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

Zheng Wang\textsuperscript{1}\textsuperscript{*} Zhixiang Wang\textsuperscript{2}\textsuperscript{*} Yang Wu\textsuperscript{3} Jingdong Wang\textsuperscript{4} Shin’ichi Satoh\textsuperscript{1,5}

\textsuperscript{1}National Institute of Informatics \hspace{1cm} \textsuperscript{2}National Taiwan University
\textsuperscript{3}Nara Institute of Science and Technology \hspace{1cm} \textsuperscript{4}Microsoft Research \hspace{1cm} \textsuperscript{5}The University of Tokyo

\{wangz, satoh\}@nii.ac.jp, wangzx1994@gmail.com, yangwu@rsc.naist.jp, jingdw@microsoft.com

Abstract

An effective and efficient person re-identification (ReID) algorithm will alleviate painful video watching, and accelerate the investigation progress. Recently, with the explosive requirements of practical applications, a lot of research efforts have been dedicated to heterogeneous person re-identification (He-ReID). In this paper, we review the state-of-the-art methods comprehensively with respect to four main application scenarios – low-resolution, infrared, sketch and text.

We begin with a comparison between He-ReID and the general Homogeneous ReID (Ho-ReID) task. Then, we survey the models that have been widely employed in He-ReID. Available existing datasets for performing evaluation are briefly described. We then summarize and compare the representative approaches. Finally, we discuss some future research directions.

1. Introduction

As video surveillance systems are widely equipped in cities, exploiting videos to re-identify different persons, generate their trajectories across cameras and predict their behaviors plays an increasingly important role in daily investigation and anti-terrorism. More importantly, when a big criminal case happens, video investigation is almost the most crucial mean to find clues [5]. The impressive “2005 London subway bombings” and “2013 Boston marathon bombings” are uncovered with the help of surveillance videos [18]. Person re-identification (ReID) is now a very hot topic in computer vision and artificial intelligence community [35, 34, 24], which aims to find the same person across disjoint camera views. An effective and efficient ReID algorithm will alleviate painful video watching, and accelerate the investigation progress. General ReID technologies have obtained very high accuracy on public datasets in recent years due to the dramatic research progress [15, 39, 21], for example, 94.4\% rank 1 accuracy on the Market-1501 [33] dataset and 97.8\% on the CUHK03 [14] dataset, which outperform the ability of human [31]. However, the general ReID research has only been tested in a desired domain, which assumes that all person images are captured in the daytime with visible spectrum and these images have sufficient pixels to represent a person. We name this kind of person retrieval task as Homogeneous Person Re-identification (Ho-ReID).

In real applications, however, it is impractical to assume only working on such a desired domain, which only faces the challenges of intra-modality discrepancy, such as pose and viewpoint changes [32], background and illumination variations [10], and occlusions [36]. For example, when a criminal case happens, the investigators likely 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 sketches drawn by artists shall also be used as the cues for finding a specific person from visual surveil-
There are gaps between the desired domain (daytime, visible spectrum, and high-resolution) and other domains, including Low-Resolution (LR)/Infrared (IR)/Sketch images and Text description. General homogeneous ReID (Ho-ReID) has so far only focused on the challenges within the desired domain, while He-ReID needs to handle extra challenges across the desired and other collaborative domains.

Therefore, as shown in Figure 2, more efforts should be made on matching persons’ data across different domains/modalities, bridging the gap between different sensory devices (e.g., Visual light vs. Infrared devices), different camera specifications and settings (e.g., High-resolution vs. Low-resolution data), reproduction of human memory and direct recording by a camera (e.g., Sketches/Text Description vs. Digital images). We call this kind of matching as Heterogeneous Person Re-identification (He-ReID). In most cases, He-ReID involves querying a gallery consisting of high-resolution visible light person images using a probe from another domain/modality.

Encouragingly, in the past a few years, we have already seen quite a number of remarkable progresses on He-ReID (including our contributions), which are largely assisted by a growing variety of He-ReID benchmark datasets allowing direct comparison of different methodologies. While general Ho-ReID attracts too much attention in the research community, causing very intense competition and homogeneous tendency of research, 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 He-ReID by providing a quality survey for both brave pioneers and wondering newcomers.

