Deep Learning for Person Re-identification: A Survey and Outlook

Article *in* IEEE Transactions on Pattern Analysis and Machine Intelligence · January 2021

DOI: 10.1109/TPAMI.2021.3054775

Mang Ye
Wuhan University
50 PUBLICATIONS **1,816** CITATIONS
SEE PROFILE

Jianbing Shen
Beijing Institute of Technology
262 PUBLICATIONS **9,863** CITATIONS
SEE PROFILE

Steven C. H. Hoi
Nanyang Technological University
274 PUBLICATIONS **9,928** CITATIONS
SEE PROFILE

Deep Learning for Person Re-identification: A Survey and Outlook

Mang Ye, Jianbing Shen, *Senior Member, IEEE*, Gaojie Lin, Tao Xiang
Ling Shao and Steven C. H. Hoi, *Fellow, IEEE*

Abstract—Person re-identification (Re-ID) aims at retrieving a person of interest across multiple non-overlapping cameras. With the advancement of deep neural networks and increasing demand of intelligent video surveillance, it has gained significantly increased interest in the computer vision community. By dissecting the involved components in developing a person Re-ID system, we categorize it into the closed-world and open-world settings. The widely studied closed-world setting is usually applied under various research-oriented assumptions, and has achieved inspiring success using deep learning techniques on a number of datasets. We first conduct a comprehensive overview with in-depth analysis for closed-world person Re-ID from three different perspectives, including deep feature representation learning, deep metric learning and ranking optimization. With the performance saturation under closed-world setting, the research focus for person Re-ID has recently shifted to the open-world setting, facing more challenging issues. This setting is closer to practical applications under specific scenarios. We summarize the open-world Re-ID in terms of five different aspects. By analyzing the advantages of existing methods, we design a powerful AGW baseline, achieving state-of-the-art or at least comparable performance on twelve datasets for FOUR different Re-ID tasks. Meanwhile, we introduce a new evaluation metric (mINP) for person Re-ID, indicating the cost for finding all the correct matches, which provides an additional criteria to evaluate the Re-ID system for real applications. Finally, some important yet under-investigated open issues are discussed.

Index Terms—Person Re-Identification, Pedestrian Retrieval, Literature Survey, Evaluation Metric, Deep Learning

1 Introduction

Person re-identification (Re-ID) has been widely studied as a specific person retrieval problem across non-overlapping cameras [1], [2]. Given a query person-of-interest, the goal of Re-ID is to determine whether this person has appeared in another place at a distinct time captured by a different camera, or even the same camera at a different time instant [3]. The query person can be represented by an image [4], [5], [6], a video sequence [7], [8], and even a text description [9], [10]. Due to the urgent demand of public safety and increasing number of surveillance cameras, person Re-ID is imperative in intelligent surveillance systems with significant research impact and practical importance.

Re-ID is a challenging task due to the presence of different viewpoints [11], [12], varying low-image resolutions [13], [14], illumination changes [15], unconstrained poses [16], [17], [18], occlusions [19], [20], heterogeneous modalities [10], [21], complex camera environments, background clutter [22], unreliable bounding box generations, etc. These result in varying variations and uncertainty. In addition, for practical model deployment, the dynamic updated camera network [23], [24], large scale gallery with efficient retrieval [25], group uncertainty [26], significant domain shift [27], unseen testing scenarios [28], incremental model updating [29] and changing cloths [30] also greatly increase the difficulties. These challenges lead that Re-ID is still unsolved problem. Early research efforts mainly focus on the hand-crafted feature construction with body structures [31], [32], [33], [34], [35] or distance metric learning [36], [37], [38], [39], [40], [41]. With the advancement of deep learning, person Re-ID has achieved inspiring performance on the widely used benchmarks [5], [42], [43], [44]. However, there is still a large gap between the research-oriented scenarios and practical applications [45]. This motivates us to conduct a comprehensive survey, develop a powerful baseline for different Re-ID tasks and discuss several future directions.

Though some surveys have also summarized the deep learning techniques [2], [46], [47], our survey makes three major differences: 1) We provide an in-depth and comprehensive analysis of existing deep learning methods by discussing their advantages and limitations, analyzing the state-of-the-arts. This provides insights for future algorithm design and new topic exploration. 2) We design a new powerful baseline (AGW: Attention Generalized mean pooling with Weighted triplet loss) and a new evaluation metric (mINP: mean Inverse Negative Penalty) for future developments. AGW achieves state-of-the-art performance on twelve datasets for four different Re-ID tasks. mINP provides a supplement metric to existing CMC/mAP, indicating the cost to find all the correct matches. 3) We make an attempt to discuss several important research directions with under-investigated open issues to narrow the gap between the closed-world and open-world applications, taking a step towards real-world Re-ID system design.

- *M. Ye is with the School of Computer Science, Wuhan University, China and Inception Institute of Artificial Intelligence, UAE.*
- *J. Shen and L. Shao are with the Inception Institute of Artificial Intelligence, UAE. E-mail:{mangye16, shenjianbingcg}@gmail.com*
- *G. Lin is with the School of Computer Science, Beijing Institute of Technology, China.*
- *T. Xiang is with the Centre for Vision Speech and Signal Processing, University of Surrey, UK. Email: t.xiang@surrey.ac.uk*
- *S. C. H. Hoi is with the Singapore Management University, and Salesforce Research Asia, Singapore. Email: stevenhoi@gmail.com*

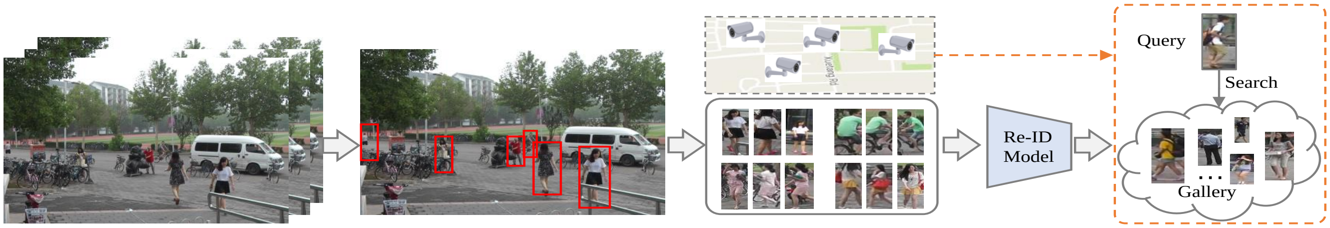


Fig. 1: The flow of designing a practical person Re-ID system, including five main steps: 1) *Raw Data Collection*, (2) *Bounding Box Generation*, 3) *Training Data Annotation*, 4) *Model Training* and 5) *Pedestrian Retrieval*.

TABLE 1: Closed-world *vs*. Open-world Person Re-ID.

Closed-world (Section 2)	Open-world (Section 3)
✓ Single-modality Data	Heterogeneous Data (§ 3.1)
✓ Bounding Boxes Generation	Raw Images/Videos (§ 3.2)
✓ Sufficient Annotated Data	Unavailable/Limited Labels (§ 3.3)
✓ Correct Annotation	Noisy Annotation (§ 3.4)
✓ Query Exists in Gallery	Open-set (§ 3.5)

Unless otherwise specified, person Re-ID in this survey refers to the pedestrian retrieval problem across multiple surveillance cameras, from a computer vision perspective. Generally, building a person Re-ID system for a specific scenario requires five main steps (as shown in Fig. 1):

1) Step 1: *Raw Data Collection*: Obtaining raw video data from surveillance cameras is the primary requirement of practical video investigation. These cameras are usually located in different places under varying environments [48]. Most likely, this raw data contains a large amount of complex and noisy background clutter.
2) Step 2: *Bounding Box Generation*: Extracting the bounding boxes which contain the person images from the raw video data. Generally, it is impossible to manually crop all the person images in large-scale applications. The bounding boxes are usually obtained by the person detection [49], [50] or tracking algorithms [51], [52].
3) Step 3: *Training Data Annotation*: Annotating the cross-camera labels. Training data annotation is usually indispensable for discriminative Re-ID model learning due to the large cross-camera variations. In the existence of large domain shift [53], we often need to annotate the training data in every new scenario.
4) Step 4: *Model Training*: Training a discriminative and robust Re-ID model with the previous annotated person images/videos. This step is the core for developing a Re-ID system and it is also the most widely studied paradigm in the literature. Extensive models have been developed to handle the various challenges, concentrating on feature representation learning [54], [55], distance metric learning [56], [57] or their combinations.
5) Step 5: *Pedestrian Retrieval*: The testing phase conducts the pedestrian retrieval. Given a person-of-interest (query) and a gallery set, we extract the feature representations using the Re-ID model learned in previous stage. A retrieved ranking list is obtained by sorting the calculated query-to-gallery similarity. Some methods have also investigated the ranking optimization to improve the retrieval performance [58], [59].

According to the five steps mentioned above, we categorize existing Re-ID methods into two main trends: *closed-world* and *open-world* settings, as summarized in Table 1. A step-by-step comparison is in the following five aspects:

1) *Single-modality* vs. *Heterogeneous Data*: For the raw data collection in Step 1, all the persons are represented by images/videos captured by single-modality visible cameras in the closed-world setting [5], [8], [31], [42], [43], [44]. However, in practical open-world applications, we might also need to process heterogeneous data, which are infrared images [21], [60], sketches [61], depth images [62], or even text descriptions [63]. This motivates the heterogeneous Re-ID in § 3.1.
2) *Bounding Box Generation* vs. *Raw Images/Videos* : For the bounding box generation in Step 2, the closed-world person Re-ID usually performs the training and testing based on the generated bounding boxes, where the bounding boxes mainly contain the person appearance information. In contrast, some practical open-world applications require end-to-end person search from the raw images or videos [55], [64]. This leads to another open-world topic, *i.e.*, end-to-end person search in § 3.2.
3) *Sufficient Annotated Data* vs. *Unavailable/Limited Labels*: For the training data annotation in Step 3, the closed-world person Re-ID usually assumes that we have enough annotated training data for supervised Re-ID model training. However, label annotation for each camera pair in every new environment is time consuming and labor intensive, incurring high costs. In open-world scenarios, we might not have enough annotated data (*i.e.*, limited labels) [65] or even without any label information [66]. This inspires the discussion of the unsupervised and semi-supervised Re-ID in § 3.3.
4) *Correct Annotation* vs. *Noisy Annotation*: For Step 4, existing closed-world person Re-ID systems usually assume that all the annotations are correct, with clean labels. However, annotation noise is usually unavoidable due to annotation error (*i.e.*, label noise) or imperfect detection/tracking results (*i.e.*, sample noise, partial Re-ID [67]). This leads to the analysis of noise-robust person Re-ID under different noise types in § 3.4.
5) *Query Exists in Gallery* vs. *Open-set*: In the pedestrian retrieval stage (Step 5), most existing closed-world person Re-ID works assume that the query must occur in the gallery set by calculating the CMC [68] and mAP [5]. However, in many scenarios, the query person may not appear in the gallery set [69], [70], or we need to perform the verification rather than retrieval [26]. This brings us to the open-set person Re-ID in § 3.5.

This survey first introduces the widely studied person Re-ID under closed-world settings in § 2. A detailed review on the datasets and the state-of-the-arts are conducted in § 2.4. We then introduce the open-world person Re-ID in § 3. An outlook for future Re-ID is presented in § 4, including a new evaluation metric (§ 4.1), a new powerful AGW baseline (§ 4.2). We discuss several under-investigated open issues for future study (§ 4.3). Conclusions will be drawn in § 5. A structure overview is shown in the supplementary.

2 Closed-world Person Re-Identification

This section provides an overview for closed-world person Re-ID. As discussed in § 1, this setting usually has the following assumptions: 1) person appearances are captured by single-modality visible cameras, either by image or video; 2) The persons are represented by bounding boxes, where most of the bounding box area belongs the same identity; 3) The training has enough annotated training data for supervised discriminative Re-ID model learning; 4) The annotations are generally correct; 5) The query person must appear in the gallery set. Typically, a standard closed-world Re-ID system contains three main components: *Feature Representation Learning* (§ 2.1), which focuses on developing the feature construction strategies; *Deep Metric Learning* (§ 2.2), which aims at designing the training objectives with different loss functions or sampling strategies; and *Ranking Optimization* (§ 2.3), which concentrates on optimizing the retrieved ranking list. An overview of the datasets and state-of-the-arts with in-depth analysis is provided in § 2.4.2.

2.1 Feature Representation Learning

We firstly discuss the feature learning strategies in closed-world person Re-ID. There are four main categories (as shown in Fig. 2): a) Global Feature (§ 2.1.1), it extracts a global feature representation vector for each person image without additional annotation cues [55]; b) Local Feature (§ 2.1.2), it aggregates part-level local features to formulate a combined representation for each person image [75], [76], [77]; c) Auxiliary Feature (§ 2.1.3), it improves the feature representation learning using auxiliary information, *e.g.*, attributes [71], [72], [78], GAN generated images [42], etc. d) Video Feature (§ 2.1.4), it learns video representation for video-based Re-ID [7] using multiple image frames and temporal information [73], [74]. We also review several specific architecture designs for person Re-ID in § 2.1.5.

2.1.1 Global Feature Representation Learning

Global feature representation learning extracts a global feature vector for each person image, as shown in Fig. 2(a). Since deep neural networks are originally applied in image classification [79], [80], global feature learning is the primary choice when integrating advanced deep learning techniques into the person Re-ID field in early years.

To capture the fine-grained cues in global feature learning, A joint learning framework consisting of a single-image representation (SIR) and cross-image representation (CIR) is developed in [81], trained with triplet loss using specific sub-networks. The widely-used ID-discriminative Embedding (IDE) model [55] constructs the training process as a multi-class classification problem by treating each identity as a distinct class. It is now widely used in Re-ID community [42], [58], [77], [82], [83]. Qian *et al.* [84] develop a multi-scale deep representation learning model to capture discriminative cues at different scales.

