

# Beyond Intra-modality: A Survey of Heterogeneous Person Re-identification

Zheng Wang<sup>†1</sup>, Zhixiang Wang<sup>†2</sup>, Yinqiang Zheng<sup>1</sup>, Yang Wu<sup>3</sup>,  
Wenjun Zeng<sup>4</sup> and Shin'ichi Satoh<sup>1</sup>

<sup>1</sup>National Institute of Informatics

<sup>2</sup>National Taiwan University

<sup>3</sup>Kyoto University

<sup>4</sup>Microsoft Research Asia

{wangz, yqzheng, satoh}@nii.ac.jp, {wangzx1994, wuyang0321}@gmail.com, wezeng@microsoft.com

## Abstract

An efficient and effective person re-identification (ReID) system relieves the users from painful and boring video watching and accelerates the process of video analysis. Recently, with the explosive demands of practical applications, a lot of research efforts have been dedicated to heterogeneous person re-identification (Hetero-ReID). In this paper, we provide a comprehensive review of state-of-the-art Hetero-ReID methods that address the challenge of inter-modality discrepancies. According to the application scenario, we classify the methods into four categories — low-resolution, infrared, sketch, and text. We begin with an introduction of ReID, and make a comparison between Homogeneous ReID (Homo-ReID) and Hetero-ReID tasks. Then, we describe and compare existing datasets for performing evaluations, and survey the models that have been widely employed in Hetero-ReID. We also summarize and compare the representative approaches from two perspectives, *i.e.*, the application scenario and the learning pipeline. We conclude by a discussion of some future research directions. *Follow-up updates are available at <https://github.com/lightChaserX/Awesome-Hetero-reID>*

## 1 Introduction

Person re-identification (ReID) is a human-centric AI technology that finds person of interest in a large amount of videos. It facilitates various applications that require painful and boring video watching, including searching for video shots related to an actor of interest from TV series, finding a lost child in a shopping mall from camera videos, and re-identifying suspects in video surveillance systems. Its efficiency and effectiveness accelerate the process of video analysis. In recent years, we have made significant advances in general ReID. The performance is remarkably high on some public datasets [Luo *et al.*, 2019; Wang *et al.*, 2019b; Wang *et al.*, 2019a; Zheng *et al.*, 2019; Zhou *et al.*, 2019], for example, 96.1% Rank-1 accuracy on the Market-1501

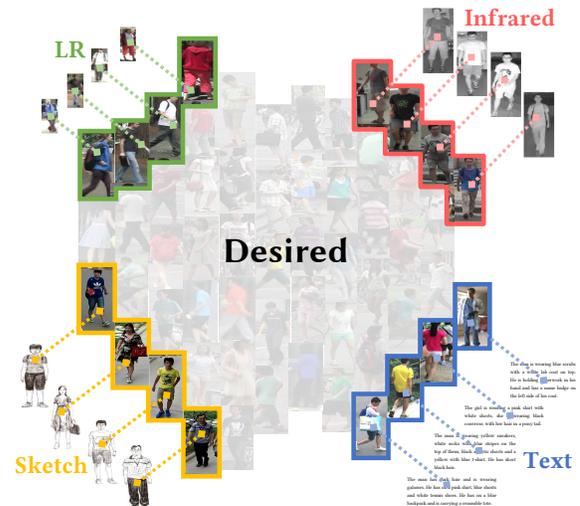


Figure 1: Scope of heterogeneous person re-identification studied in this survey. ‘LR’ indicates low-resolution. The examples are selected respectively from Hetero-ReID LR dataset (MLR-VIPeR), Hetero-ReID IR dataset (SYSU-MM01), Hetero-ReID Sketch dataset (PKU-Sketch) and Hetero-ReID Text dataset (CUHK-PEDES). The dataset details can refer to Table 2.

dataset [Zheng *et al.*, 2015a], which even outperforms human [Zhang *et al.*, 2017]. However, the general ReID assumes an ideal scenario for brevity where all person images are captured in the daytime with the visible spectrum, and have sufficient details to represent a person. Considering all these images are in the desired modality, we name person ReID under this scenario Homogeneous Person Re-identification (Homo-ReID). It just counts the challenges of intra-modality discrepancies, such as pose and viewpoint changes [Zheng *et al.*, 2019; Zhang *et al.*, 2019], background and illumination variations [Zeng *et al.*, 2020], and occlusions [Zheng *et al.*, 2015b; Huang *et al.*, 2015].

In real-world applications, however, it is impractical to assume the data source is always in such a desired modality. For example, to find a criminal, the system have to check person images in low-resolution (LR), or captured by infrared (IR) cameras when the illumination condition is not sufficient. Moreover, witnesses’ descriptions (in terms of text) and

<sup>†</sup>Equal contribution

sketches drawn by artists shall also be used as the cues. These scenarios face challenges beyond the intra-modality discrepancy investigated in Homo-ReID as shown in Figure 1. Algorithms developed under the desired modality lose their effectiveness when applied to these scenarios.

To improve the practical capability of the ReID system, we should make more efforts to bridge the gap between different camera specifications and settings (*e.g.*, low- vs. high-resolution data), different sensory devices (*e.g.*, infrared vs. visible light devices), and reproduction of human memory and direct recording by a camera (*e.g.*, sketch/text description vs. digital images). We define this practical person retrieval Heterogeneous Person Re-identification (Hetero-ReID).

Encouragingly, in the past few years, we have already seen quite a number of remarkable signs of progress on Hetero-ReID, which are primarily assisted by a growing variety of Hetero-ReID benchmark datasets allowing direct comparison of different methodologies. While Homo-ReID attracts much attention in the research community, we feel that it is the right time to call for more attention and research efforts to the important and still largely unexplored area of Hetero-ReID by providing a quality survey. This paper offers a comprehensive and up-to-date review of the diverse and growing array of Hetero-ReID techniques.