This paper provides a comprehensive and up-to-date review of the diverse and growing array of He-ReID techniques. We categorize them into different modalities they operate across. As Figure 2 shows, we consider four cross-modality application scenarios: Low-resolution (LR), Infrared (IR), Sketch, and Text. Additionally, we introduce and organize the available datasets for performance evaluation in each category, summarize and compare the representative approaches, and close by drawing overall conclusions and making recommendations for potential research directions which are worth pursuing.

2. Heterogeneous vs. Homogeneous ReID

We make a brief comparison between the Ho-ReID and He-ReID, as Table 1 shows. For the re-identification media type, general Ho-ReID only exploits the desired RGB images, while He-ReID takes additional LR/IR/Sketch images and Text description into account. For the participant, the Ho-ReID technology only employs the resources from machine intelligence, while He-ReID also brings the input from human, such as the sketch image drawn by a painter or text description of the suspect by the witness. Hence, He-ReID is a more valuable task for practical video investigation applications.

For the main challenge, Ho-ReID only needs to deal with the intra-modality discrepancy, such as viewpoint variations, image misaligned or occlusion and illumination changes, while He-ReID should mainly bridge the gaps deriving from both intra- and inter-modality discrepancies. Since additional inter-modality discrepancies introduced, which are larger than intra-modality discrepancies [25], He-ReID is more challenging than Ho-ReID. As intra-modality and inter-modality discrepancies are essentially different, the methods designed for Ho-ReID cannot be directly used in He-ReID. Currently, the performances of Ho-ReID methods are high, while the performances of He-ReID methods are low. Hence, we should investigate how to eliminate the inter-modality discrepancy for the He-ReID task.

Figure 3 makes a comparison of state-of-the-art performances of Ho-ReID and He-ReID tasks. The results are on the typical datasets, i.e., the Market-1501 dataset [33] for Ho-ReID, the MLR-VIPER dataset [8] for He-ReID LR task, the SYSU-MM01 dataset [27] for He-ReID IR task, the Sketch Re-ID dataset [19] for He-ReID Sketch task, and the CUHK-PEDES dataset [13] for He-ReID Text task. We can find that all of the state-of-the-art performances of He-ReID tasks are still not so satisfied as that of the Ho-ReID task. He-ReID is a more challenging problem, compared with general Ho-ReID problem, since we need to fill the gap between the LR/IR/Sketch/Text domain and the desired image domain.
Table 1: Comparison between Ho-ReID and He-ReID. ‘#Publications’ represents for the number of publications related to Ho-ReID or He-ReID.

<table>
<thead>
<tr>
<th>Media Type</th>
<th>Desired image</th>
<th>Participant</th>
<th>Main challenge</th>
<th>Performance</th>
<th>#Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ho-ReID</td>
<td>+ LR/IR/Sketch/Text</td>
<td>Machine</td>
<td>Intra-modality</td>
<td>High</td>
<td>&gt; 800</td>
</tr>
<tr>
<td>He-ReID</td>
<td></td>
<td>Human &amp; Machine</td>
<td>Intra- + Inter-modality</td>
<td>Low</td>
<td>&lt; 50</td>
</tr>
</tbody>
</table>

Figure 3: The state-of-the-art performances of Ho-ReID and He-ReID. The compared methods are respectively AlignedReID [31] for the Ho-ReID task, SING [8] for the He-ReID LR task, D^2RL [25] for the He-ReID IR task, CDAFL [19] for the He-ReID Sketch task and CNN-LSTM [12] for the He-ReID Text task. There is a big performance gap between the LR/IR/Sketch/Text domain and the desired image domain. Note that we use the CMC-1 value to represent the performance. Since we would like to show the big performance gap between Ho-ReID and He-ReID tasks, so we do not make a comparison of different methods within a task.

However, compared with Ho-ReID, there are not so many publications related to He-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 He-ReID.