Attention Information. Attention schemes have been widely studied in literature to enhance representation learning [85]. 1) *Group 1: Attention within the person image.* Typical strategies include the pixel level attention [86] and the channel-wise feature response re-weighting [86], [87], [88], [89], or background suppressing [22]. The spatial information is integrated in [90]. 2) *Group 2: attention across multiple person images.* A context-aware attentive feature learning method is proposed in [91], incorporating both an intra-sequence and inter-sequence attention for pair-wise feature alignment and refinement. The attention consistency property is added in [92], [93]. Group similarity [94], [95] is another popular approach to leverage the cross-image attention, which involves multiple images for local and global similarity modeling. The first group mainly enhances the robustness against misalignment/imperfect detection, and the second improves the feature learning by mining the relations across multiple images.

2.1.2 Local Feature Representation Learning

It learns part/region aggregated features, making it robust against misalignment [77], [96]. The body parts are either automatically generated by human parsing/pose estimation (Group 1) or roughly horizontal division (Group 2).

With automatic body part detection, the popular solution is to combine the full body representation and local part features [97], [98]. Specifically, the multi-channel aggregation [99], multi-scale context-aware convolutions [100], multi-stage feature decomposition [17] and bilinear-pooling [97] are designed to improve the local feature learning. Rather than feature level fusion, the part-level similarity combination is also studied in [98]. Another popular solution is to enhance the robustness against background clutter, using the pose-driven matching [101], pose-guided part attention module [102], semantically part alignment [103], [104].

For horizontal-divided region features, multiple part-level classifiers are learned in Part-based Convolutional Baseline (PCB) [77], which now serves as a strong part feature learning baseline in the current state-of-the-art [28], [105], [106]. To capture the relations across multiple body parts, the Siamese Long Short-Term Memory (LSTM) architecture [96], second-order non-local attention [107], Interaction-and-Aggregation (IA) [108] are designed to reinforce the feature learning.

The first group uses human parsing techniques to obtain semantically meaningful body parts, which provides well-align part features. However, they require an additional pose detector and are prone to noisy pose detections [77]. The second group uses a uniform partition to obtain the horizontal stripe parts, which is more flexible, but it is sensitive to heavy occlusions and large background clutter.

2.1.3 Auxiliary Feature Representation Learning

Auxiliary feature representation learning usually requires additional annotated information (*e.g.*, semantic attributes [71]) or generated/augmented training samples to reinforce the feature representation [19], [42].

Semantic Attributes. A joint identity and attribute learning baseline is introduced in [72]. Su *et al.* [71] propose a deep attribute learning framework by incorporating the predicted semantic attribute information, enhancing the

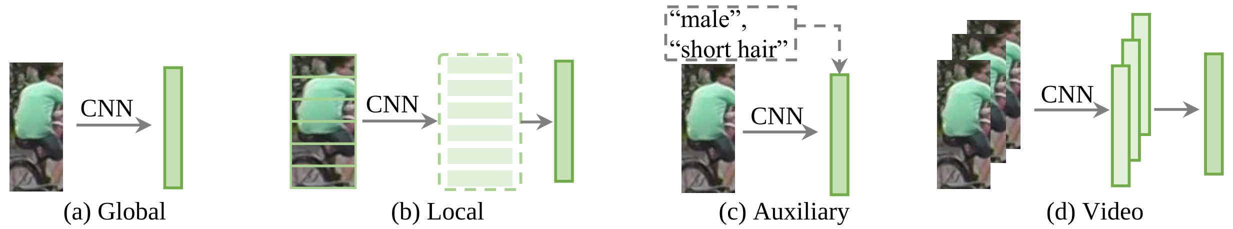


Fig. 2: Four different feature learning strategies. a) Global Feature, learning a global representation for each person image in § 2.1.1; b) Local Feature, learning part-aggregated local features in § 2.1.2; c) Auxiliary Feature, learning the feature representation using auxiliary information, *e.g.*, attributes [71], [72] in § 2.1.3 and d) Video Feature , learning the video representation using multiple image frames and temporal information [73], [74] in § 2.1.4.

generalizability and robustness of the feature representation in a semi-supervised learning manner. Both the semantic attributes and the attention scheme are incorporated to improve part feature learning [109]. Semantic attributes are also adopted in [110] for video Re-ID feature representation learning. They are also leveraged as the auxiliary supervision information in unsupervised learning [111].

Viewpoint Information. The viewpoint information is also leveraged to enhance the feature representation learning [112], [113]. Multi-Level Factorisation Net (MLFN) [112] also tries to learn the identity-discriminative and view-invariant feature representations at multiple semantic levels. Liu *et al.* [113] extract a combination of view-generic and view-specific learning. An angular regularization is incorporated in [114] in the viewpoint-aware feature learning.

Domain Information. A Domain Guided Dropout (DGD) algorithm [54] is designed to adaptively mine the domain-sharable and domain-specific neurons for multi-domain deep feature representation learning. Treating each camera as a distinct domain, Lin *et al.* [115] propose a multi-camera consistent matching constraint to obtain a globally optimal representation in a deep learning framework. Similarly, the camera view information or the detected camera location is also applied in [18] to improve the feature representation with camera-specific information modeling.

GAN Generation. This section discusses the use of GAN generated images as the auxiliary information. Zheng *et al.* [42] start the first attempt to apply the GAN technique for person Re-ID. It improves the supervised feature representation learning with the generated person images. Pose constraints are incorporated in [116] to improve the quality of the generated person images, generating the person images with new pose variants. A pose-normalized image generation approach is designed in [117], which enhances the robustness against pose variations. Camera style information [118] is also integrated in the image generation process to address the cross camera variations. A joint discriminative and generative learning model [119] separately learns the appearance and structure codes to improve the image generation quality. Using the GAN generated images is also a widely used approach in unsupervised domain adaptation Re-ID [120], [121], approximating the target distribution.

Data Augmentation. For Re-ID, custom operations are random resize, cropping and horizontal flip [122]. Besides, adversarially occluded samples [19] are generated to augment the variation of training data. A similar random erasing strategy is proposed in [123], adding random noise to the input images. A batch DropBlock [124] randomly drops a region block in the feature map to reinforce the attentive feature learning. Bak *et al.* [125] generate the virtual humans rendered under different illumination conditions. These methods enrich the supervision with the augmented samples, improving the generalizability on the testing set.

2.1.4 Video Feature Representation Learning

Video-based Re-ID is another popular topic [126], where each person is represented by a video sequence with multiple frames. Due to the rich appearance and temporal information, it has gained increasing interest in the Re-ID community. This also brings in additional challenges in video feature representation learning with multiple images.

The primary challenge is to accurately capture the temporal information. A recurrent neural network architecture is designed for video-based person Re-ID [127], which jointly optimizes the final recurrent layer for temporal information propagation and the temporal pooling layer. A weighted scheme for spatial and temporal streams is developed in [128]. Yan *et al.* [129] present a progressive/sequential fusion framework to aggregate the frame-level human region representations. Semantic attributes are also adopted in [110] for video Re-ID with feature disentangling and frame re-weighting. Jointly aggregating the frame-level feature and spatio-temporal appearance information is crucial for video representation learning [130], [131], [132].

Another major challenge is the unavoidable outlier tracking frames within the videos. Informative frames are selected in a joint Spatial and Temporal Attention Pooling Network (ASTPN) [131], and the contextual information is integrated in [130]. A co-segmentation inspired attention model [132] detects salient features across multiple video frames with mutual consensus estimation. A diversity regularization [133] is employed to mine multiple discriminative body parts in each video sequence. An affine hull is adopted to handle the outlier frames within the video sequence [83]. An interesting work [20] utilizes the multiple video frames to auto-complete occluded regions. These works demonstrate that handling the noisy frames can greatly improve the video representation learning.

It is also challenging to handle the varying lengths of video sequences, Chen *et al.* [134] divide the long video sequences into multiple short snippets, aggregating the top-ranked snippets to learn a compact embedding. A clip-level learning strategy [135] exploits both spatial and temporal dimensional attention cues to produce a robust clip-level representation. Both the short- and long-term relations [136] are integrated in a self-attention scheme.

2.1.5 Architecture Design

Framing person Re-ID as a specific pedestrian retrieval problem, most existing works adopt the network architectures [79], [80] designed for image classification as the backbone. Some works have tried to modify the backbone architecture to achieve better Re-ID features. For the widely used ResNet50 backbone [80], the important modifications include changing the last convolutional stripe/size to 1 [77], employing adaptive average pooling in the last pooling layer [77], and adding bottleneck layer with batch normalization after the pooling layer [82].

Accuracy is the major concern for specific Re-ID network architecture design to improve the accuracy, Li *et al.* [43] start the first attempt by designing a filter pairing neural network (FPNN), which jointly handles misalignment and occlusions with part discriminative information mining. Wang *et al.* [89] propose a BraidNet with a specially designed WConv layer and Channel Scaling layer. The WConv layer extracts the difference information of two images to enhance the robustness against misalignments and Channel Scaling layer optimizes the scaling factor of each input channel. A Multi-Level Factorisation Net (MLFN) [112] contains multiple stacked blocks to model various latent factors at a specific level, and the factors are dynamically selected to formulate the final representation. An efficient fully convolutional Siamese network [137] with convolution similarity module is developed to optimize multi-level similarity measurement. The similarity is efficiently captured and optimized by using the depth-wise convolution.

Efficiency is another important factor for Re-ID architecture design. An efficient small scale network, namely Omni-Scale Network (OSNet) [138], is designed by incorporating the point-wise and depth-wise convolutions. To achieve multi-scale feature learning, a residual block composed of multiple convolutional streams is introduced.

With the increasing interest in auto-machine learning, an Auto-ReID [139] model is proposed. Auto-ReID provides an efficient and effective automated neural architecture design based on a set of basic architecture components, using a part-aware module to capture the discriminative local Re-ID features. This provides a potential research direction in exploring powerful domain-specific architectures.

2.2 Deep Metric Learning

Metric learning has been extensively studied before the deep learning era by learning a Mahalanobis distance function [36], [37] or projection matrix [40]. The role of metric learning has been replaced by the loss function designs to guide the feature representation learning. We will first review the widely used loss functions in § 2.2.1 and then summarize the training strategies with specific sampling designs § 2.2.2.

2.2.1 Loss Function Design

This survey only focuses on the loss functions designed for deep learning [56]. An overview of the distance metric learning designed for hand-crafted systems can be found in [2], [143]. There are three widely studied loss functions with their variants in the literature for person Re-ID, including the identity loss, verification loss and triplet loss. An illustration of three loss functions is shown in Fig. 3.

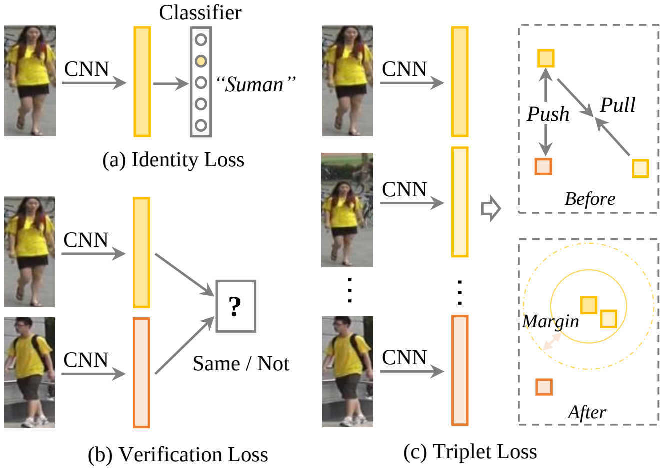


Fig. 3: Three kinds of widely used loss functions in the literature. (a) Identity Loss [42], [82], [118], [140] ; (b) Verification Loss [94], [141] and (c) Triplet Loss [14], [22], [57]. Many works employ their combinations [87], [137], [141], [142].

Identity Loss. It treats the training process of person Re-ID as an image classification problem [55], *i.e.*, each identity is a distinct class. In the testing phase, the output of the pooling layer or embedding layer is adopted as the feature extractor. Given an input image x_i with label y_i, the predicted probability of x_i being recognized as class y_i is encoded with a softmax function, represented by $p(y_i|x_i)$. The identity loss is then computed by the cross-entropy

$$\mathcal{L}_{id} = -\frac{1}{n}\sum\nolimits_{i=1}^{n} \log(p(y_i|x_i)), \tag{1}$$

where n represents the number of training samples within each batch. The identity loss has been widely used in existing methods [19], [42], [82], [92], [95], [106], [118], [120], [140], [144]. Generally, it is easy to train and automatically mine the hard samples during the training process, as demonstrated in [145]. Several works have also investigated the softmax variants [146], such as the sphere loss in [147] and AM softmax in [95]. Another simple yet effective strategy, *i.e.*, label smoothing [42], [122], is generally integrated into the standard softmax cross-entropy loss. Its basic idea is to avoid the model fitting to over-confident annotated labels, improving the generalizability [148].

Verification Loss. It optimizes the pairwise relationship, either with a contrastive loss [96], [120] or binary verification loss [43], [141]. The contrastive loss improves the relative pairwise distance comparison, formulated by

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

where d_{ij} represents the Euclidean distance between the embedding features of two input samples x_i and x_j. δ_{ij} is a binary label indicator ($\delta_{ij} = 1$ when x_i and x_j belong to the same identity, and $\delta_{ij} = 0$, otherwise). ρ is a margin parameter. There are several variants, *e.g.*, the pairwise comparison with ranking SVM in [81].

Binary verification [43], [141] discriminates the positive and negative of a input image pair. Generally, a differential feature f_{ij} is obtained by $f_{ij} = (f_j - f_j)^2$ [141], where f_i and f_j are the embedding features of two samples x_i and x_j. The verification network classifies the differential feature into positive or negative. We use $p(\delta_{ij}|f_{ij})$ to represent the probability of an input pair (x_i and x_j) being recognized as δ_{ij} (0 or 1). The verification loss with cross-entropy is

$$\mathcal{L}_{veri}(i,j) = -\delta_{ij}\log(p(\delta_{ij}|f_{ij})) - (1-\delta_{ij})\log(1-p(\delta_{ij}|f_{ij})). \quad (3)$$

The verification is often combined with the identity loss to improve the performance [94], [96], [120], [141].