**The differences between this survey and the previous ones.** To our knowledge, there are few reviews in the ReID field. [Nambiar *et al.*, 2019] explored the ReID applications built on gait sequences, which is a special and different focus. [Vezzani *et al.*, 2013; Zheng *et al.*, 2016; Gou *et al.*, 2018] focused on Homo-ReID. [Vezzani *et al.*, 2013] made a multidimensional overview of Homo-ReID. [Zheng *et al.*, 2016] described critical future directions and briefed some important yet under-developed issues. [Gou *et al.*, 2018] conducted a systematic evaluation with different features and metrics. These works well position the current progress of Homo-ReID. However, they have no investigation on Hetero-ReID and confine to address the intra-modality discrepancy. [Leng *et al.*, 2019; Ye *et al.*, 2020] started to look into some inter-modality challenges, but provided a limited summary of current efforts or problems present in Hetero-ReID. Note that in other areas, face recognition [Ouyang *et al.*, 2016] and image retrieval [Wang *et al.*, 2016a] also raise the urgent requirements of addressing the inter-modality discrepancy. This survey seeks to provide a summary of current research from different perspectives.

**Goals of this survey.** We aim to (i) present the idea of Hetero-ReID and raise awareness of this problem; (ii) thoroughly review the literature on Hetero-ReID and provide a panorama for researchers in other fields, with which they can quickly understand and step into the new area; (iii) give comprehensive guidance of future directions to peers in our community.

**Contributions.** (i) We highlight the difference between Hetero-ReID and Homo-ReID, and conduct a systematic review of Hetero-ReID beyond intra-modality, where the inter-modality discrepancy is the main challenge. (ii) We categorize related works into different modalities on which they operate and summarize the state-of-the-arts. As Figure 1 shows, we consider four cross-modality application scenarios: Low-

	Homo-ReID	Hetero-ReID
Media Type	Desired image	+ LR / IR / Sketch / Text
Participant	Machine	Machine (+Human)
Main Challenge	Intra-modality	Intra- + Inter-modality
#Publications	>1000	<100
Performance	96.1	42.50 / 28.90 / 34.00 / 53.14

Table 1: Comparison between Homo-ReID and Hetero-ReID. ‘#Publications’ represents the number of related publications. We use CMC-1 values (%) to show the performances (higher is better). Note that the results are recoded by the end of 2019.

resolution (LR), Infrared (IR), Sketch, and Text. (iii) We analyze representative methods and identify future directions in this research field to share the vision and expand the horizons of Hetero-ReID.

## 2 Person Re-identification

### 2.1 ReID Diagram

ReID system always consists of a feature extraction part and a descriptor matching part. At first, feature vectors of a gallery  $\mathcal{G} = \{f_{g_1}, f_{g_2}, \dots, f_{g_N}\}$  are extracted by the pre-trained feature extraction part. When we input a probe into the system, the probe’s feature vector  $f_p$  is obtained by the feature extraction part. After that, the feature vector  $f_p$  matches against feature vectors  $\mathcal{G}$  by the descriptor matching part. For Homo-ReID, both probe samples and gallery samples are from the desired modality (daytime, visible spectrum, and high-resolution images). While for Hetero-ReID, the probe sample or gallery samples are from another modality (see Figure 2).

### 2.2 Hetero-ReID vs. Homo-ReID

Homo-ReID only faces the challenges of persons’ appearance changes, and person images are captured in the desired modality with only intra-modality discrepancy. Following the diagram of ReID, researches tried to design robust and discriminative features or learn effective distance metrics. For more details, the readers can refer to Homo-ReID surveys [Gou *et al.*, 2018]. Hetero-ReID should face more challenges beyond intra-modality discrepancies, such as HR-to-LR discrepancy, RGB-to-IR discrepancy, Photo-to-Sketch discrepancy, and Image-to-Text discrepancy. These

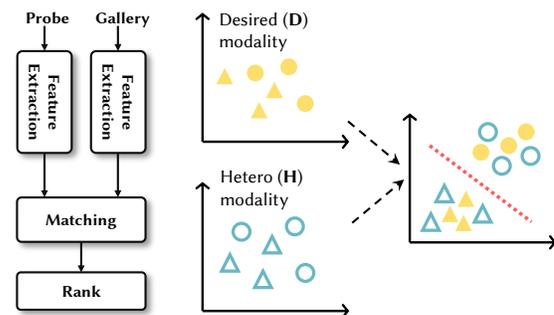


Figure 2: The diagram of Hetero-ReID.

extremely large discrepancies are generated from different modalities. Hence, the approaches of Hetero-ReID should pay more attention to remove the inter-modality discrepancy. Note that we don't attribute the RGB-depth based ReID to our Hetero-ReID, although there are a couple of researches investigated this kind of task [Barbosa *et al.*, 2012; Mogelmoose *et al.*, 2013; Haque *et al.*, 2016; Wu *et al.*, 2017a; Munaro *et al.*, 2014]. In realistic applications, the depth channel works as a complement for the RGB channels, and a depth image of a person will not be used to retrieval in an RGB image gallery.