3. He-ReID Methods

3.1. He-ReID LR Application Scenario

It is common that the resolution of surveillance person image varies a lot, due to variations in the person-camera distance and camera deployment settings [26]. He-ReID LR application scenario attempts to compare images with different resolutions, where one is the normal resolution and the other is low-resolution. [9] was the first to investigate the problem of LR person re-identification. They proposed a semi-coupled low-rank discriminant dictionary learning method, constructing a mapping from the features of normal and LR person images. [16] designed a joint multi-scale learning framework by simultaneously learning metrics on image domains of different scales. [22] changed the problem to be low-resolution with different scales. They observed that scale-distance functions can be classified, and then learned a discriminating surface separating these functions in function space to identify persons. [8] developed a super-resolution and identity joint learning method to improve the He-ReID LR performance. [26] first cascaded multiple SR-GANs in series to promote the ability of scale-adaptive upscaling, then plugged-in a re-identification network to supplement the ability of image feature representation. Similar to [9], [11] designed a semi-coupled projective dictionary learning model to bridge the gap across different resolutions.

3.2. He-ReID IR Application Scenario

As we know, the visible cameras cannot capture valid appearance information under poor illumination environments (e.g. during the night), which limits the applicability in practical surveillance applications. Therefore, He-ReID IR application scenario provides a good supplement for nighttime surveillance applications [30]. [27] was the first to investigate the problem of IR person re-identification. They discussed and evaluated three common network structures for this task, including one-stream structure, two-stream structure and asymmetric FC layer structure. They also proposed deep zero-padding to help one-stream network and make the implicit network structure more suitable. [28] proposed a hierarchical cross-modality matching model by jointly optimizing the modality-specific and modality-shared metrics. [30] also utilized an end-to-end framework to learn discriminative feature representations, which is a dual-path network with a bi-directional dual-constrained top-ranking loss. [4] proposed to learn discriminative common representations. The network consists of a generator for learning image representations and as discriminator which tries to discriminate between RGB and IR image modalities. [25] presented a new framework, taking advantage of CycleGAN to reduce modality discrepancy from image-level and advantage of sophisticated Ho-ReID models to reduce appearance discrepancy from feature-level.

3.3. He-ReID Sketch Application Scenario

When a criminal is witnessed but not be photographed by a surveillance camera, He-ReID Sketch application scenario can automatically search all the surveillance videos to locate this criminal according to an artist’s drawing. Conse-
quently, police can cut down the number of suspects quickly and focus on those potential suspects [19]. [19] was the first and only one to investigate the problem of Sketch person re-identification. They proposed a cross-domain adversarial feature learning approach to jointly learn the identity features and domain-invariant features.

3.4. He-ReID Text Application Scenario

Searching person in a database with free-form natural language description has wide applications in video surveillance and activity analysis. To search possible criminals from large-scale videos automatically is in urgent need [13]. In the early years, some researches started to investigate the re-identification with attribute-based queries [5, 29, 23]. He-ReID Text application scenario denotes a situation that the queries are natural language descriptions. [13] evaluated and compared a wide range of possible models, and proposed a Recurrent Neural Network with Gated Neural Attention mechanism. [12] designed an identity-aware two-stage framework for this problem, which first learned to embed cross-modal features with a novel cross-modal loss and refined the matching results with a latent co-attention mechanism.

3.5. Summary

For He-ReID methods, we make 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 the deep learning methods are in the high-speed development period. The deep learning methods have their superior on shared feature learning and image generating.

- Different methods focused on different stages to fill the gaps between desired domain and other heterogeneous domains. Some of them [16, 27] try to learn a metric to fill the gap of features of different modalities. Some of them [9, 28] attempt to learn a shared features. Some of them [25, 26] unify the modality in the data level. In Section 5, we will make an analysis and discussion for the methods.

- The existing researches in each application scenario still have much limitations. For the He-ReID LR application scenario, recent researches are mainly evaluated on the simulated datasets. For the He-ReID Sketch application scenario, only one research touched this direction, due to the limitation of dataset. The researches on He-ReID Text application scenario are also not so many, because only one dataset is constructed.

4. Available Datasets

We list available datasets for He-ReID in Table 2, including the application scenario, the data construction type, the total number of cameras, identities and samples. To make a comparison, we also list some typical Ho-ReID datasets.