Triplet loss. It treats the Re-ID model training process as a retrieval ranking problem. The basic idea is that the distance between the positive pair should be smaller than the negative pair by a pre-defined margin [57]. Typically, a triplet contains one anchor sample x_i, one positive sample x_j with the same identity, and one negative sample x_k from a different identity. The triplet loss with a margin parameter is represented by

$$\mathcal{L}_{tri}(i,j,k) = \max(\rho + d_{ij} - d_{ik}, 0), \quad (4)$$

where $d(\cdot)$ measures the Euclidean distance between two samples. The large proportion of easy triplets will dominate the training process if we directly optimize above loss function, resulting in limited discriminability. To alleviate this issue, various informative triplet mining methods have been designed [14], [22], [57], [97]. The basic idea is to select the informative triplets [57], [149]. Specifically, a moderate positive mining with a weight constraint is introduced in [149], which directly optimizes the feature difference. Hermans *et al.* [57] demonstrate that the online hardest positive and negative mining within each training batch is beneficial for discriminative Re-ID model learning. Some methods also studied the point to set similarity strategy for informative triplet mining [150], [151]. This enhances robustness against the outlier samples with a soft hard-mining scheme.

To further enrich the triplet supervision, a quadruplet deep network is developed in [152], where each quadruplet contains one anchor sample, one positive sample and two mined negative samples. The quadruplets are formulated with a margin-based online hard negative mining. Optimizing the quadruplet relationship results in smaller intra-class variation and larger inter-class variation.

The combination of triplet loss and identity loss is one of the most popular solutions for deep Re-ID model learning [28], [87], [90], [93], [103], [104], [116], [137], [142], [153], [154]. These two components are mutually beneficial for discriminative feature representation learning.

OIM loss. In addition to the above three kinds of loss functions, an Online Instance Matching (OIM) loss [64] is designed with a memory bank scheme. A memory bank $\{v_k, k = 1, 2, \cdots, c\}$ contains the stored instance features, where c denotes the class number. The OIM loss is then formulated by

$$\mathcal{L}_{oim} = -\frac{1}{n}\sum\nolimits_{i=1}^{n} \log\frac{\exp(v_i^T f_i/\tau)}{\sum_{k=1}^{c}\exp(v_k^T f_i/\tau)}, \quad (5)$$

where v_i represents the corresponding stored memory feature for class y_i, and τ is a temperature parameter that controls the similarity space [145]. $v_i^T f_i$ measures the online instance matching score. The comparison with a memorized feature set of unlabelled identities is further included to calculate the denominator [64], handling the large instance number of non-targeted identities. This memory scheme is also adopted in unsupervised domain adaptive Re-ID [106].

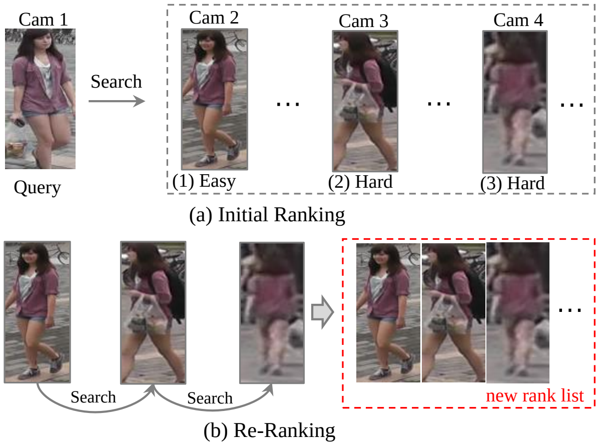


Fig. 4: An illustration of re-ranking in person Re-ID. Given a query example, an initial rank list is retrieved, where the hard matches are ranked in the bottom. Using the top-ranked easy positive match (1) as query to search in the gallery, we can get the hard match (2) and (3) with similarity propagation in the gallery set.

2.2.2 *Training strategy*

The batch sampling strategy plays an important role in discriminative Re-ID model learning. It is challenging since the number of annotated training images for each identity varies significantly [5]. Meanwhile, the severely imbalanced positive and negative sample pairs increases additional difficulty for the training strategy design [40].

The most commonly used training strategy for handling the imbalanced issue is identity sampling [57], [122]. For each training batch, a certain number of identities are randomly selected, and then several images are sampled from each selected identity. This batch sampling strategy guarantees the informative positive and negative mining.

To handle the imbalance issue between the positive and negative, adaptive sampling is the popular approach to adjust the contribution of positive and negative samples, such as Sample Rate Learning (SRL) [89], curriculum sampling [87]. Another approach is sample re-weighting, using the sample distribution [87] or similarity difference [52] to adjust the sample weight. An efficient reference constraint is designed in [155] to transform the pairwise/triplet similarity to a sample-to-reference similarity, addressing the imbalance issue and enhancing the discriminability, which is also robust to outliers.

To adaptively combine multiple loss functions, a multi-loss dynamic training strategy [156] adaptively reweights the identity loss and triplet loss, extracting appropriate component shared between them. This multi-loss training strategy leads to consistent performance gain.

2.3 Ranking Optimization

Ranking optimization plays a crucial role in improving the retrieval performance in the testing stage. Given an initial ranking list, it optimizes the ranking order, either by automatic gallery-to-gallery similarity mining [58], [157] or human interaction [158], [159]. Rank/Metric fusion [160], [161] is another popular approach for improving the ranking performance with multiple ranking list inputs.

2.3.1 Re-ranking

The basic idea of re-ranking is to utilize the gallery-to-gallery similarity to optimize the initial ranking list, as shown in Fig. 4. The top-ranked similarity pulling and bottom-ranked dissimilarity pushing is proposed in [157]. The widely-used k-reciprocal reranking [58] mines the contextual information. Similar idea for contextual information modeling is applied in [25]. Bai *et al.* [162] utilize the geometric structure of the underlying manifold. An expanded cross neighborhood re-ranking method [18] is introduced by integrating the cross neighborhood distance. A local blurring re-ranking [95] employs the clustering structure to improve neighborhood similarity measurement.

Query Adaptive. Considering the query difference, some methods have designed the query adaptive retrieval strategy to replace the uniform searching engine to improve the performance [163], [164]. Andy *et al.* [163] propose a query adaptive re-ranking method using locality preserving projections. An efficient online local metric adaptation method is presented in [164], which learns a strictly local metric with mined negative samples for each probe.

Human Interaction. It involves using human feedback to optimize the ranking list [158]. This provides reliable supervision during the re-ranking process. A hybrid human-computer incremental learning model is presented in [159], which cumulatively learns from human feedback, improving the Re-ID ranking performance on-the-fly.

2.3.2 Rank Fusion

Rank fusion exploits multiple ranking lists obtained with different methods to improve the retrieval performance [59]. Zheng *et al.* [165] propose a query adaptive late fusion method on top of a "L" shaped observation to fuse methods. A rank aggregation method by employing the similarity and dissimilarity is developed in [59]. The rank fusion process in person Re-ID is formulated as a consensus-based decision problem with graph theory [166], mapping the similarity scores obtained by multiple algorithms into a graph with path searching. An Unified Ensemble Diffusion (UED) [161] is recently designed for metric fusion. UED maintains the advantages of three existing fusion algorithms, optimized by a new objective function and derivation. The metric ensemble learning is also studied in [160].

2.4 Datasets and Evaluation

2.4.1 Datasets and Evaluation Metrics

Datasets. We first review the widely used datasets for the closed-world setting, including 11 image datasets (VIPeR [31], iLIDS [167], GRID [168], PRID2011 [126], CUHK01-03 [43], Market-1501 [5], DukeMTMC [42], Airport [169] and MSMT17 [44]) and 7 video datasets (PRID-2011 [126], iLIDS-VID [7], MARS [8], Duke-Video [144], Duke-Tracklet [170], LPW [171] and LS-VID [136]). The statistics of these datasets are shown in Table 2. This survey only focuses on the general large-scale datsets for deep learning methods. A comprehensive summarization of the Re-ID datasets can be found in [169] and their website[1]. Several observations can be made in terms of the dataset collection over recent years:

TABLE 2: Statistics of some commonly used datasets for closed-world person Re-ID. "both" means that it contains both hand-cropped and detected bounding boxes. "C&M" means both CMC and mAP are evaluated.

			Image datasets				
Dataset	Time	#ID	#image	#cam.	Label	Res.	Eval.
VIPeR	2007	632	1,264	2	hand	fixed	CMC
iLIDS	2009	119	476	2	hand	vary	CMC
GRID	2009	250	1,275	8	hand	vary	CMC
PRID2011	2011	200	1,134	2	hand	fixed	CMC
CUHK01	2012	971	3,884	2	hand	fixed	CMC
CUHK02	2013	1,816	7,264	10	hand	fixed	CMC
CUHK03	2014	1,467	13,164	2	both	vary	CMC
Market-1501	2015	1,501	32,668	6	both	fixed	C&M
DukeMTMC	2017	1,404	36,411	8	both	fixed	C&M
Airport	2017	9,651	39,902	6	auto	fixed	C&M
MSMT17	2018	4,101	126,441	15	auto	vary	C&M
			Video datasets				
Dataset	time	#ID	#track(#bbox)	#cam.	label	Res.	Eval
PRID-2011	2011	200	400 (40k)	2	hand	fixed	CMC
iLIDS-VID	2014	300	600 (44k)	2	hand	vary	CMC
MARS	2016	1261	20,715 (1M)	6	auto	fixed	C&M
Duke-Video	2018	1,812	4,832 (-)	8	auto	fixed	C&M
Duke-Tracklet	2018	1,788	12,647 (-)	8	auto	C&M	
LPW	2018	2,731	7,694(590K)	4	auto	fixed	C&M
LS-VID	2019	3,772	14,943 (3M)	15	auto	fixed	C&M

1) The dataset scale (both #image and #ID) has increased rapidly. Generally, the deep learning approach can benefit from more training samples. This also increases the annotation difficulty needed in closed-world person Re-ID. 2) The camera number is also greatly increased to approximate the large-scale camera network in practical scenarios. This also introduces additional challenges for model generalizability in a dynamically updated network. 3) The bounding boxes generation is usually performed automatically detected/tracked, rather than manually cropped. This simulates the real-world scenario with tracking/detection errors.

Evaluation Metrics. To evaluate a Re-ID system, Cumulative Matching Characteristics (CMC) [68] and mean Average Precision (mAP) [5] are two widely used measurements.

CMC-k (*a.k.a*, Rank-k matching accuracy) [68] represents the probability that a correct match appears in the top-k ranked retrieved results. CMC is accurate when only one ground truth exists for each query, since it only considers the first match in evaluation process. However, the gallery set usually contains multiple groundtruths in a large camera network, and CMC cannot completely reflect the discriminability of a model across multiple cameras.

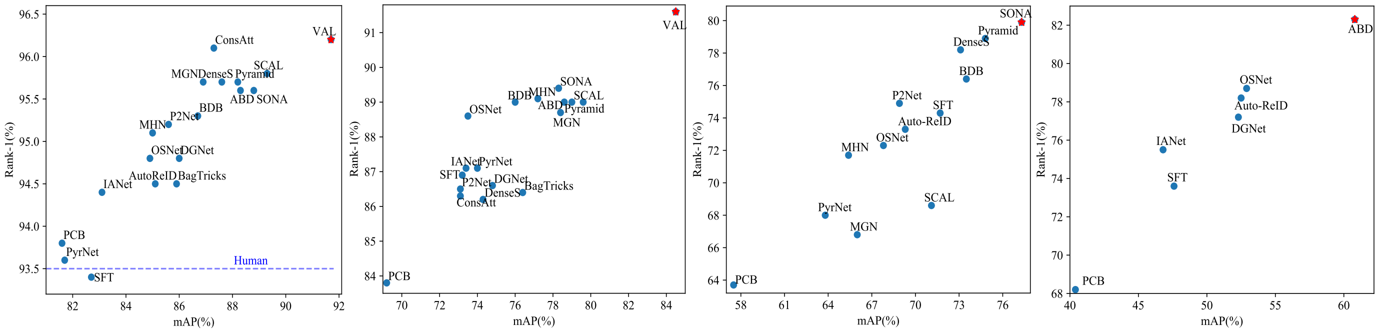
Another metric, *i.e.*, mean Average Precision (mAP) [5], measures the average retrieval performance with multiple grountruths. It is originally widely used in image retrieval. For Re-ID evaluation, it can address the issue of two systems performing equally well in searching the first ground truth (might be easy match as in Fig. 4), but having different retrieval abilities for other hard matches.

Considering the efficiency and complexity of training a Re-ID model, some recent works [138], [139] also report the FLoating-point Operations Per second (FLOPs) and the network parameter size as the evaluation metrics. These two metrics are crucial when the training/testing device has limited computational resources.

2.4.2 In-depth Analysis on State-of-The-Arts

We review the state-of-the-arts from both image-based and video-based perspectives. We include methods published in top CV venues over the past three years.

[1] https://github.com/NEU-Gou/awesome-reid-dataset



(a) SOTA on Market-1501 [5]
(b) SOTA on DukeMTMC [42]
(c) SOTA on CUHK03 [43]
(d) SOTA on MSMT17 [44]

Fig. 5: State-of-the-arts (SOTA) on four image-based person Re-ID datasets. Both the Rank-1 accuracy (%) and mAP value (%) are reported. For CUHK03 [43], the detected data under the setting [58] is reported. For Market-1501, the single query setting is used. The best result is highlighted with a red star. All the listed results do not use re-ranking or additional annotated information.