We make a comparison between Homo-ReID and Hetero-ReID, as Table 1 shows. For the media type, Homo-ReID only exploits desired RGB images, while Hetero-ReID takes additional LR/IR/Sketch images or Text description into account. For the participant, Homo-ReID only employs the resources from machine intelligence, while Hetero-ReID also brings in the input from human intelligence, such as sketch images drawn by painters or text descriptions of the suspect by witnesses. For the main challenge, Homo-ReID only needs to deal with the intra-modality discrepancy, such as viewpoint variations, image misaligned or occlusion and illumination changes, while Hetero-ReID should bridge the gaps deriving from both intra- and inter-modality discrepancies. Since additional inter-modality discrepancies are larger than intra-modality discrepancies, Hetero-ReID is more challenging. As intra- and inter-modality discrepancies are essentially different, the methods designed for Homo-ReID cannot be directly used in Hetero-ReID.

There is also a big performance gap between Homo-ReID and Hetero-ReID. The state-of-the-art performances (CMC-1) are reported in Table 1. They are respectively CAR [Zhou *et al.*, 2019] on Market-1501 [Zheng *et al.*, 2015a] for Homo-ReID, RAIN [Chen *et al.*, 2019] on MLR-VIPER [Jiao *et al.*, 2018] for Hetero-ReID LR, D<sup>2</sup>RL [Wang *et al.*, 2019c] on SYSU-MM01 [Wu *et al.*, 2017b] for Hetero-ReID IR, CDAFL [Pang *et al.*, 2018] on PKU-Sketch [Pang *et al.*, 2018] for Hetero-ReID Sketch, and A-GANet [Liu *et al.*, 2019] on CUHK-PEDES [Li *et al.*, 2017b] for Hetero-ReID Text. The performances of Hetero-ReID are still not so satisfied as that of Homo-ReID. However, compared with Homo-ReID, there are not so many publications related to Hetero-ReID. Considering the realistic value and research significance, we feel that it is the right time to call for more attention and research efforts to the area of Hetero-ReID.

### 3 Available Datasets and Evaluation Metrics

#### 3.1 Datasets

We summarize available datasets for Hetero-ReID in Table 2, including the application scenario, the total number of cameras, identities, and samples. To make a comparison, we also list two typical Homo-ReID datasets, including Market-1501 and MSMT17. The Hetero-ReID datasets are categorized into four types, and we also take notice of the data construction manner, *i.e.*, simulated or really collected. Figure 1 shows some samples for typical datasets. We have the following summaries: (i) Although many Hetero-ReID LR datasets were constructed, most of them are simulated *w.r.t.*

No.	Dataset	Appl.	Type	#Cam.	#ID	#Sam.
	Market-1501 [Zheng <i>et al.</i> , 2015a]	Desired	Real	6	1,501	32,668
	MSMT17 [Wei <i>et al.</i> , 2018]	Desired	Real	15	4,101	126,441
1	CAVIAR [Cheng <i>et al.</i> , 2011]	LR	Real	2	50	1,000
2	LR-VIPeR [Li <i>et al.</i> , 2015]	LR	Simulated	2	632	1,264
3	LR-3DPES [Li <i>et al.</i> , 2015]	LR	Simulated	8	192	1,011
4	LR-i-LIDS [Jing <i>et al.</i> , 2015]	LR	Simulated	2	119	238
5	LR-PRID [Jing <i>et al.</i> , 2015]	LR	Simulated	2	100	200
6	SALR-VIPeR [Wang <i>et al.</i> , 2016b]	LR	Simulated	2	632	1,264
7	SALR-PRID [Wang <i>et al.</i> , 2016b]	LR	Simulated	2	450	900
8	MLR-VIPeR [Jiao <i>et al.</i> , 2018]	LR	Simulated	2	632	1,264
9	MLR-SYSU [Jiao <i>et al.</i> , 2018]	LR	Simulated	2	502	3,012
10	MLR-CUHK03 [Jiao <i>et al.</i> , 2018]	LR	Simulated	2	1,467	14,000
11	SYSU-MM01 [Wu <i>et al.</i> , 2017b]	IR	Real	6	491	38,271
12	RegDB [Nguyen <i>et al.</i> , 2017]	IR	Real	2	412	8,240
13	PKU-Sketch [Pang <i>et al.</i> , 2018]	Sketch	Real	2	200	400
14	CUHK-PEDES [Li <i>et al.</i> , 2017b]	Text	Real	-	13,003	80,412

Table 2: Released and freely available datasets. ‘Appl.’, ‘#Cam.’, ‘#ID’ and ‘#Sam.’ respectively represent for the application scenario, the total number of cameras, identities and samples.

from Homo-ReID datasets. (ii) Only two Hetero-ReID IR datasets were released. SYSU-MM01 is an active IR dataset, while RegDB is a passive one. Only one Hetero-ReID Sketch dataset and one Hetero-ReID Text dataset were constructed. (iii) Compared with Homo-ReID, there is still a lot of room to construct more available datasets. For Hetero-ReID LR, it requires to construct practical datasets instead of the simulated ones. For the other three kinds of application scenarios, datasets under different conditions/styles or with larger scales should be constructed.

#### 3.2 Evaluation Metrics

Evaluation metrics are the same for Homo-ReID and Hetero-ReID. When evaluating Hetero-ReID methods, we often choose the Cumulative Matching Characteristics (CMC) [Wang *et al.*, 2007] curve and the mean average precision (mAP) [Zheng *et al.*, 2015a] as the evaluation metrics. In practice, when only one ground truth exists in the gallery, or we are caring about the true matches ranking in top positions of the returning list, the CMC curve is an acceptable selection. On the other hand, if multiple true matches exist in the gallery and we would like to retrieval all the results, mAP is a good selection. Since mAP cares more about the ability of retrieval recall, while CMC does not. Hence, CMC and mAP are always used together for Hetero-ReID.