Table 2: Released and freely available datasets for Ho-ReID and He-ReID. The notation ‘Des.’ indicates that the dataset is constructed by desired RGB images. The ‘Sim.’ indicates that the dataset is constructed by simulated images. ‘Appl.’ indicates the application scenario the dataset related to. ‘#Cam.’, ‘#ID’ and ‘#Sam.’ respectively represent the total number of cameras, identities and samples.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Appl. Type</th>
<th>#Cam.</th>
<th>#ID</th>
<th>#Sam.</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIPeR [6]</td>
<td>Des.</td>
<td>2</td>
<td>632</td>
<td>1,264</td>
</tr>
<tr>
<td>3DPES [1]</td>
<td>Des.</td>
<td>8</td>
<td>192</td>
<td>1,011</td>
</tr>
<tr>
<td>i-LIDS [37]</td>
<td>Des.</td>
<td>2</td>
<td>119</td>
<td>238</td>
</tr>
<tr>
<td>PRID 2011 [7]</td>
<td>Des.</td>
<td>2</td>
<td>934</td>
<td>1,134</td>
</tr>
<tr>
<td>PRID 450S [20]</td>
<td>Des.</td>
<td>2</td>
<td>450</td>
<td>900</td>
</tr>
<tr>
<td>CUHK03 [14]</td>
<td>Des.</td>
<td>2</td>
<td>1,360</td>
<td>13,164</td>
</tr>
<tr>
<td>Market-1501 [33]</td>
<td>Des.</td>
<td>6</td>
<td>1,501</td>
<td>32,668</td>
</tr>
<tr>
<td>DukeMTMC-reID [38]</td>
<td>Des.</td>
<td>8</td>
<td>1,404</td>
<td>36,411</td>
</tr>
</tbody>
</table>

4.1. Ho-ReID Datasets

The VIPeR [6] dataset contains 1,264 outdoor images obtained from two views of 632 persons. Generally, the set of 632 image pairs are randomly split into two sets of 316 image pairs each, one for training and another for testing. The 3DPES [1] dataset includes 1,011 images of 192 individuals captured from 8 outdoor cameras with signif-
icantly different viewpoints. In this dataset each person has 2 to 26 images. The i-LIDS [37] dataset consists of 119 persons with a total of 476 shots captured by multiple non-overlapping cameras with an average of four images for each person. The PRID 2011 [7] dataset consists of person images recorded from two different cameras. Camera A contains 385 persons and camera B contains 749 persons, with 200 persons appearing in both views. The PRID 450S [20] dataset contains 450 single-shot image pairs captured over two spatially disjoint camera views. All images are normalized to 168x80 pixels. The CUHK03 [14] dataset contains 14,096 images of 1,467 identities. Each identity is captured from two cameras in the CUHK campus, and has an average of 4.8 images in each camera. Both experimental results on ‘labeled’ and ‘detected’ data are presented. The Market-1501 [33] dataset is currently one of the largest benchmark dataset for Ho-ReID. It contains 32,668 labeled bounding boxes of 1,501 identities captured from 6 different cameras. The DukeMTMC-reID [38] dataset contains 1812 identities captured by 8 cameras. A number of 1404 identities appear in more than two cameras. The rest 408 are distractor images.

4.2. He-ReID Datasets

- **He-ReID LR datasets:** The CAVIAR [3] dataset contains images of 72 individuals captured from 2 cameras in a shopping mall. The resolution of images captured from one camera is much lower than that in the other one. 22 people are removed, since they are only captured in a single camera with no low-resolution images. The remaining data include 1,000 images of 50 people, with 10 normal resolution images and 10 LR images per person. The LR-VIPeR and tbLR-3DPES [16] datasets are respectively simulated from the VIPeR and 3DPES datasets. For both of the datasets, half of all the images of each person from both datasets are selected and replaced with the LR images which were subsampled to a quarter of their original image size. The LR-i-LIDS and LR-PRID [9] datasets are respectively simulated from i-LIDS and PRID 2011 datasets. For both of the datasets, one image from each person as the HR image set, and select another image from each person and perform down-sampling and smoothing operations to generate LR image set. The sampling rate is 1/8. The SALR-VIPeR and SALR-PRID [22] datasets are respectively simulated from VIPeR and PRID 450S datasets. For both of the datasets, images from one camera are set as the HR probe set, whose resolution remains unchanged. While images from the other camera are set as the LR gallery set, which are down-sampled randomly to different scales. The scale ratios range from 0.1 to 0.25. The MLR-VIPeR, MLR-CUHK03 and MLR-SYSU [8] datasets are respectively simulated from VIPeR, CUHK03, and SYSU datasets. For all of the datasets, SYSU [2] has totally 24,446 images of 502 people captured by two cameras. Three images per person per camera are selected and created LR images. Images from one camera view are down-sampled by a ratio randomly picked from \{1/2, 1/3, 1/4\} whilst images of another view are remaining the same.