Image-based Re-ID. There are a large number of published papers for image-based Re-ID[2]. We mainly review the works published in 2019 as well as some representative works in 2018. Specifically, we include PCB [77], MGN [172], PyrNet [6], Auto-ReID [139], ABD-Net [173], BagTricks [122], OSNet [138], DGNet [119], SCAL [90], MHN [174], P2Net [104], BDB [124], SONA [107], SFT [95], ConsAtt [93], DenseS [103], Pyramid [156], IANet [108], VAL [114]. We summarize the results on four datasets (Fig. 5). This overview motivates five major insights, as discussed below.

First, with the advancement of deep learning, most of the image-based Re-ID methods have achieved higher rank-1 accuracy than humans (93.5% [175]) on the widely used Market-1501 dataset. In particular, VAL [114] obtains the best mAP of 91.6% and Rank-1 accuracy of 96.2% on Market-1501 dataset. The major advantage of VAL is the usage of viewpoint information. The performance can be further improved when using re-ranking or metric fusion. The success of deep learning on these closed-world datasets also motivates the shift focus to more challenging scenarios, *i.e.*, large data size [136] or unsupervised learning [176].

Second, part-level feature learning is beneficial for discriminative Re-ID model learning. Global feature learning directly learns the representation on the whole image without the part constraints [122]. It is discriminative when the person detection/ tracking can accurately locate the human body. When the person images suffer from large background clutter or heavy occlusions, part-level feature learning usually achieves better performance by mining discriminative body regions [67]. Due to its advantage in handling misalignment/occlusions, we observe that most of the state-of-the-art methods developed recently adopt the features aggregation paradigm, combining the part-level and full human body features [139], [156].

Third, attention is beneficial for discriminative Re-ID model learning. We observe that all the methods (ConsAtt [93], SCAL [90], SONA [107], ABD-Net [173]) achieving the best performance on each dataset adopt an attention scheme. The attention captures the relationship between different convolutional channels, multiple feature maps, hierarchical layers, different body parts/regions, and even multiple images. Meanwhile, discriminative [173], diverse [133], consistent [93] and high-order [107] properties are incorporated to enhance the attentive feature learning. Considering the powerful attention schemes and the specificity of the Re-ID problem, it is highly possible that attentive deeply learned systems will continue dominating the Re-ID community, with more domain specific properties.

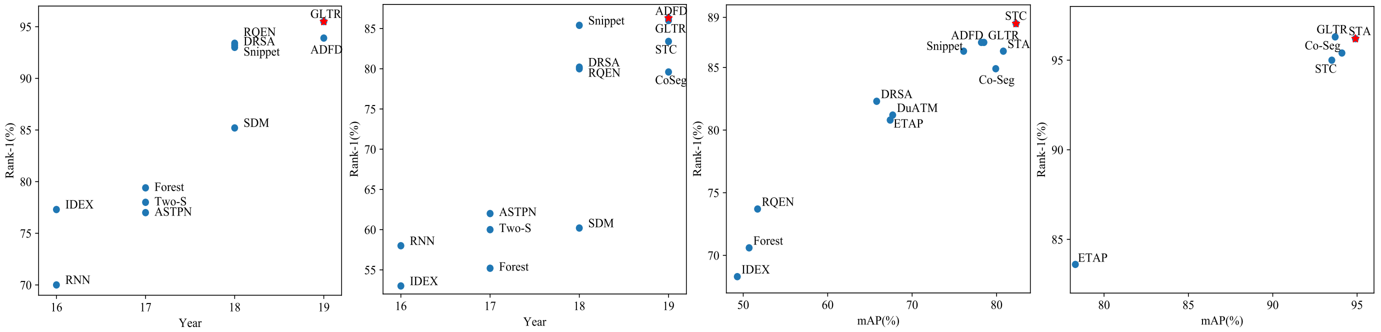
Fourth, multi-loss training can improve the Re-ID model learning. Different loss functions optimize the network from a multi-view perspective. Combining multiple loss functions can improve the performance, evidenced by the multi-loss training strategy in the state-of-the-art methods, including ConsAtt [93], ABD-Net [173] and SONA [107]. In addition, a dynamic multi-loss training strategy is designed in [156] to adaptively integrated two loss functions. The combination of identity loss and triplet loss with hard mining is the primary choice. Moreover, due to the imbalanced issue, sample weighting strategy generally improves the performance by mining informative triplets [52], [89].

Finally, there is still much room for further improvement due to the increasing size of datasets, complex environment, limited training samples. For example, the Rank-1 accuracy (82.3%) and mAP (60.8%) on the newly released MSMT17 dataset [44] are much lower than that on Market-1501 (Rank-1: 96.2% and mAP 91.7%) and DukeMTMC (Rank-1: 91.6% and mAP 84.5%). On some other challenging datasets with limited training samples (*e.g.*, GRID [168] and VIPeR [31]), the performance is still very low. In addition, Re-ID models usually suffers significantly on cross-dataset evaluation [28], [54], and the performance drops dramatically under adversarial attack [177]. We are optimistic that there would be important breakthroughs in person Re-ID, with increasing discriminability, robustness, and generalizability.

Video-based Re-ID. Video-based Re-ID has received less interest, compared to image-based Re-ID. We review the deeply learned Re-ID models, including CoSeg [132], GLTR [136], STA [135], ADFD [110], STC [20], DRSA [133], Snippet [134], ETAP [144], DuATM [91], SDM [178], TwoS [128], ASTPN [131], RQEN [171], Forest [130], RNN [127] and IDEX [8]. We also summarize the results on four video Re-ID datasets, as shown in Fig. 6. From these results, the following observations can be drawn.

First, a clear trend of increasing performance can be seen over the years with the development of deep learning techniques. Specifically, the Rank-1 accuracy increases from 70% (RNN [127] in 2016) to 95.5% (GLTR [136] in 2019) on PRID-2011 dataset, and from 58% (RNN [127]) to 86.3%

[2] https://paperswithcode.com/task/person-re-identification



(a) SOTA on PRID-2011 [126] (b) SOTA on iLIDS-VID [7] (c) SOTA on MARS [8] (d) SOTA on Duke-Video [144]

Fig. 6: State-of-the-arts (SOTA) on four widely used video-based person Re-ID datasets. The Rank-1 accuracies (%) over years are reported. mAP values (%) on MARS [8] and Duke-Video [144] are reported. For Duke-Video, we refer to the settings in [144]. The best result is highlighted with a red star. All the listed results do not use re-ranking or additional annotated information.

(ADFD [110]) on iLIDS-VID dataset. On the large-scale MARS dataset, the Rank-1 accuracy/mAP increase from 68.3%/49.3% (IDEX [8]) to 88.5%/82.3% (STC [20]). On the Duke-Video dataset [144], STA [135] also achieves a Rank-1 accuracy of 96.2%, and the mAP is 94.9%.

Second, spatial and temporal modeling is crucial for discriminative video representation learning. We observe that all the methods (STA [135], STC [20], GLTR [136]) design spatial-temporal aggregation strategies to improve the video Re-ID performance. Similar to image-based Re-ID, the attention scheme across multiple frames [110], [135] also greatly enhances the discriminability. Another interesting observation in [20] demonstrates that utilizing multiple frames within the video sequence can fill in the occluded regions, which provides a possible solution for handling the challenging occlusion problem in the future.

Finally, the performance on these datases has reached a saturation state, usually about less than 1% accuracy gain on these four video datasets. However, there is still large room for improvements on the challenging cases. For example, on the newly collected video dataset, LS-VID [136], the Rank-1 accuracy/mAP of GLTR [136] are only 63.1%/44.43%, while GLTR [136] can achieve state-of-the-art or at least comparable performance on the other four daatsets. LS-VID [136] contains significantly more identities and video sequences. This provides a challenging benchmark for future breakthroughs in video based Re-ID.

3 Open-world Person Re-Identification

This section reviews open-world person Re-ID as discussed in § 1, including heterogeneous Re-ID by matching person images across heterogeneous modalities (§ 3.1), end-to-end Re-ID from the raw images/videos (§ 3.2), semi-/unsupervised learning with limited/unavailable annotated labels (§ 3.3), robust Re-ID model learning with noisy annotations (§ 3.4) and open-set person Re-ID when the correct match does not occur in the gallery (§ 3.5).

3.1 Heterogeneous Re-ID

This subsection summarizes four main kinds of heterogeneous Re-ID, including Re-ID between depth and RGB images (§ 3.1.1), text-to-image Re-ID (§ 3.1.2), visible-to-infrared Re-ID (§ 3.1.3) and cross resolution Re-ID (§ 3.1.4).

3.1.1 Depth-based Re-ID

Depth images capture the body shape and skeleton information. This provides the possibility for Re-ID under illumination/clothes changing environments, which is also important for personalized human interaction applications.

A recurrent attention-based model is proposed in [179] to address the depth-based person identification. In a reinforcement learning framework, they combine the convolutional and recurrent neural networks to identify small, discriminative local regions of the human body.

Karianakis *et al.* [180] leverage the large RGB datasets to design a split-rate RGB-to-Depth transfer method, which bridges the gap between the depth images and the RGB images. Their model further incorporates a temporal attention to enhance video representation for depth Re-ID.

Some methods [62], [181] have also studied the combination of RGB and depth information to improve the Re-ID performance, addressing the clothes-changing challenge.

3.1.2 Text-to-Image Re-ID

Text-to-image Re-ID addresses the matching between a text description and RGB images [63]. It is imperative when the visual image of query person cannot be obtained, and only a text description can be alternatively provided.

A gated neural attention model [63] with recurrent neural network learns the shared features between the text description and the person images. This enables the end-to-end training for text to image pedestrian retrieval. Cheng *et al.* [182] propose a global discriminative image-language association learning method, capturing the identity discriminative information and local reconstructive image-language association under a reconstruction process. A cross projection learning method [183] also learns a shared space with image-to-text matching. A deep adversarial graph attention convolution network is designed in [184] with graph relation mining. However, the large semantic gap between the text descriptions and the visual images is still challenging. Meanwhile, how to combine the texts and hand-painting sketch image is also worth studying in the future.

3.1.3 Visible-Infrared Re-ID

Visible-Infrared Re-ID handles the cross-modality matching between the daytime visible and night-time infrared images. It is important in low-lighting conditions, where the images can only be captured by infrared cameras [21], [60], [185].

Wu *et al.* [21] start the first attempt to address this issue, by proposing a deep zero-padding framework [21] to adaptively learn the modality sharable features. A two stream network is introduced in [142], [186] to model the modality-sharable and -specific information, addressing the intra- and cross-modality variations simultaneously. Besides the cross-modality shared embedding learning [187], the classifier-level discrepancy is also investigated in [188]. Recent methods [189], [190] adopt the GAN technique to generate cross-modality person images to reduce the cross-modality discrepancy at both image and feature level. Hierarchical cross-Modality disentanglement factors are modeled in [191]. A dual-attentive aggregation learning method is presented in [192] to capture multi-level relations.

3.1.4 Cross-Resolution Re-ID

Cross-Resolution Re-ID conducts the matching between low-resolution and high-resolution images, addressing the large resolution variations [13], [14]. A cascaded SR-GAN [193] generates the high-resolution person images in a cascaded manner, incorporating the identity information. Li *et al.* [194] adopt the adversarial learning technique to obtain resolution-invariant image representations.

3.2 End-to-End Re-ID

End-to-end Re-ID alleviates the reliance on additional step for bounding boxes generation. It involves the person Re-ID from raw images or videos, and multi-camera tracking.

Re-ID in Raw Images/Videos This task requires that the model jointly performs the person detection and re-identification in a single framework [55], [64]. It is challenging due to the different focuses of two major components.

Zheng *et al.* [55] present a two-stage framework, and systematically evaluate the benefits and limitations of person detection for the later stage person Re-ID. Xiao *et al.* [64] design an end-to-end person search system using a single convolutional neural network for joint person detection and re-identification. A Neural Person Search Machine (NPSM) [195] is developed to recursively refine the searching area and locate the target person by fully exploiting the contextual information between the query and the detected candidate region. Similarly, a contextual instance expansion module [196] is learned in a graph learning framework to improve the end-to-end person search. A query-guided end-to-end person search system [197] is developed using the Siamese squeeze-and-excitation network to capture the global context information with query-guided region proposal generation. A localization refinement scheme with discriminative Re-ID feature learning is introduced in [198] to generate more reliable bounding boxes. An Identity DiscriminativE Attention reinforcement Learning (IDEAL) method [199] selects informative regions for auto-generated bounding boxes, improving the Re-ID performance.

Yamaguchi *et al.* [200] investigate a more challenging problem, *i.e.*, searching for the person from raw videos with text description. A multi-stage method with spatio-temporal person detection and multi-modal retrieval is proposed. Further exploration along this direction is expected.

Multi-camera Tracking End-to-end person Re-ID is also closely related to multi-person, multi-camera tracking [52]. A graph-based formulation to link person hypotheses is proposed for multi-person tracking [201], where the holistic features of the full human body and body pose layout are combined as the representation for each person. Ristani *et al.* [52] learn the correlation between the multi-target multi-camera tracking and person Re-ID by hard-identity mining and adaptive weighted triplet learning. Recently, a locality aware appearance metric (LAAM) [202] with both intra- and inter-camera relation modeling is proposed.

3.3 Semi-supervised and Unsupervised Re-ID

3.3.1 Unsupervised Re-ID

Early unsupervised Re-ID mainly learns invariant components, *i.e.*, dictionary [203], metric [204] or saliency [66], which leads to limited discriminability or scalability.

For deeply unsupervised methods, cross-camera label estimation is one the popular approaches [176], [205]. Dynamic graph matching (DGM) [206] formulates the label estimation as a bipartite graph matching problem. To further improve the performance, global camera network constraints [207] are exploited for consistent matching. Liu *et al.* progressively mine the labels with step-wise metric promotion [204]. A robust anchor embedding method [83] iteratively assigns labels to the unlabelled tracklets to enlarge the anchor video sequences set. With the estimated labels, deep learning can be applied to learn Re-ID models.