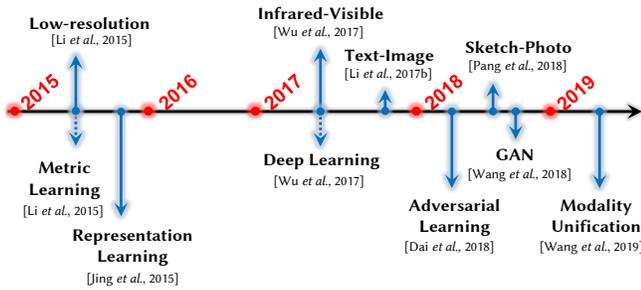


Figure 3: Milestones of existing Hetero-ReID studies. As shown in the figure, Hetero-ReID studies started in 2015. Application scenarios, pipelines and new technical trends are highlighted.

## 4 Hetero-ReID Methods

In this section, we first provide the milestones of existing Hetero-ReID studies, and then make a survey of Hetero-ReID methods from two perspectives. From the perspective of application scenario, we classify all related methods into four kinds of Hetero-ReID application scenarios. From the perspective of learning pipeline, we categorize typical methods into three types.

### 4.1 Milestones of Existing Hetero-ReID Studies

Through the unremitting efforts of AI researchers, Hetero-ReID has achieved remarkable success in different aspects. We draw a timeline to introduce important milestones for Hetero-ReID and present them in Figure 3. In particular, ‘Low-resolution’, ‘Infrared-Visible’, ‘Text-Image’, and ‘Sketch-Photo’ indicate milestones of application scenarios. Their typical methods are introduced in § 4.2. ‘Metric Learning’, ‘Representation Learning’, and ‘Modality Unification’ indicate milestones of learning pipelines, which are discussed in § 4.3. Besides, milestones of new technical trends, such as deep learning, adversarial learning, and GAN, are pointed out as well.

### 4.2 From the Perspective of Application Scenario

**Hetero-ReID LR.** Commonly, the resolution of person image changes a lot, due to the variations in the person-camera distance and camera deployment settings. Hetero-ReID LR application scenario attempts to compare images with different resolutions, where low-resolution leads to an extreme loss of appearance information and a large increase of discrepancy. [Jing et al., 2015] was the first to investigate the Hetero-ReID LR task. They designed a semi-coupled low-rank discriminant dictionary learning method, constructing a relationship mapping from the features of normal person images to that of LR person images. [Li et al., 2015] proposed a joint multi-scale learning framework by learning metrics on feature domains of two different image scales simultaneously. [Wang et al., 2016b] changed the problem to be low-resolution with different scales. They observed that scale-distance functions could be classified, and then learned a surface in function space to separate functions to identify persons. Jiao et al. [Jiao et al., 2018] developed a super-resolution and identity joint learning method to improve the Hetero-ReID LR performance. Similar to [Jing et al., 2015],

Reference	Conf.	Appl.	Techn.	Datasets	Pipeline
[Wang et al., 2016b]	IJCAI	LR	Subspace Learning	1 6 7	RL
[Li et al., 2015]	ICCV	LR	Metric Learning	1 2 3	ML
[Jing et al., 2015]	CVPR	LR	Dictionary Learning	2 4 5	RL
[Li et al., 2018]	AAAI	LR	Dictionary Learning	2 4	RL
[Jiao et al., 2018]	AAAI	LR	Super Resolution	1 8 9 10	MU
[Wang et al., 2018]	IJCAI	LR	Super Resolution	1 6 7	MU
[Mao et al., 2019]	IJCAI	LR	Super Resolution	1 8 10	MU
[Chen et al., 2019]	AAAI	LR	Resolution Adaptation	1 8 10	MU
[Ye et al., 2018a]	AAAI	IR	Metric Learning	12	ML
[Wu et al., 2017b]	ICCV	IR	Deep Zero-Padding	11	RL
[Ye et al., 2018b]	IJCAI	IR	Feature Learning	11 12	RL
[Dai et al., 2018]	IJCAI	IR	Feature Embedding	11	RL
[Wang et al., 2019c]	CVPR	IR	Image Generation	11 12	MU
[Pang et al., 2018]	MM	Sketch	Sketch Feature Learning	13	RL
[Li et al., 2017b]	CVPR	Text	Affinity Learning	14	ML
[Li et al., 2017a]	ICCV	Text	Feature Learning	14	RL
[Chen et al., 2018b]	WACV	Text	Patch Matching	14	ML
[Chen et al., 2018a]	ECCV	Text	Association Learning	14	RL
[Liu et al., 2019]	MM	Text	Adversarial Learning	14	RL

Table 3: Representative methods employed in Hetero-ReID. In the datasets column, the number denotes for the certain dataset in Table 2.

[Li et al., 2018] designed a semi-coupled projective dictionary learning model to bridge the gap across different resolutions. [Chen et al., 2019] combined a resolution adaptation network and a re-identification network together to solve Hetero-ReID LR problem. [Wang et al., 2018] first cascaded multiple SR-GANs in series to promote the ability of scale-adaptive super-resolution, then plugged-in a re-identification network to enhance the ability of person representation. [Mao et al., 2019] jointly trained a Foreground-Focus Super-Resolution module and a Resolution-Invariant Feature Extractor, and then obtained a strong resolution invariant representation. [Cheng et al., 2020] discovered the underlying association knowledge between image SR and ReID, and leveraged it as an extra learning constraint for enhancing the compatibility of SR and ReID models.