- **He-ReID IR datasets:** The SYSU-MM01 [27] dataset contains images captured by 6 cameras, including two IR cameras and four RGB ones. Some cameras are in the indoor environments and some are in outdoor environments. Note that all the IR cameras are the active IR cameras. The training set contains 395 persons, with 22258 visible images and 11909 thermal images. The testing set contains 96 persons, with 3803 thermal images for query and 301 randomly selected visible images as gallery set. The RegDB [17] dataset is collected by two cameras, which contains 412 persons. Note that the IR camera is the passive one. For each person, 10 different visible light images are captured by a visible camera, and 10 different thermal images are obtained by a thermal camera.

- **He-ReID Sketch dataset:** The Sketch Re-ID [19] dataset contains 200 persons, each of which has one sketch and two photos. A total of 5 artists draw all persons’ sketches and every artist has his own painting style. Overall, it has 150 persons for training and 50 persons for testing.

- **He-ReID Text dataset:** The CUHK-PEDES [13] dataset contains 40,206 images of 13,003 person identities. Each image is described by two sentences. There are 11,003 persons, 34,054 images and 68,108 sentence descriptions in the training set. The validation set and test set consist of 3,078 and 3,074 images, respectively, and both of them contain 1,000 persons.

4.3. Summary

For He-ReID datasets, we make following summaries:

- Although many He-ReID LR datasets were constructed, most of them are simulated w.r.t. from general Ho-ReID datasets.

- Only two He-ReID IR datasets exist. SYSU-MM01 [27] is an active IR dataset, while RegDB [27] is passive one. Only one He-ReID Sketch dataset and one He-ReID Text dataset are constructed.

- Compared with Ho-ReID, there is still a lot of room to construct more available datasets. For He-ReID LR, it requires to construct practical dataset 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.

5. Discussion

Representative methods employed in He-ReID are listed in Table 3. For easy comparison, we list most of the
Figure 4: **Summary of three pipelines.** (a) just utilizes the metric learning method to learn how to match representations from separate representation models. (b) focuses on learning the common representation models of different modalities. (c) pays attention to generating the modality-unified samples and then learning the common representation models. In the figure, ‘D Domain’ stands for the samples from the desired domain. ‘H Domain’ stands for the samples from the heterogeneous domain.

Table 3: **Representative methods employed in He-ReID.** The notation ‘Conf.’ indicates the conference that the corresponding paper was published in. ‘Appl.’ indicates the application scenario the paper studies in. ‘Sep.RL+ML’, ‘Shared RL’ and ‘MU+SharedRL’ respectively represent for the pipelines the corresponding utilized. ‘Sep.RL+ML’ stands for separate representation learning and metric learning (pipeline (a)). ‘Shared RL’ stands for shared representation learning (pipeline (b)). ‘MU+SharedRL’ stands for modality unifying and shared representation learning (pipeline (c)). ‘Datasets’ indicates the evaluated datasets in the paper, where the number denotes for the certain dataset in Table 2.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Appl.</th>
<th>Year</th>
<th>Conf.</th>
<th>Method</th>
<th>Datasets</th>
<th>Sep.RL+ML</th>
<th>Shared RL</th>
<th>MU+SharedRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li et al. [16]</td>
<td>LR</td>
<td>2015</td>
<td>ICCV</td>
<td>Metric Learning</td>
<td>1 2 3</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wang et al. [22]</td>
<td>LR</td>
<td>2016</td>
<td>IJCAI</td>
<td>Subspace Learning</td>
<td>1 6 7</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jiao et al. [8]</td>
<td>LR</td>
<td>2018</td>
<td>AAAI</td>
<td>Super Resolution</td>
<td>1 8 9 10</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wang et al. [26]</td>
<td>LR</td>
<td>2018</td>
<td>IJCAI</td>
<td>Super Resolution</td>
<td>1 6 7</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wu et al. [27]</td>
<td>IR</td>
<td>2017</td>
<td>ICCV</td>
<td>Deep Zero-Padding</td>
<td>11</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ye et al. [28]</td>
<td>IR</td>
<td>2018</td>
<td>AAAI</td>
<td>Metric Learning</td>
<td>12</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ye et al. [30]</td>
<td>IR</td>
<td>2018</td>
<td>IJCAI</td>
<td>Feature Learning</td>
<td>11 12</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wang et al. [25]</td>
<td>IR</td>
<td>2019</td>
<td>CVPR</td>
<td>Image Generation</td>
<td>11 12</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pang et al. [19]</td>
<td>Sketch</td>
<td>2018</td>
<td>ACM MM</td>
<td>Feature Learning</td>
<td>13</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al. [13]</td>
<td>Text</td>
<td>2017</td>
<td>CVPR</td>
<td>Affinity Learning</td>
<td>14</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Li et al. [12]</td>
<td>Text</td>
<td>2017</td>
<td>ICCV</td>
<td>Feature Learning</td>
<td>14</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The above information of the works, including the published paper, year, conference, application scenario, method and strategies. We can find that all of the representative works are published in recent five years. It means that He-ReID is a relatively new research topic.

