person re-identification with deep similarity-guided graph neural network. In Proceedings of the European conference on computer vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 486–504.
85. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
86. Wang, G.A.; Zhang, T.; Yang, Y.; Cheng, J.; Chang, J.; Liang, X.; Hou, Z.G. Cross-modality paired-images generation for RGB-infrared person re-identification. *AAAI Conf. Artif. Intell.* **2020**, *34*, 12144–12151. [CrossRef]
87. 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, Republic of Korea, 27 October–2 November 2019; Number 18, pp. 3623–3632.
88. Zhong, X.; Lu, T.; Huang, W.; Ye, M.; Jia, X.; Lin, C.W. Grayscale enhancement colorization network for visible-infrared person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1418–1430. [CrossRef]
89. 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. *AAAI Conf. Artif. Intell.* **2021**, *35*, 3520–3528. [CrossRef]
90. Van Erven, T.; Harremos, P. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820. [CrossRef]
91. Joyce, J.M. Kullback-leibler divergence. In *International Encyclopedia of Statistical Science*; Springer: Cham, Switzerland, 2011; pp. 720–722.
92. Li, D.; Wei, X.; Hong, X.; Gong, Y. Infrared-visible cross-modal person re-identification with an x modality. *AAAI Conf. Artif. Intell.* **2020**, *34*, 4610–4617. [CrossRef]
93. Liu, H.; Miao, Z.; Yang, B.; Ding, R. A base-derivative framework for cross-modality RGB-infrared person re-identification. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; Number 22, pp. 7640–7646.
94. 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. [CrossRef]
95. Liu, H.; Xia, D.; Jiang, W.; Xu, C. Towards Homogeneous Modality Learning and Multi-Granularity Information Exploration for Visible-Infrared Person Re-Identification. *arXiv* **2022**, arXiv:2204.04842.
96. Huang, Z.; Liu, J.; Li, L.; Zheng, K.; Zha, Z.J. Modality-Adaptive Mixup and Invariant Decomposition for RGB-Infrared Person Re-Identification. *arXiv* **2022**, arXiv:2203.01735.
97. Puterman, M.L. Markov decision processes. *Handb. Oper. Res. Manag. Sci.* **1990**, *2*, 331–434.
98. Hershey, J.R.; Olsen, P.A. Approximating the Kullback Leibler divergence between Gaussian mixture models. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Honolulu, HI, USA, 16–20 April 2007; Volume 4, p. IV-317.
99. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
100. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
101. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
102. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

103. Park, H.; Lee, S.; Lee, J.; Ham, B. Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 12046–12055.
104. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 4330–4339.
105. 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**, *33*, 4676–4687. [CrossRef]
106. Ye, M.; Wang, Z.; Lan, X.; Yuen, P.C. Visible thermal person re-identification via dual-constrained top-ranking. In Proceedings of the IJCAI-18, Stockholm, Sweden, 13–19 July 2018; Volume 1, p. 2.
107. Ye, M.; Lan, X.; Li, J.; Yuen, P. Hierarchical discriminative learning for visible thermal person re-identification. *AAAI Conf. Artif. Intell.* **2018**, *32*. [CrossRef]
108. 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. [CrossRef]
109. 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.* **2019**, *15*, 407–419. [CrossRef]
110. Hao, Y.; Wang, N.; Li, J.; Gao, X. HSME: Hypersphere manifold embedding for visible thermal person re-identification. *AAAI Conf. Artif. Intell.* **2019**, *33*, 8385–8392. [CrossRef]
111. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 1522–1531.
112. Meshky, N.M.; Iodice, S.; Mikolajczyk, K. Domain adversarial training for infrared-colour person re-identification. In Proceedings of the 9th International Conference on Imaging for Crime Detection and Prevention (ICDP-2019), London, UK, 16–18 December 2019.
113. Dai, P.; Ji, R.; Wang, H.; Wu, Q.; Huang, Y. Cross-modality person re-identification with generative adversarial training. In Proceedings of the IJCAI-18, Stockholm, Sweden, 13–19 July 2018; Volume 1, p. 6.
114. Shuai, Z.; Li, S.; Gao, Y.; Wu, F. Adversarial Learning Based on Global and Local Features for Cross-Modal Person Re-identification. In Proceedings of the 2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE), Hangzhou, China, 5–7 November 2021; Number 34, pp. 1–4.
115. Hu, W.; Liu, B.; Zeng, H.; Hou, Y.; Hu, H. Adversarial Decoupling and Modality-invariant Representation Learning for Visible-Infrared Person Re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5095–5109. [CrossRef]
116. Fu, X.; Huang, F.; Zhou, Y.; Ma, H.; Xu, X.; Zhang, L. Cross-Modal Cross-Domain Dual Alignment Network for RGB-Infrared Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6874–6887. [CrossRef]
117. Yin, J.; Ma, Z.; Xie, J.; Nie, S.; Liang, K.; Guo, J. DFˆ 2AM: Dual-level Feature Fusion and Affinity Modeling for RGB-Infrared Cross-modality Person Re-identification. *arXiv* **2021**, arXiv:2104.00226.