**Hetero-ReID IR.** The visible cameras are not able to capture clear and valid appearance information under poor illumination environments (e.g. during the nighttime), which limits the applicability of ReID in practical surveillance applications. Hetero-ReID IR application scenario provides a good supplement for nighttime surveillance applications, which also introduces a large intra-modality discrepancy. [Wu et al., 2017b] was the first to investigate Hetero-ReID task in

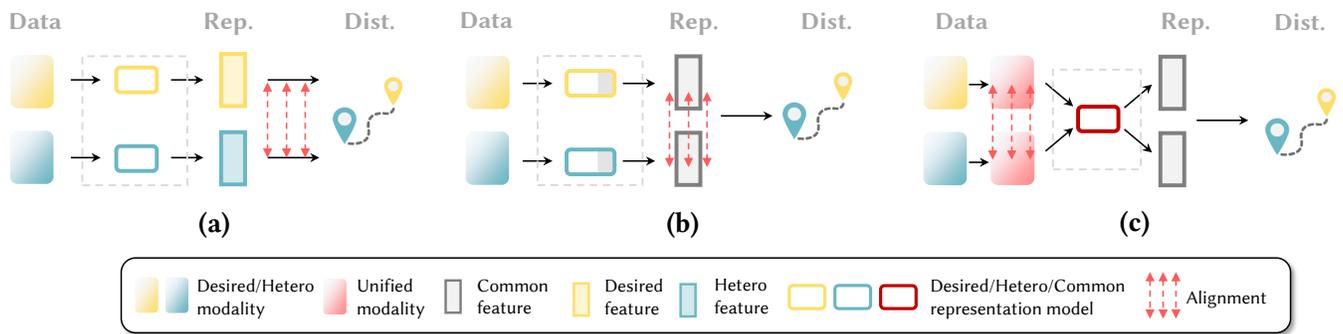


Figure 4: Summary of three learning pipelines. (a) utilizes the metric learning method to learn how to match representations from separate representation models, *i.e.*, ‘ML’; (b) focuses on learning the common representation of different modalities, *i.e.*, ‘RL’; (c) pays attention to generating the modality-unified samples and then learning the common representation, *i.e.*, ‘MU’.

the Hetero-ReID IR task. They proposed a deep zero-padding module to improve the one-stream network, making the implicit network structure more suitable. [Ye *et al.*, 2018a] jointly optimized the modality-specific and modality-shared metrics and designed a hierarchical cross-modality matching model. [Ye *et al.*, 2018b; Ye *et al.*, 2019] attempted to learn discriminative features with a bi-directional dual-constrained top-ranking loss function. [Dai *et al.*, 2018] proposed to learn discriminative common representations with a generator for learning image representations and a discriminator for discriminating the modalities of RGB and IR images. [Wang *et al.*, 2019c] presented a new framework, taking advantage of CycleGAN to reduce modality discrepancy from image-level and advantage of sophisticated Homo-ReID models to reduce appearance discrepancy from feature-level. [Wang *et al.*, 2020] jointly exploited pixel alignment and feature alignment by generators. [Li *et al.*, 2020] introduced an auxiliary X modality as an assistant and an X-Infrared-Visible three-mode learning framework. [Kansal *et al.*, 2020] designed a network with disentanglement loss to distill identity features and dispel spectrum features. [Choi *et al.*, 2020] also attempted to disentangle ID-discriminative factors and ID-excluded factors from images, and they designed an ID-preserving person image generation network to implement the idea. [Jin *et al.*, 2020] also proposed to distill identity-relevant feature from the removed spectrum style information and reconstitute it to the network to ensure high discrimination. [Lu *et al.*, 2020] explored the modality-shared information and the modality-specific characteristics to boost the performance. [Yang *et al.*, 2020] proposed a bi-directional random walk scheme to mining more reliable relationships between images by traversing heterogeneous manifolds in the feature space of each modality. Note that the method [Yang *et al.*, 2020] performs as a re-ranking method thus obtains relative high performances.

**Hetero-ReID Sketch.** When someone reports a lost child case while the child was not identified by the surveillance camera, we call for an algorithm to automatically pick out the potential candidates according to a sketch by an artist’s drawing. Consequently, investigators will be able to narrow down and hence expedite their search. [Pang *et al.*, 2018] was the first and only one to investigate Hetero-ReID task in the

Sketch application scenario. They not only raised a dataset, but also proposed to learn identity features and modality-invariant features by exploiting a cross-modal adversarial feature learning framework.

**Hetero-ReID Text.** Searching a person with free-form natural language descriptions can be widely applied. In the early years, some researches started to investigate the ReID with attribute-based queries [Feris *et al.*, 2014; Ye *et al.*, 2015; Wang *et al.*, 2015; Yin *et al.*, 2018]. Hetero-ReID Text application scenario denotes a situation that the queries are natural language descriptions. [Li *et al.*, 2017b] evaluated and compared a wide range of possible models, and proposed an RNN with Gated Neural Attention mechanism. [Li *et al.*, 2017a] learned to embed cross-modal features, and then refined the matching results with a co-attention mechanism. [Zhang and Lu, 2018] designed a cross-modal projection matching loss and a cross-modal projection classification loss for learning discriminative image-text embeddings. [Chen *et al.*, 2018b] proposed a patch-word matching model and designed an adaptive threshold mechanism into the model. [Chen *et al.*, 2018a] attempted to build global and local image-language associations, which enforce semantic consistencies between local visual and linguistic features. [Liu *et al.*, 2019] proposed a deep adversarial graph attention convolution network.

**Summary.** From the perspective of application scenario, we make the following summaries:

- *Most of the methods selected a deep learning framework.* It is probably because all the heterogeneous application scenarios are raised in recent years, and deep learning methods are in their high-speed development period. In general, deep learning methods are superior in shared feature learning and classification tasks when a significant amount of training samples are available.
- *Different methods have different focuses.* To fill the gaps between the desired modality and other heterogeneous modalities, some methods (*e.g.* [Li *et al.*, 2015]) try to learn a metric, others (like [Ye *et al.*, 2018a]) attempt to learn shared features, and there are also methods (for example [Wang *et al.*, 2019c]) unifying the modality on the data level. In § 4.3, we will make an analysis and discussion.