For end-to-end unsupervised Re-ID, an iterative clustering and Re-ID model learning is presented in [205]. Similarly, the relations among samples are utilized in a hierarchical clustering framework [208]. Soft multi-label learning [209] mines the soft label information from a reference set for unsupervised learning. A Tracklet Association Unsupervised Deep Learning (TAUDL) framework [170] jointly conducts the within-camera tracklet association and model the cross-camera tracklet correlation. Similarly, an unsupervised camera-aware similarity consistency mining method [210] is also presented in a coarse-to-fine consistency learning scheme. The intra-camera mining and inter-camera association is applied in a graph association framework [211]. The semantic attributes are also adopted in Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) framework [111]. However, it is still challenging for model updating with newly arriving unlabelled data.

Besides, several methods have also tried to learn a part-level representation based on the observation that it is easier to mine the label information in local parts than that of a whole image. A PatchNet [153] is designed to learn discriminative patch features by mining patch level similarity. A Self-similarity Grouping (SSG) approach [212] iteratively conducts grouping (exploits both the global body and local parts similarity for pseudo labeling) and Re-ID model training in a self-paced manner.

Semi-/Weakly supervised Re-ID. With limited label information, a one-shot metric learning method is proposed in [213], which incorporates a deep texture representation and a color metric. A stepwise one-shot learning method (EUG) is proposed in [144] for video-based Re-ID, gradually selecting a few candidates from unlabeled tracklets to enrich the labeled tracklet set. A multiple instance attention learning framework [214] uses the video-level labels for representation learning, alleviating the reliance on full annotation.

3.3.2 Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) transfers the knowledge on a labeled source dataset to the unlabeled target dataset [53]. Due to the large domain shift and powerful supervision in source dataset, it is another popular approach for unsupervised Re-ID without target dataset labels.

Target Image Generation. Using GAN generation to transfer the source domain images to target-domain style is a popular approach for UDA Re-ID. With the generated images, this enables supervised Re-ID model learning in the unlabeled target domain. Wei *et al.* [44] propose a Person Transfer Generative Adversarial Network (PTGAN), transferring the knowledge from one labeled source dataset to the unlabeled target dataset. Preserved self-similarity and domain-dissimilarity [120] is trained with a similarity preserving generative adversarial network (SPGAN). A Hetero-Homogeneous Learning (HHL) method [215] simultaneously considers the camera invariance with homogeneous learning and domain connectedness with heterogeneous learning. An adaptive transfer network [216] decomposes the adaptation process into certain imaging factors, including illumination, resolution, camera view, etc. This strategy improves the cross-dataset performance. Huang *et al.* [217] try to suppress the background shift to minimize the domain shift problem. Chen *et al.* [218] design an instance-guided context rendering scheme to transfer the person identities from source domain into diverse contexts in the target domain. Besides, a pose disentanglement scheme is added to improve the image generation [121]. A mutual mean-teacher learning scheme is also developed in [219]. However, the scalability and stability of the image generation for practical large-scale changing environment are still challenging.

Bak *et al.* [125] generate a synthetic dataset with different illumination conditions to model realistic indoor and outdoor lighting. The synthesized dataset increases generalizability of the learned model and can be easily adapted to a new dataset without additional supervision [220].

Target Domain Supervision Mining. Some methods directly mine the supervision on the unlabeled target dataset with a well trained model from source dataset. An exemplar memory learning scheme [106] considers three invariant cues as the supervision, including exemplar-invariance, camera invariance and neighborhood-invariance. The Domain-Invariant Mapping Network (DIMN) [28] formulates a meta-learning pipeline for the domain transfer task, and a subset of source domain is sampled at each training episode to update the memory bank, enhancing the scalability and discriminability. The camera view information is also applied in [221] as the supervision signal to reduce the domain gap. A self-training method with progressive augmentation [222] jointly captures the local structure and global data distribution on the target dataset. Recently, a self-paced contrastive learning framework with hybrid memory [223] is developed with great success, which dynamically generates multi-level supervision signals.

The spatio-temporal information is also utilized as the supervision in TFusion [224]. TFusion transfers the spatio-temporal patterns learned in the source domain to the target domain with a Bayesian fusion model. Similarly, Query-Adaptive Convolution (QAConv) [225] is developed to improve cross-dataset accuracy.

TABLE 3: Statistics of SOTA unsupervised person Re-ID on two image-based datasets. “Source” represents if it utilizes the source annotated data in training the target Re-ID model. “Gen.” indicates if it contains an image generation process. Rank-1 accuracy (%) and mAP (%) are reported.

			Market-1501		*DukeMTMC*	
Methods	Source	Gen.	R1	mAP	R1	mAP
CAMEL [226] ICCV17	Model	No	54.5	26.3	-	-
PUL [205] TOMM18	Model	No	45.5	20.5	30.0	16.4
PTGAN [120] CVPR18	Data	Yes	58.1	26.9	46.9	26.4
TJ-AIDL[†][111] CVPR18	Data	No	58.2	26.5	44.3	23.0
HHL [215] ECCV18	Data	Yes	62.2	31.4	46.9	27.2
MAR[‡] [209] CVPR19	Data	No	67.7	40.0	67.1	48.0
ENC [106] CVPR19	Data	No	75.1	43.0	63.3	40.4
ATNet [216] CVPR19	Data	Yes	55.7	25.6	45.1	24.9
PAUL[‡] [153] CVPR19	Model	No	68.5	40.1	72.0	53.2
SBGAN [217] ICCV19	Data	Yes	58.5	27.3	53.5	30.8
UCDA [221] ICCV19	Data	No	64.3	34.5	55.4	36.7
CASC[‡] [210] ICCV19	Model	No	65.4	35.5	59.3	37.8
PDA [121] ICCV19	Data	Yes	75.2	47.6	63.2	45.1
CR-GAN [218] ICCV19	Data	Yes	77.7	54.0	68.9	48.6
PAST [222] ICCV19	Model	No	78.4	54.6	72.4	54.3
SSG [212] ICCV19	Model	No	80.0	58.3	73.0	53.4
HCT [208] CVPR20	Model	No	80.0	56.4	69.6	50.7
SNR [227] CVPR20	Data	No	82.8	61.7	76.3	58.1
MMT [219] ICLR20	Data	No	87.7	71.2	78.0	65.1
MEB-Net [228] ECCV20	Data	No	89.9	76.0	79.6	66.1
SpCL [223] NeurIPS20	Data	No	**90.3**	**76.7**	**82.9**	**68.8**

- [†] TJ-AIDL [111] requires additional attribute annotation.
- [§] DAS [125] generates synthesized virtual humans under vairous lightings.
- [‡] PAUL [153], MAR [209] and CASC [210] use MSMT17 as source dataset.

3.3.3 State-of-The-Arts for Unsupervised Re-ID

Unsupervised Re-ID has achieved increasing attention in recent years, evidenced by the increasing number of publications in top venues. We review the SOTA for unsupervised deeply learned methods on two widely-used image-based Re-ID datasets. The results are summarized in Table 3. From these results, the following insights can be drawn.

First, the unsupervised Re-ID performance has increased significantly over the years. The Rank-1 accuracy/mAP increases from 54.5%/26.3% (CAMEL [226]) to 90.3%/76.7% (SpCL [223]) on the Market-1501 dataset within three years. The performance for DukeMTMC dataset increases from 30.0%/16.4% to 82.9%/68.8%. The gap between the supervised upper bound and the unsupervised learning is narrowed significantly. This demonstrates the success of unsupervised Re-ID with deep learning.

Second, current unsupervised Re-ID is still underdeveloped and it can be further improved in the following aspects: 1) The powerful attention scheme in supervised Re-ID methods has rarely been applied in unsupervised Re-ID. 2) Target domain image generation has been proved effective in some methods, but they are not applied in two best methods (PAST [222], SSG [212]). 3) Using the annotated source data in the training process of the target domain is beneficial for cross-dataset learning, but it is also not included in above two methods. These observations provide the potential basis for further improvements.

Third, there is still a large gap between the unsupervised and supervised Re-ID. For example, the rank-1 accuracy of supervised ConsAtt [93] has achieved 96.1% on the Market-1501 dataset, while the highest accuracy of unsupervised SpCL [223] is about 90.3%. Recently, He *et al.* [229] have demonstrated that unsupervised learning with large-scale unlabeled training data has the ability to outperform the supervised learning on various tasks [230]. We expect that several breakthroughs in future unsupervised Re-ID.

3.4 Noise-Robust Re-ID

Re-ID usually suffers from unavoidable noise due to data collection and annotation difficulty. We review noise-robust Re-ID from three aspects: *Partial Re-ID* with heavy occlusion, *Re-ID with sample noise* caused by detection or tracking errors, and *Re-ID with label noise* caused by annotation error.

Partial Re-ID. This addresses the Re-ID problem with heavy occlusions, *i.e.*, only part of the human body is visible [231]. A fully convolutional network [232] is adopted to generate fix-sized spatial feature maps for the incomplete person images. Deep Spatial feature Reconstruction (DSR) is further incorporated to avoid explicit alignment by exploiting the reconstructing error. Sun *et al.* [67] design a Visibility-aware Part Model (VPM) to extract sharable region-level features, thus suppressing the spatial misalignment in the incomplete images. A foreground-aware pyramid reconstruction scheme [233] also tries to learn from the unoccluded regions. The Pose-Guided Feature Alignment (PGFA) [234] exploits the pose landmarks to mine discriminative part information from occlusion noise. However, it is still challenging due to the severe partial misalignment, unpredictable visible regions and distracting unshared body regions. Meanwhile, how to adaptively adjust the matching model for different queries still needs further investigation.

Re-ID with Sample Noise. This refers to the problem of the person images or the video sequence containing outlying regions/frames, either caused by poor detection/inaccurate tracking results. To handle the outlying regions or background clutter within the person image, pose estimation cues [17], [18] or attention cues [22], [66], [199] are exploited. The basic idea is to suppress the contribution of the noisy regions in the final holistic representation. For video sequences, set-level feature learning [83] or frame level re-weighting [134] are the commonly used approaches to reduce the impact of noisy frames. Hou *et al.* [20] also utilize multiple video frames to auto-complete occluded regions. It is expected that more domain-specific sample noise handling designs in the future.

Re-ID with Label Noise. Label noise is usually unavoidable due to annotation error. Zheng *et al.* adopt a label smoothing technique to avoid label overfiting issues [42]. A Distribution Net (DNet) that models the feature uncertainty is proposed in [235] for robust Re-ID model learning against label noise, reducing the impact of samples with high feature uncertainty. Different from the general classification problem, robust Re-ID model learning suffers from limited training samples for each identity [236]. In addition, the unknown new identities increase additional difficulty for the robust Re-ID model learning.

3.5 Open-set Re-ID and Beyond

Open-set Re-ID is usually formulated as a person verification problem, *i.e.*, discriminating whether or not two person images belong to the same identity [69], [70]. The verification usually requires a learned condition τ, *i.e.*, $sim(query, gallery) > \tau$. Early researches design hand-crafted systems [26], [69], [70]. For deep learning methods, an Adversarial PersonNet (APN) is proposed in [237], which jointly learns a GAN module and the Re-ID feature extractor. The basic idea of this GAN is to generate realistic target-like images (imposters) and enforce the feature extractor is robust to the generated image attack. Modeling feature uncertainty is also investigated in [235]. However, it remains quite challenging to achieve a high true target recognition and maintain low false target recognition rate [238].

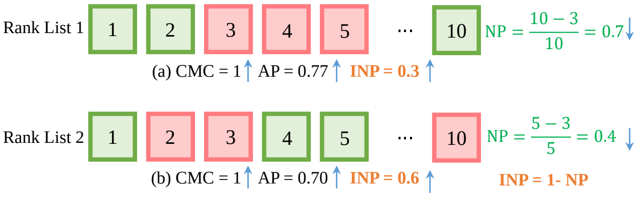

Rank List 1
1
2
3
4
5
···
10
NP = $\frac{10-3}{10}$ = 0.7↓
(a) CMC = 1↑ AP = 0.77↑ INP = 0.3↑
Rank List 2
1
2
3
4
5
···
10
NP = $\frac{5-3}{5}$ = 0.4 ↓
(b) CMC = 1↑ AP = 0.70↑ INP = 0.6↑
INP = 1- NP

Fig. 7: Difference between the widely used CMC, AP and the negative penalty (NP) measurements. True matching and false matching are bounded in green and red boxes, respectively. Assume that only three correct matches exist in the gallery, rank list 1 gets better AP, but gets much worse NP than rank list 2. The main reason is that rank list 1 contains too many false matchings before finding the hardest true matching. For consistency with CMC and mAP, we compute the inverse negative penalty (INP), *e.g.*, INP = 1- NP. Larger INP means better performance.

Group Re-ID. It aims at associating the persons in groups rather than individuals [167]. Early researches mainly focus on group representation extraction with sparse dictionary learning [239] or covariance descriptor aggregation [240]. The multi-grain information is integrated in [241] to fully capture the characteristics of a group. Recently, the graph convoltuional network is applied in [242], representing the group as a graph. The group similarity is also applied in the end-to-end person search [196] and the individual re-identification [197], [243] to improve the accuracy. However, group Re-ID is still challenging since the group variation is more complicated than the individuals.

Dynamic Multi-Camera Network. Dynamic updated multi-camera network is another challenging issue [23], [24], [27], [29], which needs model adaptation for new cameras or probes. A human in-the-loop incremental learning method is introduced in [24] to update the Re-ID model, adapting the representation for different probe galleries. Early research also applies the active learning [27] for continuous Re-ID in multi-camera network. A continuous adaptation method based on sparse non-redundant representative selection is introduced in [23]. A transitive inference algorithm [244] is designed to exploit the best source camera model based on a geodesic flow kernel. Multiple environmental constraints (*e.g.*, Camera Topology) in dense crowds and social relationships are integrated for an open-world person Re-ID system [245]. The model adaptation and environmental factors of cameras are crucial in practical dynamic multi-camera network. Moreover, how to apply the deep learning technique for the dynamic multi-camera network is still less investigated.