Among all the application scenarios, He-ReID Sketch and Text are less investigated. As described in the sections above, we consider that if more datasets are available, more works will be published.

In essence, He-ReID is a cross-modality task. Most of the methods can be categorized into three pipelines, as Figure 4 shows. 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. Pipeline (b) focuses on learning shared feature models of different modalities, and training data are from both modalities. Pipeline (c) pays attention to generating the unified-modality samples, for example, using the super-resolution method to generate high-resolution images from low-resolution images, or using the image generation method to generate infrared images from RGB images and RGB images from infrared images.

We analysis all listed representative methods and contribute the methods to these three pipelines. It also shows in Table 3. For each kind of application scenario, from top to bottom, the methods become more effective. We can find that pipeline (c) performs better than pipeline (b), and better than pipeline (a). Hence, unifying the modality of samples is the most effective way to fill the modality gap. For the He-ReID LR application scenario, the super-resolution methods can be used, such as [26]. Thanks to the developments of GAN, unifying the modalities is considerable as [25].

6. Conclusion and Future Directions

This paper attempted to provide an overview of recent developments of Heterogeneous Person Re-identification (He-ReID). It can cover most of the literature on He-ReID. We summarized the widely employed methods, released datasets, and compared the existing methods. We believe that He-ReID will continue to be an active and promising research area with broad potential applications. Many issues in He-ReID, however, are still open.

Dataset Construction. For He-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 He-ReID datasets are also required. It is somehow urgent for us to built new datasets, and push forward this research area.

Human Interaction and Crowd Sourcing. For He-ReID Sketch and Text application scenarios, human intelligence join into the process of re-identification. Human intelligence is sometime subjective and incomplete. So we should consider how to mine and integrate useful information to help search the target in surveillance system. On the hand, a lot of witnesses will provide their cues. Each person may have a different view, so the crowd sourcing cues may be diverse to each other. We should to design strategy to remove conflict and filter valuable information.

Investigation on unifying the modality. For He-ReID LR application scenario, some work [26] attempted to use super-resolution methods to unify the modality, where they transfer the heterogeneous domain LR image to desired domain. For He-ReID IR application scenario, with the help of CycleGAN, one work [25] not only transferred the image from heterogeneous (IR) domain to desired (RGB) domain, but also transferred the image from desired domain to heterogeneous domain. However, for the He-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 domain as Figure 5 shows, for example, a middle-level resolution for the He-ReID LR application scenario and a hyperspectral image for the He-ReID IR application scenario.

Integrating multiple He-ReID application scenarios. For a practical system, it may require different kinds of inputs to search out the target. The He-ReID application scenarios can be equipped in different stages. If we can integrate different kinds of inputs together, 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-domain research task.

References


[29] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, Jun Chen, and Jun Liu. Specific person retrieval via incomplete