- *The existing researches in each application scenario still have many limitations.* For the Hetero-ReID LR application scenario, recent researches are mainly evaluated on the simulated datasets. For the Hetero-ReID Sketch application scenario, only one research touched this direction, due to the hardness of building a benchmark.

### 4.3 From the Perspective of Learning Pipeline

Representative methods employed in Hetero-ReID are compared in Table 3. We list some key and valuable information, including the reference paper, conference (‘Conf.’), application scenario (‘Appl.’), technology (‘Techn.’), used datasets, and Learning pipeline. We can find that all representative works are published in the recent five years. It means that Hetero-ReID is a relatively new research topic. Among all the application scenarios, Hetero-ReID Sketch is less investigated. As described in the sections above, we consider that if more datasets are available, more works could be conducted and published. Particularly, to address the inter-modality discrepancy, in our opinion, representative methods can be categorized into three kinds of learning pipelines.

Figure 4 shows the diagrams of the learning pipelines. Pipeline (a) employs the metric learning method to learn how to match the features from separate representation learning models, and the representation learning models are trained separately with single modality data samples (represented as ‘ML’). Pipeline (b) focuses on learning shared feature models of different modalities, and training data come from both modalities (represented as ‘RL’). Pipeline (c) pays attention to generating the unified-modality samples (represented as ‘MU’), for example, using a super-resolution method to generate high-resolution images from low-resolution images, or using some image generation method to generate infrared images from RGB images and RGB images from infrared images.

As shown in Table 3, for each kind of application scenario, from top to bottom, the methods become more effective. Generally, pipeline ‘MU’ performs better than pipeline ‘RL’, and better than pipeline ‘ML’. Hence, unifying the modalities of samples is an effective way to fill the modality gap. For the Hetero-ReID LR application scenario, the super-resolution methods can be used, such as [Wang *et al.*, 2018; Mao *et al.*, 2019]. Thanks to the developments of GAN, unifying the modalities is considerable as [Wang *et al.*, 2019c].

## 5 Conclusion and Future Directions

This paper overviews recent developments of Hetero-ReID. It covers most of the literature on Hetero-ReID. We summarized available datasets, the widely employed methods, and compared the existing techniques. We believe that Hetero-ReID will continue to be an active and promising research area with broad potential applications. Many issues in Hetero-ReID, however, still remain.

- *Dataset Construction.* For Hetero-ReID LR application scenario, the researchers can only get simulated datasets. For the other types of applications, although the researchers can achieve practical datasets, the choice is limited. In addition, large scale Hetero-ReID datasets are

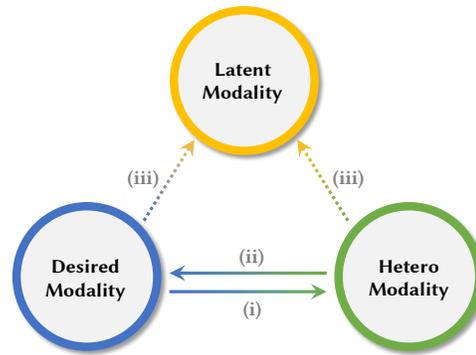


Figure 5: Three kinds of modality transfer. To unify the modality, we can transfer (i) from heterogeneous modality to desired modality, (ii) from desired modality to heterogeneous modality, and (iii) from both desired and heterogeneous modalities to a latent modality.

also required. It is somehow urgent for us to build new datasets, and push forward this research area. On the other hand, synthesizing more realistic datasets (like [Zeng *et al.*, 2020]) is also a good way to make one step forward.

- *Taking Advantages of Homo-ReID Datasets and Methods.* There are sufficient datasets and methods in Homo-ReID. It is easy to address the intra-modality discrepancy in the desired modality. Obviously, the feature space of the desired modality can be well-constructed. Since the heterogeneous modality samples are not enough, it is hard to build a perfect heterogeneous feature space. However, it is reasonable to project the Hetero-ReID samples into the desired modality (like [Yang *et al.*, 2020]).
- *Human Interaction and Crowd-sourcing.* For Hetero-ReID Sketch and Text application scenarios, human intelligence joins into the process of ReID. Human intelligence is sometimes subjective and incomplete. So we should consider how to mine and integrate useful information to help search the target in the surveillance system. On the other hand, a lot of witnesses will provide their cues. Each person may have a different view so that the crowd-sourcing cues may be diverse to each other. We should design a strategy to remove conflict and filter valuable information.
- *Investigation on Unifying the Modality.* For Hetero-ReID LR application scenario, some methods [Wang *et al.*, 2018; Mao *et al.*, 2019] attempted to use super-resolution technologies to unify the modality, where they transfer the heterogeneous modality LR image to the desired modality. For Hetero-ReID IR application scenario, with the help of CycleGAN, one work [Wang *et al.*, 2019c] not only transferred the image from heterogeneous (IR) modality to desired (RGB) modality but also transferred the image from desired modality to heterogeneous modality. However, for the Hetero-ReID Text and Sketch application scenarios, no work has investigated to unify the data modality. On the other hand, we can also investigate to unify the modality to a latent modality as Figure 5 shows, for example, a middle-level resolution for the Hetero-ReID LR application scenario and a hyperspectral image for the Hetero-ReID IR application scenario.