4 An Outlook: Re-ID in Next Era

This section firstly presents a new evaluation metric in § 4.1, a strong baseline (in § 4.2) for person Re-ID. It provides an important guidance for future Re-ID research. Finally, we discuss some under-investigated open issues in § 4.3.

TABLE 4: Comparison with the state-of-the-arts on single-modality image-based Re-ID. Rank-1 accuracy (%), mAP (%) and mINP (%) are reported on two public datasets.

	Market-1501 [5]			DukeMTMC [42]		
Method	R1	mAP	mINP	R1	mAP	mINP
BagTricks [122] CVPR19W	94.5	85.9	59.4	86.4	76.4	40.7
ABD-Net [173] ICCV19	**95.6**	**88.3**	**66.2**	89.0	78.6	42.1
B (ours)	94.2	85.4	58.3	86.1	76.1	40.3
B + Att [246]	94.9	86.9	62.2	87.5	77.6	41.9
B + WRT	94.6	86.8	61.9	87.1	77.0	41.4
B + GeM [247]	94.4	86.3	60.1	87.3	77.3	41.9
B + WRT + GeM	94.9	87.1	62.5	88.2	78.1	43.4
AGW (Full)	95.1	87.8	65.0	**89.0**	**79.6**	**45.7**

4.1 mINP: A New Evaluation Metric for Re-ID

For a good Re-ID system, the target person should be retrieved as accurately as possible, *i.e.*, all the correct matches should have low rank values. Considering that the target person should not be neglected in the top-ranked retrieved list, especially for multi-camera network, so as to accurately track the target. When the target person appears in the gallery set at multiple time stamps, the rank position of the hardest correct match determines the workload of the inspectors for further investigation. However, the currently widely used CMC and mAP metrics cannot evaluate this property, as shown in Fig. 7. With the same CMC, rank list 1 achieves a better AP than rank list 2, but it requires more efforts to find all the correct matches. To address this issue, we design a computationally efficient metric, namely a negative penalty (NP), which measures the penalty to find the hardest correct match

$$\mathrm{NP}_i = \frac{R_i^{hard} - |G_i|}{R_i^{hard}}, \tag{6}$$

where R_i^{hard} indicates the rank position of the hardest match, and $|G_i|$ represents the total number of correct matches for query i. Naturally, a smaller NP represents better performance. For consistency with CMC and mAP, we prefer to use the inverse negative penalty (INP), an inverse operation of NP. Overall, the mean INP of all the queries is represented by

$$\mathrm{mINP} = \frac{1}{n}\sum\nolimits_i (1 - \mathrm{NP}_i) = \frac{1}{n}\sum\nolimits_i \frac{|G_i|}{R_i^{hard}}. \tag{7}$$

The calculation of mINP is quite efficient and can be seamlessly integrated in the CMC/mAP calculating process. mINP avoids the domination of easy matches in the mAP/CMC evaluation. One limitation is that mINP value difference for large gallery size would be much smaller compared to small galleries. But it still can reflect the relative performance of a Re-ID model, providing a supplement to the widely-used CMC and mAP metrics.

4.2 A New Baseline for Single-/Cross-Modality Re-ID

According to the discussion in § 2.4.2, we design a new *AGW*[3] baseline for person Re-ID, which achieves competitive performance on both single-modality (image and video) and cross-modality Re-ID tasks. Specifically, our new baseline is designed on top of BagTricks [122], and AGW contains the following three major improved components:

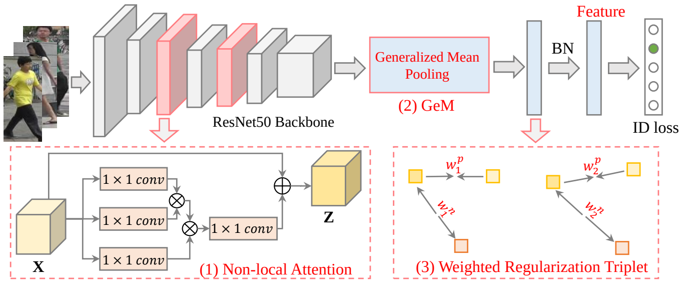


Fig. 8: The framework of the proposed AGW baseline using the widely used ResNet50 [80] as the backbone network.

(1) **Non-local Attention (Att) Block**. As discussed in § 2.4.2, the attention scheme plays a crucial role in discriminative Re-ID model learning. We adopt the powerful non-local attention block [246] to obtain the weighted sum of the features at all positions, represented by

$$\mathbf{z}_i = W_z * \phi(\mathbf{x}_i) + \mathbf{x}_i, \tag{8}$$

where W_z is a weight matrix to be learned, $\phi(\cdot)$ represents a non-local operation, and $+\mathbf{x}_i$ formulates a residual learning strategy. Details can be found in [246]. We adopt the default setting from [246] to insert the non-local attention block.

(2) **Generalized-mean (GeM) Pooling**. As a fine-grained instance retrieval, the widely-used max-pooling or average pooling cannot capture the domain-specific discriminative features. We adopt a learnable pooling layer, named *generalized-mean (GeM) pooling* [247], formulated by

$$\mathbf{f} = [f_1 \cdots f_k \cdots f_K]^T, f_k = (\frac{1}{|\mathcal{X}_k|}\sum\nolimits_{x_i \in \mathcal{X}_k} x_i^{p_k})^{\frac{1}{p_k}}, \tag{9}$$

where f_k represents the feature map, and K is number of feature maps in the last layer. $\mathcal{X}_k$ is the set of $W \times H$ activations for feature map $k \in \{1, 2, \cdots, K\}$. p_k is a pooling hyper-parameter, which is learned in the back-propagation process [247]. The above operation approximates max pooling when $p_k \to \infty$ and average pooling when $p_k = 1$.

(3) **Weighted Regularization Triplet (WRT) loss**. In addition to the baseline identity loss with softmax cross-entropy, we integrate with another weighted regularized triplet loss,

$$\mathcal{L}_{wrt}(i) = \log(1 + \exp(\sum\nolimits_j w_{ij}^p d_{ij}^p - \sum\nolimits_k w_{ik}^n d_{ik}^n)). \tag{10}$$

$$w_{ij}^p = \frac{\exp(d_{ij}^p)}{\sum_{d_{ij}^p \in \mathcal{P}_i} \exp(d_{ij}^p)}, w_{ik}^n = \frac{\exp(-d_{ik}^n)}{\sum_{d_{ik}^n \in \mathcal{N}_i} \exp(-d_{ik}^n)}, \tag{11}$$

where (i, j, k) represents a hard triplet within each training batch. For anchor i, $\mathcal{P}_i$ is the corresponding positive set, and $\mathcal{N}_i$ is the negative set. d_{ij}^p/d_{ik}^n represents the pairwise distance of a positive/negative sample pair. The above weighted regularization inherits the advantage of relative distance optimization between positive and negative pairs, but it avoids introducing any additional margin parameters. Our weighting strategy is similar to [248], but our solution does not introduce additional hyper-parameters.

The overall framework of AGW is shown in Fig 8. Other components are exactly the same as [122]. In the testing phase, the output of BN layer is adopted as the feature representation for Re-ID. The implementation details and more experimental results are in the supplementary material.

[3] Details are in https://github.com/mangye16/ReID-Survey and comprehensive comparison is shown in the supplementary material.

TABLE 5: Comparison with the state-of-the-arts on two image Re-ID datasets, including CUHK03 and MSMT17. Rank-1 accuracy (%), mAP (%) and mINP (%) are reported.

	CUHK03 [43]			MSMT17 [44]		
Method	R1	mAP	mINP	R1	mAP	mINP
BagTricks [122] CVPR19W	58.0	56.6	43.8	63.4	45.1	12.4
AGW (Full)	**63.6**	**62.0**	**50.3**	**68.3**	**49.3**	**14.7**

TABLE 6: Comparison with the state-of-the-arts on four video-based Re-ID datasets, including MARS [8], DukeVideo [144], PRID2011 [126] and iLIDS-VID [7]. Rank-1 accuracy (%), mAP (%) and mINP (%) are reported.

	MARS [8]			DukeVideo [144]		
Method	R1	mAP	mINP	R1	mAP	mINP
BagTricks [122] CVPR19W	85.8	81.6	62.0	92.6	92.4	88.3
CoSeg [132] ICCV19	84.9	79.9	57.8	95.4	94.1	89.8
AGW (Ours)	87.0	82.2	62.8	94.6	93.4	89.2
AGW$_+$ (Ours)	**87.6**	**83.0**	**63.9**	**95.4**	**94.9**	**91.9**

	PRID2011 [126]			iLIDS-VID [7]		
Method	R1	R5	mINP	R1	R5	mINP
BagTricks [122] CVPR19W	84.3	93.3	88.5	74.0	93.3	82.2
AGW (Ours)	87.8	96.6	91.7	78.0	97.0	85.5
AGW$_+$ (Ours)	**94.4**	**98.4**	**95.4**	**83.2**	**98.3**	**89.0**

TABLE 7: Comparison with the state-of-the-arts on two partial Re-ID datasets, including Partial-REID and Partial-iLIDS. Rank-1, -3 accuracy (%) and mINP (%) are reported.

Method	Partial-REID			Partial-iLIDS		
	R1	R3	mINP	R1	R3	mINP
DSR [232] CVPR18	50.7	70.0	-	58.8	67.2	-
SFR [249] ArXiv18	56.9	78.5	-	63.9	74.8	-
VPM [67] CVPR19	67.7	**81.9**	-	**67.2**	76.5	-
BagTricks [122] CVPR19W	62.0	74.0	45.4	58.8	73.9	68.7
AGW	**69.7**	80.0	**56.7**	64.7	**79.8**	**73.3**

TABLE 8: Comparison with the state-of-the-arts on cross-modality visible-infrared Re-ID. Rank-1 accuracy (%), mAP (%) and mINP (%) are reported on two public datasets.

	RegDB [60]		SYSU-MM01 [21]			
	Visible-Thermal		*All Search*		*Indoor Search*	
Method	R1	mAP	R1	mAP	R1	mAP
Zero-Pad [21] ICCV17	17.75	18.90	14.8	15.95	20.58	26.92
HCML [186] AAAI18	24.44	20.08	14.32	16.16	24.52	30.08
eBDTR [142] TIFS19	34.62	33.46	27.82	28.42	32.46	42.46
HSME [187] AAAI19	50.85	47.00	20.68	23.12	-	-
D^{2}RL [189] CVPR19	43.4	44.1	28.9	29.2	-	-
AlignG [190] ICCV19	57.9	53.6	42.4	40.7	45.9	54.3
Hi-CMD [191] CVPR20	70.93	66.04	34.9	35.9		
AGW (Ours)	**70.05** mINP = 50.19	**66.37**	**47.50** mINP =35.30	**47.65**	**54.17** mINP = 59.23	**62.97**

Results on Single-modality Image Re-ID. We first evaluate each component on two image-based datasets (Market-1501 and DukeMTMC) in Table 4. We also list two state-of-the-art methods, BagTricks [122] and ABD-Net [173]. We report the results on CUHK03 and MSMT17 datasets in Table 5. We obtain the following two observations:

1) All the components consistently contribute the accuracy gain, and AGW performs much better than the original BagTricks under various metrics. AGW provides a strong baseline for future improvements. We have also tried to incorporate part-level feature learning [77], but extensive experiments show that it does not improve the performance. How to aggregate part-level feature learning with AGW needs further study in the future. 2) Compared to the current state-of-the-art, ABD-Net [173], AGW performs favorably in most cases. In particular, we achieve much higher mINP on DukeMTMC dataset, 45.7% *vs.* 42.1%. This demonstrates that AGW requires less effort to find all the correct matches, verifying the ability of mINP.

Results on Single-modality Video Re-ID. We also evaluate the proposed AGW on four widely used single modality video-based datasets (MARS [8], DukeVideo [144], PRID2011 [126] and iLIDS-VID [7], as shown in Table 6. We also compare two state-of-the-art methods, BagTricks [122] and Co-Seg [132]. For video data, we develop a variant (AGW$_+$) to capture the temporal information with frame-level average pooling for sequence representation. Meanwhile, constraint random sampling strategy [133] is applied for training. Compared to Co-Seg [132], our AGW$_+$ obtains better Rank-1, mAP and mINP in most cases.

Results on Partial Re-ID. We also test the performance of AGW on two partial Re-ID datasets, as shown in Table 7. The experimental setting are from DSR [232]. We also achieve comparable performance with the state-of-the-art VPM method [67]. This experiment further demonstrates the superiority of AGW for the open-world partial Re-ID task. Meanwhile, the mINP also shows the applicability for this open-world Re-ID problem.

Results on Cross-modality Re-ID. We also test the performance of AGW using a two-stream architecture on the cross-modality visible-infrared Re-ID task. The comparison with the current state-of-the-arts on two datasets is shown in Table 8. We follow the settings in AlignG [190] to perform the experiments. Results show that AGW achieves much higher accuracy than existing cross-modality Re-ID models, verifying the effectiveness for the open-world Re-ID task.

4.3 Under-Investigated Open Issues

We discuss the open-issues from five different aspects according to the five steps in §1, including uncontrollable data collection, human annotation minimization, domain-specific/generalizable architecture design, dynamic model updating and efficient model deployment.

4.3.1 Uncontrollable Data Collection

Most existing Re-ID works evaluate their method on a well-defined data collection environment. However, the data collection in real complex environment is uncontrollable. The data might be captured from unpredictable modality, modality combinations, or even cloth changing data [30].

Multi-Heterogeneous Data. In real applications, the Re-ID data might be captured from multiple heterogeneous modalities, *i.e.*, the resolutions of person images vary a lot [193], both the query and gallery sets may contain different modalities (visible, thermal [21], depth [62] or text description [10]). This results in a challenging multiple heterogeneous person Re-ID. A good person Re-ID system would be able to automatically handle the changing resolutions, different modalities, various environments and multiple domains. Future work with broad generalizability is expected, evaluating their method for different Re-ID tasks.