118. Huang, N.; Liu, J.; Zhang, Q.; Han, J. Exploring modality-shared appearance features and modality-invariant relation features for cross-modality person re-identification. *arXiv* **2021**, arXiv:2104.11539.
119. Chen, C.; Ye, M.; Qi, M.; Wu, J.; Jiang, J.; Lin, C.W. Structure-Aware Positional Transformer for Visible-Infrared Person Re-Identification. *IEEE Trans. Image Process.* **2022**, *31*, 2352–2364. [CrossRef]
120. Gao, W.; Liu, L.; Zhu, L.; Zhang, H. Visible–infrared person re-identification based on key-point feature extraction and optimization. *J. Vis. Commun. Image Represent.* **2022**, *85*, 103511. [CrossRef]
121. Beauchemin, S.S.; Barron, J.L. The computation of optical flow. *ACM Comput. Surv. (CSUR)* **1995**, *27*, 433–466. [CrossRef]
122. Chen, D.; Li, H.; Xiao, T.; Yi, S.; Wang, X. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1169–1178.
123. McLaughlin, N.; Del Rincon, J.M.; Miller, P. Recurrent convolutional network for video-based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1325–1334.
124. Zhang, W.; Yu, X.; He, X. Learning bidirectional temporal cues for video-based person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 2768–2776. [CrossRef]
125. Xu, S.; Cheng, Y.; Gu, K.; Yang, Y.; Chang, S.; Zhou, P. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4733–4742.
126. Chung, D.; Tahboub, K.; Delp, E.J. A two stream siamese convolutional neural network for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1983–1991.
127. Gao, J.; Nevatia, R. Revisiting temporal modeling for video-based person reid. *arXiv* **2018**, arXiv:1805.02104.
128. Gu, X.; Chang, H.; Ma, B.; Zhang, H.; Chen, X. Appearance-preserving 3d convolution for video-based person re-identification. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2020; pp. 228–243.
129. Li, J.; Zhang, S.; Huang, T. Multi-scale 3d convolution network for video based person re-identification. *AAAI Conf. Artif. Intell.* **2019**, *33*, 8618–8625. [CrossRef]

130. Liu, Y.; Yuan, Z.; Zhou, W.; Li, H. Spatial and temporal mutual promotion for video-based person re-identification. *AAAI Conf. Artif. Intell.* **2019**, *33*, 8786–8793. [CrossRef]
131. Aich, A.; Zheng, M.; Karanam, S.; Chen, T.; Roy-Chowdhury, A.K.; Wu, Z. Spatio-Temporal Representation Factorization for Video-based Person Re-Identification. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; Number 41, pp. 152–162. [CrossRef]
132. Hou, R.; Chang, H.; Ma, B.; Huang, R.; Shan, S. BiCnet-TKS: Learning Efficient Spatial-Temporal Representation for Video 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. 2014–2023. [CrossRef]
133. Danielsson, P.E. Euclidean distance mapping. *Comput. Graph. Image Process.* **1980**, *14*, 227–248. [CrossRef]
134. Malkauthekar, M. Analysis of Euclidean distance and Manhattan distance measure in Face recognition. In Proceedings of the Third International Conference on Computational Intelligence and Information Technology (CIIT 2013), Mumbai, India, 18–19 October 2013; pp. 503–507.
135. Gultom, S.; Sriadhi, S.; Martiano, M.; Simarmata, J. Comparison analysis of K-means and K-medoid with Ecluidience distance algorithm, Chanberra distance, and Chebyshev distance for big data clustering. *Proc. Iop Conf. Ser. Mater. Sci. Eng.* **2018**, *420*, 012092. [CrossRef]
136. Sahu, L.; Mohan, B.R. An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop. In Proceedings of the 2014 9th International Conference on Industrial and Information Systems (ICIIS), Gwalior, India, 15–17 December 2014; pp. 1–5.
137. Yang, L.; Jin, R. Distance metric learning: A comprehensive survey. *Mich. State Univ.* **2006**, *2*, 4.
138. Wang, F.; Liu, H. Understanding the behaviour of contrastive loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2495–2504.
139. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
140. Chen, W.; Chen, X.; Zhang, J.; Huang, K. Beyond triplet loss: A deep quadruplet network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 403–412.
141. Kasun, L.L.C.; Zhou, H.; Huang, G.B.; Vong, C.M. *Representational Learning with ELMs for Big Data*; University of Macau: Macau, China, 2013.