- *Integrating Multiple Hetero-ReID Application Scenarios.* For a practical system, it may require different kinds of inputs to search out the target. The Hetero-ReID application scenarios can be equipped in different stages. If we can integrate different kinds of inputs, more valuable information could be used for retrieval, since different inputs have different attentions and views for the target. It would raise a novel multiple cross-modality research task.
- *Considering the Privacy Issue.* It is fine to conduct academic research on public datasets. But with the leakage of personal image data, privacy concerns are raising nowadays when algorithms need to be applied to practical applications. Recently, pioneer works [Mirjalili and Ross, 2017] have explored to hide some of private information presented in the images. Further research on the principles of visual cryptography, signal mixing and image perturbation to protect users' privacy on person templates are essential for addressing public concern on privacy. In particular, for Hetero-ReID, since it contains multi-modality data, person templates of different modalities or a latent modality should be investigated to protect.

## Acknowledgements

This work was supported partly by the JST CREST Grant JPMJCR1686, partly by the Grant-in-Aid for JSPS Fellows 18F18378, and partly by the Microsoft Collaborative Research Grant.

## References

- [Barbosa *et al.*, 2012] Igor Barros Barbosa, Marco Cristani, Alessio Del Bue, Loris Bazzani, and Vittorio Murino. Re-identification with rgb-d sensors. In *ECCV*, 2012.
- [Chen *et al.*, 2018a] Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. Improving deep visual representation for person re-identification by global and local image-language association. In *ECCV*, 2018.
- [Chen *et al.*, 2018b] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving text-based person search by spatial matching and adaptive threshold. In *WACV*, 2018.
- [Chen *et al.*, 2019] Yun-Chun Chen, Yu-Jhe Li, Xiaofei Du, and Yu-Chiang Frank Wang. Learning resolution-invariant deep representations for person re-identification. In *AAAI*, 2019.
- [Cheng *et al.*, 2011] Dong Seon Cheng, Marco Cristani, Michele Stoppa, Loris Bazzani, and Vittorio Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011.
- [Cheng *et al.*, 2020] Zhiyi Cheng, Qi Dong, Shaogang Gong, and Xiatian Zhu. Inter-task association critic for cross-resolution person re-identification. In *CVPR*, 2020.
- [Choi *et al.*, 2020] Seokeon Choi, Sumin Lee, Youngeun Kim, Taekyung Kim, and Changick Kim. Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In *CVPR*, 2020.
- [Dai *et al.*, 2018] Pingyang Dai, Rongrong Ji, Haibin Wang, Qiong Wu, and Yuyu Huang. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, 2018.
- [Feris *et al.*, 2014] Rogerio Feris, Russel Bobbitt, Lisa Brown, and Sharath Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *ICMR*, 2014.
- [Gou *et al.*, 2018] Mengran Gou, Ziyang Wu, Angels Rates-Borras, Octavia Camps, Richard J Radke, et al. A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets. *TPAMI*, 2018.
- [Haque *et al.*, 2016] Albert Haque, Alexandre Alahi, and Li Fei-Fei. Recurrent attention models for depth-based person identification. In *CVPR*, 2016.
- [Huang *et al.*, 2015] Bingyue Huang, Jun Chen, Yimin Wang, Chao Liang, Zheng Wang, and Kaimin Sun. Sparsity-based occlusion handling method for person re-identification. In *MMM*, 2015.
- [Jiao *et al.*, 2018] Jiening Jiao, Wei-Shi Zheng, Ancong Wu, Xiatian Zhu, and Shaogang Gong. Deep low-resolution person re-identification. In *AAAI*, 2018.
- [Jin *et al.*, 2020] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, 2020.
- [Jing *et al.*, 2015] Xiao-Yuan Jing, Xiaoke Zhu, Fei Wu, Xingye You, Qinglong Liu, Dong Yue, Ruimin Hu, and Baowen Xu. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning. In *CVPR*, 2015.
- [Kansal *et al.*, 2020] Kajal Kansal, AV Subramanyam, Zheng Wang, and Shin'ichi Satoh. Sdl: Spectrum-disentangled representation learning for visible-infrared person re-identification. *TCSVT*, 2020.
- [Leng *et al.*, 2019] Qingming Leng, Mang Ye, and Qi Tian. A survey of open-world person re-identification. *TCSVT*, 2019.
- [Li *et al.*, 2015] Xiang Li, Wei-Shi Zheng, Xiaojuan Wang, Tao Xiang, and Shaogang Gong. Multi-scale learning for low-resolution person re-identification. In *ICCV*, 2015.
- [Li *et al.*, 2017a] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *ICCV*, 2017.
- [Li *et al.*, 2017b] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017.
- [Li *et al.*, 2018] Kai Li, Zhengming Ding, Sheng Li, and Yun Fu. Discriminative semi-coupled projective dictionary learning for low-resolution person re-identification. In *AAAI*, 2018.
- [Li *et al.*, 2020] Diangang Li, Xing Wei, Xiaopeng Hong, and Yihong Gong. Infrared-visible cross-modal person re-identification with an x modality. In *AAAI*, 2020.
- [Liu *et al.*, 2019] Jiawei Liu, Zheng-Jun Zha, Richang Hong, Meng Wang, and Yongdong Zhang. Deep adversarial graph attention convolution network for text-based person search. In *ACM MM*, 2019.
- [Lu *et al.*, 2020] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, 2020.
- [Luo *et al.*, 2019] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. A strong baseline and batch normalization neck for deep person re-identification. *TMM*, 2019.
- [Mao *et al.*, 2019] Shunan Mao, Shiliang Zhang, and Ming Yang. Resolution-invariant person re-identification. In *IJCAI*, 2019.
- [Mirjalili and Ross, 2017] Vahid Mirjalili and Arun Ross. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *IJCB*, 2017.