Cloth-Changing Data. In practical surveillance system, it is very likely to contain a large number of target persons with changing clothes. A cloth-Clothing Change Aware Network (CCAN) [250] addresses this issue by separately extracting the face and body context representation, and similar idea is applied in [251]. Yang *et al.* [30] present a spatial polar transformation (SPT) to learn cross-cloth invariant representation. However, they still rely heavily on the face and body appearance, which might be unavailable and unstable in real scenarios. It would be interesting to further explore the possibility of other discriminative cues (e.g., gait, shape) to address the cloth-changing issue.

4.3.2 Human Annotation Minimization

Besides the unsupervised learning, active learning or human interaction [24], [27], [154], [159] provides another possible solution to alleviate the reliance on human annotation.

Active Learning. Incorporating human interaction, labels are easily provided for newly arriving data and the model can be subsequently updated [24], [27]. A pairwise subset selection framework [252] minimizes human labeling effort by firstly constructing an edge-weighted complete k-partite graph and then solving it as a triangle free subgraph maximization problem. Along this line, a deep reinforcement active learning method [154] iteratively refines the learning policy and trains a Re-ID network with human-in-the-loop supervision. For video data, an interpretable reinforcement learning method with sequential decision making [178] is designed. The active learning is crucial in practical Re-ID system design, but it has received less attention in the research community. Additionally, the newly arriving identities is extremely challenging, even for human. Efficient human in-the-loop active learning is expected in the future.

Learning for Virtual Data. This provides an alternative for minimizing the human annotation. A synthetic dataset is collected in [220] for training, and they achieve competitive performance on real-world datasets when trained on this synthesized dataset. Bak *et al.* [125] generate a new synthetic dataset with different illumination conditions to model realistic indoor and outdoor lighting. A large-scale synthetic PersonX dataset is collected in [105] to systematically study the effect of viewpoint for a person Re-ID system. Recently, the 3D person images are also studied in [253], generating the 3D body structure from 2D images. However, how to bridge the gap between synthesized images and real-world datasets remains challenging.

4.3.3 Domain-Specific/Generalizable Architecture Design

Re-ID Specific Architecture. Existing Re-ID methods usually adopt architectures designed for image classification as the backbone. Some methods modify the architecture to achieve better Re-ID features [82], [122]. Very recently, researchers have started to design domain specific architectures, *e.g.*, OSNet with omni-scale feature learning [138]. It detects the small-scale discriminative features at a certain scale. OSNet is extremely lightweight and achieves competitive performance. With the advancement of automatic neural architecture search (*e.g.*, Auto-ReID [139]), more domain-specific powerful architectures are expected to address the task-specific Re-ID challenges. Limited training samples in Re-ID also increase the difficulty in architecture design.

Domain Generalizable Re-ID. It is well recognized that there is a large domain gap between different datsets [56], [225]. Most existing methods adopt domain adaptation for cross-dataset training. A more practical solution would be learning a domain generalized model with a number of source datasets, such that the learned model can be generalized to new unseen datasets for discriminative Re-ID without additional training [28]. Hu *et al.* [254] studied the cross-dataset person Re-ID by introducing a part-level CNN framework. The Domain-Invariant Mapping Network (DIMN) [28] designs a meta-learning pipeline for domain generalizable Re-ID, learning a mapping between a person image and its identity classifier. The domain generalizability is crucial to deploy the learned Re-ID model under an unknown scenario.

4.3.4 Dynamic Model Updating

Fixed model is inappropriate for practical dynamically updated surveillance system. To alleviate this issue, dynamic model updating is imperative, either to a new domain/camera or adaptation with newly collected data.

Model Adaptation to New Domain/Camera. Model adaptation to a new domain has been widely studied in the literature as a domain adaptation problem [125], [216]. In practical dynamic camera network, a new camera may be temporarily inserted into an existing surveillance system. Model adaptation is crucial for continuous identification in a multi-camera network [23], [29]. To adapt a learned model to a new camera, a transitive inference algorithm [244] is designed to exploit the best source camera model based on a geodesic flow kernel. However, it is still challenging when the newly collected data by the new camera has totally different distributions. In addition, the privacy and efficiency issue [255] also need further consideration.

Model Updating with Newly Arriving Data. With the newly collected data, it is impractical to training the previously learned model from the scratch [24]. An incremental learning approach together with human interaction is designed in [24]. For deeply learned model, an addition using covariance loss [256] is integrated in the overall learning function. However, this problem is not well studied since the deep model training require large amount of training data. Besides, the unknown new identities in the newly arriving data is hard to be identified for the model updating.

4.3.5 Efficient Model Deployment

It is important to design efficient and adaptive models to address scalability issue for practical model deployment.

Fast Re-ID. For fast retrieval, hashing has been extensively studied to boost the searching speed, approximating the nearest neighbor search [257]. Cross-camera Semantic Binary Transformation (CSBT) [258] transforms the original high-dimensional feature representations into compact low-dimensional identity-preserving binary codes. A Coarse-to-Fine (CtF) hashing code search strategy is developed in [259], complementarily using short and long codes. However, the domain-specific hashing still needs further study.

Lightweight Model. Another direction for addressing the scalability issue is to design a lightweight Re-ID model. Modifying the network architecture to achieve a lightweight model is investigated in [86], [138], [139]. Model distillation is another approach, *e.g.*, a multi-teacher adaptive similarity distillation framework is proposed in [260], which learns a user-specified lightweight student model from multiple teacher models, without access to source domain data.

Resource Aware Re-ID. Adaptively adjusting the model according to the hardware configurations also provides a solution to handle the scalability issue. Deep Anytime Re-ID (DaRe) [14] employs a simple distance based routing strategy to adaptively adjust the model, fitting to hardware devices with different computational resources.

5 Concluding Remarks

This paper presents a comprehensive survey with in-depth analysis from a both closed-world and open-world perspectives. We first introduce the widely studied person Re-ID under the closed-world setting from three aspects: feature representation learning, deep metric learning and ranking optimization. With powerful deep learning, the closed-world person Re-ID has achieved performance saturation on several datasets. Correspondingly, the open-world setting has recently gained increasing attention, with efforts to address various practical challenges. We also design a new AGW baseline, which achieves competitive performance on four Re-ID tasks under various metrics. It provides a strong baseline for future improvements. This survey also introduces a new evaluation metric to measure the cost of finding all the correct matches. We believe this survey will provide important guidance for future Re-ID research.

References

[1] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE TPAMI*, vol. 40, no. 2, pp. 392–408, 2018.

[2] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[3] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *CVPR*, 2006, pp. 1528–1535.

[4] J. Almazan, B. Gajic, N. Murray, and D. Larlus, "Re-id done right: towards good practices for person re-identification," *arXiv preprint arXiv:1801.05339*, 2018.

[5] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[6] N. Martinel, G. Luca Foresti, and C. Micheloni, "Aggregating deep pyramidal representations for person re-identification," in *CVPR Workshops*, 2019, pp. 0–0.

[7] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *ECCV*, 2014, pp. 688–703.

[8] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video benchmark for large-scale person re-identification," in *ECCV*, 2016.

[9] M. Ye, C. Liang, Z. Wang, Q. Leng, J. Chen, and J. Liu, "Specific person retrieval via incomplete text description," in *ACM ICMR*, 2015, pp. 547–550.

[10] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *ICCV*, 2017, pp. 1890–1899.

[11] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *ICCV*, 2015, pp. 4516–4524.

[12] S. Bak, S. Zaidenberg, B. Boulay, and F. Bremond, "Improving person re-identification by viewpoint cues," in *AVSS*, 2014, pp. 175–180.

[13] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in *ICCV*, 2015, pp. 3765–3773.

[14] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, "Resource aware person re-identification across multiple resolutions," in *CVPR*, 2018, pp. 8042–8051.

[15] Y. Huang, Z.-J. Zha, X. Fu, and W. Zhang, "Illumination-invariant person re-identification," in *ACM MM*, 2019, pp. 365–373.

[16] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *CVPR*, 2016, pp. 1354–1362.

[17] H. Zhao, M. Tian, S. Sun, and et al, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017, pp. 1077–1085.

[18] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *CVPR*, 2018, pp. 420–429.

[19] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *CVPR*, 2018, pp. 5098–5107.

[20] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Vrstc: Occlusion-free video person re-identification," in *CVPR*, 2019, pp. 7183–7192.

[21] A. Wu, W.-s. Zheng, H.-X. Yu, S. Gong, and J. Lai, "Rgb-infrared cross-modality person re-identification," in *ICCV*, 2017, pp. 5380–5389.

[22] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *CVPR*, 2018, pp. 1179–1188.

[23] A. Das, R. Panda, and A. K. Roy-Chowdhury, "Continuous adaptation of multi-camera person identification models through sparse non-redundant representative selection," *CVIU*, vol. 156, pp. 66–78, 2017.

[24] N. Martinel, A. Das, C. Micheloni, and A. K. Roy-Chowdhury, "Temporal model adaptation for person re-identification," in *ECCV*, 2016, pp. 858–877.

[25] J. Garcia, N. Martinel, A. Gardel, I. Bravo, G. L. Foresti, and C. Micheloni, "Discriminant context information analysis for post-ranking person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1650–1665, 2017.

[26] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE TPAMI*, vol. 38, no. 3, pp. 591–606, 2015.

[27] A. Das, R. Panda, and A. Roy-Chowdhury, "Active image pair selection for continuous person re-identification," in *ICIP*, 2015, pp. 4263–4267.

[28] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *CVPR*, 2019, pp. 719–728.

[29] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *ECCV*, 2014, pp. 330–345.

[30] Q. Yang, A. Wu, and W. Zheng, "Person re-identification by contour sketch under moderate clothing change." *IEEE TPAMI*, 2019.

[31] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *ECCV*, 2008, pp. 262–275.

[32] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010, pp. 2360–2367.

[33] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *ECCV*, 2014, pp. 536–551.

[34] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015, pp. 2197–2206.

[35] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *CVPR*, 2016, pp. 1363–1372.

[36] M. Kostinger, M. Hirzer, P. Wohlhart, and et al, "Large scale metric learning from equivalence constraints," in *CVPR*, 2012, pp. 2288–2295.

[37] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR*, 2011, pp. 649–656.

[38] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *ECCV*, 2014, pp. 1–16.

[39] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *ECCV*, 2012, pp. 780–793.

[40] S. Liao and S. Z. Li, "Efficient psd constrained asymmetric metric learning for person re-identification," in *ICCV*, 2015, pp. 3685–3693.

[41] H.-X. Yu, A. Wu, and W.-S. Zheng, "Unsupervised person re-identification by deep asymmetric metric embedding," *IEEE TPAMI*, 2018.

[42] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017, pp. 3754–3762.

[43] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014, pp. 152–159.

[44] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *CVPR*, 2018, pp. 79–88.

[45] Q. Leng, M. Ye, and Q. Tian, "A survey of open-world person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2019.

[46] D. Wu, S.-J. Zheng, X.-P. Zhang, C.-A. Yuan, F. Cheng, Y. Zhao, Y.-J. Lin, Z.-Q. Zhao, Y.-L. Jiang, and D.-S. Huang, "Deep learning-based methods for person re-identification: A comprehensive review," *Neurocomputing*, 2019.

[47] B. Lavi, M. F. Serj, and I. Ullah, "Survey on deep learning techniques for person re-identification task," *arXiv preprint arXiv:1807.05284*, 2018.

[48] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern recognition letters*, vol. 34, no. 1, pp. 3–19, 2013.

[49] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE TPAMI*, no. 7, pp. 1239–1258, 2009.

[50] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *CVPR*, 2009, pp. 304–311.

[51] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele, "Arttrack: Articulated multi-person tracking in the wild," in *CVPR*, 2017, pp. 6457–6465.

[52] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *CVPR*, 2018, pp. 6036–6046.

[53] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *ICCV*, 2013, pp. 3567–3574.

[54] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *CVPR*, 2016, pp. 1249–1258.

[55] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *CVPR*, 2017, pp. 1367–1376.

[56] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in *ICPR*, 2014, pp. 34–39.

[57] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[58] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017, pp. 1318–1327.

[59] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Transactions on Multimedia (TMM)*, vol. 18, no. 12, pp. 2553–2566, 2016.

[60] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.

[61] W.-H. Li, Z. Zhong, and W.-S. Zheng, "One-pass person re-identification by sketch online discriminant analysis," *Pattern Recognition*, vol. 93, pp. 237–250, 2019.

[62] A. Wu, W.-S. Zheng, and J.-H. Lai, "Robust depth-based person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 6, pp. 2588–2603, 2017.

[63] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *CVPR*, 2017, pp. 1345–1353.

[64] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *CVPR*, 2017, pp. 3415–3424.

[65] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *CVPR*, 2014, pp. 3550–3557.

[66] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013, pp. 3586–3593.

[67] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *CVPR*, 2019, pp. 393–402.

[68] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *ICCV*, 2007, pp. 1–8.

[69] H. Wang, X. Zhu, T. Xiang, and S. Gong, "Towards unsupervised open-set person re-identification," in *ICIP*, 2016, pp. 769–773.

[70] X. Zhu, B. Wu, D. Huang, and W.-S. Zheng, "Fast open-world person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 5, pp. 2286 – 2300, 2018.

[71] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *ECCV*, 2016, pp. 475–491.

[72] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang, "Improving person re-identification by attribute and identity learning," *arXiv preprint arXiv:1703.07220*, 2017.

[73] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in *ICCV*, 2015, pp. 3810–3818.

[74] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, "Video person re-identification by temporal residual learning," *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 3, pp. 1366–1377, 2018.

[75] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *CVPR*, 2017, pp. 3219–3228.

[76] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Transactions on Image Processing (TIP)*, 2019.

[77] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *ECCV*, 2018, pp. 480–496.

[78] T. Matsukawa and E. Suzuki, "Person re-identification using cnn features learned from combination of attributes," in *ICPR*, 2016, pp. 2428–2433.