- [Mogelmose *et al.*, 2013] Andreas Mogelmose, Chris Bahnsen, Thomas Moeslund, Albert Clapes, and Sergio Escalera. Tri-modal person re-identification with rgb, depth and thermal features. In *CVPRW*, 2013.
- [Munaro *et al.*, 2014] Matteo Munaro, Alberto Basso, Andrea Fosfati, Luc Van Gool, and Emanuele Menegatti. 3d reconstruction of freely moving persons for re-identification with a depth sensor. In *ICRA*, 2014.
- [Nambiar *et al.*, 2019] Athira Nambiar, Alexandre Bernardino, and Jacinto C Nascimento. Gait-based person re-identification: A survey. *ACM Computing Surveys*, 2019.
- [Nguyen *et al.*, 2017] Dat Nguyen, Hyung Hong, Ki Kim, and Kang Park. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors*, 2017.
- [Ouyang *et al.*, 2016] Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, Xueming Li, Chen Change Loy, and Xiaogang Wang. A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution. *Image and Vision Computing*, 2016.
- [Pang *et al.*, 2018] Lu Pang, Yaowei Wang, Yi-Zhe Song, Tiejun Huang, and Yonghong Tian. Cross-domain adversarial feature learning for sketch re-identification. In *ACM MM*, 2018.
- [Vezzani *et al.*, 2013] Roberto Vezzani, Davide Baltieri, and Rita Cucchiara. People reidentification in surveillance and forensics: A survey. *ACM Computing Surveys*, 2013.
- [Wang *et al.*, 2007] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. Shape and appearance context modeling. In *ICCV*, 2007.
- [Wang *et al.*, 2015] Zheng Wang, Ruimin Hu, Yi Yu, Chao Liang, and Wenxin Huang. Multi-level fusion for person re-identification with incomplete marks. In *ACM MM*, 2015.
- [Wang *et al.*, 2016a] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*, 2016.
- [Wang *et al.*, 2016b] Zheng Wang, Ruimin Hu, Yi Yu, Junjun Jiang, Chao Liang, and Jinqiao Wang. Scale-adaptive low-resolution person re-identification via learning a discriminating surface. In *IJCAI*, 2016.
- [Wang *et al.*, 2018] Zheng Wang, Mang Ye, Fan Yang, Xiang Bai, and Shin'ichi Satoh. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *IJCAI*, 2018.
- [Wang *et al.*, 2019a] Zheng Wang, Junjun Jiang, Yang Wu, Mang Ye, Xiang Bai, and Shin'ichi Satoh. Learning sparse and identity-preserved hidden attributes for person re-identification. *TMM*, 2019.
- [Wang *et al.*, 2019b] Zheng Wang, Junjun Jiang, Yi Yu, and Shin'ichi Satoh. Incremental re-identification by cross-direction and cross-ranking adaption. *TMM*, 2019.
- [Wang *et al.*, 2019c] Zhixiang Wang, Zheng Wang, Yinqiang Zheng, Yung-Yu Chuang, and Shin'ichi Satoh. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, 2019.
- [Wang *et al.*, 2020] Guan-An Wang, Tianzhu Zhang Yang, Jian Cheng, Jianlong Chang, Xu Liang, and Zengguang Hou. Cross-modality paired-images generation for rgb-infrared person re-identification. In *AAAI*, 2020.
- [Wei *et al.*, 2018] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.
- [Wu *et al.*, 2017a] Ancong Wu, Wei-Shi Zheng, and Jian-Huang Lai. Robust depth-based person re-identification. *TIP*, 2017.
- [Wu *et al.*, 2017b] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *ICCV*, 2017.
- [Yang *et al.*, 2020] Fan Yang, Zheng Wang, Jing Xiao, and Shin'ichi Satoh. Mining on heterogeneous manifolds for zero-shot cross-modal image retrieval. In *AAAI*, 2020.
- [Ye *et al.*, 2015] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, Jun Chen, and Jun Liu. Specific person retrieval via incomplete text description. In *ICMR*, 2015.
- [Ye *et al.*, 2018a] Mang Ye, Xiangyuan Lan, Jiawei Li, and Pong C Yuen. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*, 2018.
- [Ye *et al.*, 2018b] Mang Ye, Zheng Wang, Xiangyuan Lan, and Pong C Yuen. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, 2018.
- [Ye *et al.*, 2019] Mang Ye, Xiangyuan Lan, Zheng Wang, and Pong C Yuen. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *TIFS*, 2019.
- [Ye *et al.*, 2020] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *arXiv preprint arXiv:2001.04193*, 2020.
- [Yin *et al.*, 2018] Zhou Yin, Wei-Shi Zheng, Ancong Wu, Hong-Xing Yu, Hai Wan, Xiaowei Guo, Feiyue Huang, and Jianhuang Lai. Adversarial attribute-image person re-identification. In *IJCAI*, 2018.
- [Zeng *et al.*, 2020] Zelong Zeng, Zhixiang Wang, Zheng Wang, Yung-Yu Chuang, and Shin'ichi Satoh. Illumination-adaptive person re-identification. *TMM*, 2020.
- [Zhang and Lu, 2018] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, 2018.
- [Zhang *et al.*, 2017] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017.
- [Zhang *et al.*, 2019] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *CVPR*, 2019.
- [Zheng *et al.*, 2015a] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.
- [Zheng *et al.*, 2015b] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shaogang Gong. Partial person re-identification. In *ICCV*, 2015.
- [Zheng *et al.*, 2016] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.
- [Zheng *et al.*, 2019] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. *TIP*, 2019.
- [Zhou *et al.*, 2019] Sanping Zhou, Fei Wang, Zeyi Huang, and Junjun Wang. Discriminative feature learning with consistent attention regularization for person re-identification. In *ICCV*, 2019.