[79] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[80] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[81] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *CVPR*, 2016, pp. 1288–1296.

[82] Y. Sun, L. Zheng, W. Deng, and S. Wang, "Svdnet for pedestrian retrieval," in *ICCV*, 2017, pp. 3800–3808.

[83] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *ECCV*, 2018, pp. 170–186.

[84] X. Qian, Y. Fu, Y.-G. Jiang, T. Xiang, and X. Xue, "Multi-scale deep learning architectures for person re-identification," in *ICCV*, 2017, pp. 5399–5408.

[85] F. Yang, K. Yan, S. Lu, H. Jia, X. Xie, and W. Gao, "Attention driven person re-identification," *Pattern Recognition*, vol. 86, pp. 143–155, 2019.

[86] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *CVPR*, 2018, pp. 2285–2294.

[87] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *ECCV*, 2018, pp. 365–381.

[88] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep kronecker-product matching for person re-identification," in *CVPR*, 2018, pp. 6886–6895.

[89] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in *CVPR*, 2018, pp. 1470–1478.

[90] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou, "Self-critical attention learning for person re-identification," in *ICCV*, 2019, pp. 9637–9646.

[91] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *CVPR*, 2018, pp. 5363–5372.

[92] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *CVPR*, 2019, pp. 5735–5744.

[93] S. Zhou, F. Wang, Z. Huang, and J. Wang, "Discriminative feature learning with consistent attention regularization for person re-identification," in *ICCV*, 2019, pp. 8040–8049.

[94] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep crf for person re-identification," in *CVPR*, 2018, pp. 8649–8658.

[95] C. Luo, Y. Chen, N. Wang, and Z. Zhang, "Spectral feature transformation for person re-identification," in *ICCV*, 2019, pp. 4976–4985.

[96] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *ECCV*, 2016, pp. 135–153.

[97] Y. Suh, J. Wang, S. Tang, T. Mei, and K. Mu Lee, "Part-aligned bilinear representations for person re-identification," in *ECCV*, 2018, pp. 402–419.

[98] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *ICCV*, 2017, pp. 3219–3228.

[99] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *CVPR*, 2016, pp. 1335–1344.

[100] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *CVPR*, 2017, pp. 384–393.

[101] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *ICCV*, 2017, pp. 3960–3969.

[102] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *CVPR*, 2018, pp. 2119–2128.

[103] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *CVPR*, 2019, pp. 667–676.

[104] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han, "Beyond human parts: Dual part-aligned representations for person re-identification," in *ICCV*, 2019, pp. 3642–3651.

[105] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *CVPR*, 2019, pp. 608–617.

[106] Z. Zhong, L. Zheng, Z. Luo, S. Li, and Y. Yang, "Invariance matters: Exemplar memory for domain adaptive person re-identification," in *CVPR*, 2019, pp. 598–607.

[107] B. N. Xia, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *ICCV*, 2019, pp. 3760–3769.

[108] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *CVPR*, 2019, pp. 9317–9326.

[109] C.-P. Tay, S. Roy, and K.-H. Yap, "Aanet: Attribute attention network for person re-identifications," in *CVPR*, 2019, pp. 7134–7143.

[110] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-s. Hua, "Attribute-driven feature disentangling and temporal aggregation for video person re-identification," in *CVPR*, 2019, pp. 4913–4922.

[111] W. Jingya, Z. Xiatian, G. Shaogang, and L. Wei, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *CVPR*, 2018, pp. 2275–2284.

[112] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *CVPR*, 2018, pp. 2109–2118.

[113] F. Liu and L. Zhang, "View confusion feature learning for person re-identification," in *ICCV*, 2019, pp. 6639–6648.

[114] Z. Zhu, X. Jiang, F. Zheng, X. Guo, F. Huang, W. Zheng, and X. Sun, "Aware loss with angular regularization for person re-identification," in *AAAI*, 2020.

[115] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou, "Consistent-aware deep learning for person re-identification in a camera network," in *CVPR*, 2017, pp. 5771–5780.

[116] J. Liu, B. Ni, Y. Yan, and et al., "Pose transferrable person re-identification," in *CVPR*, 2018, pp. 4099–4108.

[117] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *ECCV*, 2018, pp. 650–667.

[118] Z. Zhong, L. Zheng, Z. Zheng, and et al., "Camera style adaptation for person re-identification," in *CVPR*, 2018, pp. 5157–5166.

[119] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *CVPR*, 2019, pp. 2138–2147.

[120] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *CVPR*, 2018, pp. 994–1003.

[121] Y.-J. Li, C.-S. Lin, Y.-B. Lin, and Y.-C. F. Wang, "Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation," in *ICCV*, 2019, pp. 7919–7929.

[122] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normneuralization neck for deep person re-identification," *arXiv preprint arXiv:1906.08332*, 2019.

[123] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," *arXiv preprint arXiv:1708.04896*, 2017.

[124] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch dropblock network for person re-identification and beyond," in *ICCV*, 2019, pp. 3691–3701.

[125] S. Bak, P. Carr, and J.-F. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification," in *ECCV*, 2018, pp. 189–205.

[126] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Image Analysis*, 2011, pp. 91–102.

[127] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *CVPR*, 2016, pp. 1325–1334.

[128] D. Chung, K. Tahboub, and E. J. Delp, "A two stream siamese convolutional neural network for person re-identification," in *ICCV*, 2017, pp. 1983–1991.

[129] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *ECCV*, 2016, pp. 701–716.

[130] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *CVPR*, 2017, pp. 4747–4756.

[131] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *ICCV*, 2017, pp. 4733–4742.

[132] A. Subramaniam, A. Nambiar, and A. Mittal, "Co-segmentation inspired attention networks for video-based person re-identification," in *ICCV*, 2019, pp. 562–572.

[133] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *CVPR*, 2018, pp. 369–378.

[134] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *CVPR*, 2018, pp. 1169–1178.

[135] Y. Fu, X. Wang, Y. Wei, and T. Huang, "Sta: Spatial-temporal attention for large-scale video-based person re-identification," in *AAAI*, 2019.

[136] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representations for video person re-identification," in *ICCV*, 2019, pp. 3958–3967.

[137] Y. Guo and N.-M. Cheung, "Efficient and deep person re-identification using multi-level similarity," in *CVPR*, 2018, pp. 2335–2344.

[138] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *ICCV*, 2019, pp. 3702–3712.

[139] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto-reid: Searching for a part-aware convnet for person re-identification," in *ICCV*, 2019, pp. 3750–3759.

[140] M. Tian, S. Yi, H. Li, S. Li, X. Zhang, J. Shi, J. Yan, and X. Wang, "Eliminating background-bias for robust person re-identification," in *CVPR*, 2018, pp. 5794–5803.

[141] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned cnn embedding for person re-identification," *arXiv preprint arXiv:1611.05666*, 2016.

[142] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 15, pp. 407–419, 2020.

[143] P. Moutafis, M. Leng, and I. A. Kakadiaris, "An overview and empirical comparison of distance metric learning methods," *IEEE Transactions on Cybernetics*, vol. 47, no. 3, pp. 612–625, 2016.

[144] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning," in *CVPR*, 2018, pp. 5177–5186.

[145] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *CVPR*, 2019, pp. 6210–6219.

[146] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *WACV*, 2018, pp. 748–756.

[147] X. Fan, W. Jiang, H. Luo, and M. Fei, “Sphereredid: Deep hypersphere manifold embedding for person re-identification,” *JVCIR*, vol. 60, pp. 51–58, 2019.

[148] R. Müller, S. Kornblith, and G. Hinton, “When does label smoothing help?” *arXiv preprint arXiv:1906.02629*, 2019.

[149] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li, “Embedding deep metric for person re-identification: A study against large variations,” in *ECCV*, 2016, pp. 732–748.

[150] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, “Point to set similarity based deep feature learning for person re-identification,” in *CVPR*, 2017, pp. 3741–3750.

[151] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai, “Hard-aware point-to-set deep metric for person re-identification,” in *ECCV*, 2018, pp. 188–204.

[152] W. Chen, X. Chen, J. Zhang, and K. Huang, “Beyond triplet loss: a deep quadruplet network for person re-identification,” in *CVPR*, 2017, pp. 403–412.

[153] Q. Yang, H.-X. Yu, A. Wu, and W.-S. Zheng, “Patch-based discriminative feature learning for unsupervised person re-identification,” in *CVPR*, 2019, pp. 3633–3642.

[154] Z. Liu, J. Wang, S. Gong, H. Lu, and D. Tao, “Deep reinforcement active learning for human-in-the-loop person re-identification,” in *ICCV*, 2019, pp. 6122–6131.

[155] J. Zhou, B. Su, and Y. Wu, “Easy identification from better constraints: Multi-shot person re-identification from reference constraints,” in *CVPR*, 2018, pp. 5373–5381.

[156] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, “Pyramidal person re-identification via multi-loss dynamic training,” in *CVPR*, 2019, pp. 8514–8522.

[157] M. Ye, C. Liang, Z. Wang, Q. Leng, and J. Chen, “Ranking optimization for person re-identification via similarity and dissimilarity,” in *ACM Multimedia (ACM MM)*, 2015, pp. 1239–1242.

[158] C. Liu, C. Change Loy, S. Gong, and G. Wang, “Pop: Person re-identification post-rank optimisation,” in *ICCV*, 2013, pp. 441–448.

[159] H. Wang, S. Gong, X. Zhu, and T. Xiang, “Human-in-the-loop person re-identification,” in *ECCV*, 2016, pp. 405–422.

[160] S. Paisitkriangkrai, C. Shen, and A. Van Den Hengel, “Learning to rank in person re-identification with metric ensembles,” in *CVPR*, 2015, pp. 1846–1855.

[161] S. Bai, P. Tang, P. H. Torr, and L. J. Latecki, “Re-ranking via metric fusion for object retrieval and person re-identification,” in *CVPR*, 2019, pp. 740–749.

[162] S. Bai, X. Bai, and Q. Tian, “Scalable person re-identification on supervised smoothed manifold,” in *CVPR*, 2017, pp. 2530–2539.

[163] A. J. Ma and P. Li, “Query based adaptive re-ranking for person re-identification,” in *ACCV*, 2014, pp. 397–412.

[164] J. Zhou, P. Yu, W. Tang, and Y. Wu, “Efficient online local metric adaptation via negative samples for person re-identification,” in *ICCV*, 2017, pp. 2420–2428.

[165] L. Zheng, S. Wang, L. Tian, F. He, Z. Liu, and Q. Tian, “Query-adaptive late fusion for image search and person re-identification,” in *CVPR*, 2015, pp. 1741–1750.

[166] A. Barman and S. K. Shah, “Shape: A novel graph theoretic algorithm for making consensus-based decisions in person re-identification systems,” in *ICCV*, 2017, pp. 1115–1124.

[167] W.-S. Zheng, S. Gong, and T. Xiang, “Associating groups of people,” in *BMVC*, 2009, pp. 1–23.

[168] C. C. Loy, C. Liu, and S. Gong, “Person re-identification by manifold ranking,” in *ICIP*, 2013, pp. 3567–3571.

[169] M. Gou, Z. Wu, A. Rates-Borras, O. Camps, R. J. Radke *et al.*, “A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets,” *IEEE TPAMI*, vol. 41, no. 3, pp. 523–536, 2018.

[170] M. Li, X. Zhu, and S. Gong, “Unsupervised person re-identification by deep learning tracklet association,” in *ECCV*, 2018, pp. 737–753.

[171] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, “Region-based quality estimation network for large-scale person re-identification,” in *AAAI*, 2018, pp. 7347–7354.

[172] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, “Learning discriminative features with multiple granularities for person re-identification,” in *ACM MM*, 2018, pp. 274–282.

[173] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, “Abd-net: Attentive but diverse person re-identification,” in *ICCV*, 2019, pp. 8351–8361.

[174] B. Chen, W. Deng, and J. Hu, “Mixed high-order attention network for person re-identification,” in *ICCV*, 2019, pp. 371–381.

[175] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, “Alignedreid: Surpassing human-level performance in person re-identification,” *arXiv preprint arXiv:1711.08184*, 2017.

[176] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, “Dynamic label graph matching for unsupervised video re-identification,” in *ICCV*, 2017, pp. 5142–5150.

[177] Z. Wang, S. Zheng, M. Song, Q. Wang, A. Rahimpour, and H. Qi, “advpattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns,” in *ICCV*, 2019, pp. 8341–8350.

[178] J. Zhang, N. Wang, and L. Zhang, “Multi-shot pedestrian re-identification via sequential decision making,” in *CVPR*, 2018, pp. 6781–6789.

[179] A. Haque, A. Alahi, and L. Fei-Fei, “Recurrent attention models for depth-based person identification,” in *CVPR*, 2016, pp. 1229–1238.

[180] N. Karianakis, Z. Liu, Y. Chen, and S. Soatto, “Reinforced temporal attention and split-rate transfer for depth-based person re-identification,” in *ECCV*, 2018, pp. 715–733.

[181] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, “Re-identification with rgb-d sensors,” in *ECCV Workshop*, 2012, pp. 433–442.

[182] D. Chen, H. Li, X. Liu, Y. Shen, J. Shao, Z. Yuan, and X. Wang, “Improving deep visual representation for person re-identification by global and local image-language association,” in *ECCV*, 2018, pp. 54–70.

[183] Y. Zhang and H. Lu, “Deep cross-modal projection learning for image-text matching,” in *ECCV*, 2018, pp. 686–701.

[184] J. Liu, Z.-J. Zha, R. Hong, M. Wang, and Y. Zhang, “Deep adversarial graph attention convolution network for text-based person search,” in *ACM MM*, 2019, pp. 665–673.

[185] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, “Visible thermal person re-identification via dual-constrained top-ranking,” in *IJCAI*, 2018, pp. 1092–1099.

[186] M. Ye, X. Lan, J. Li, and P. C. Yuen, “Hierarchical discriminative learning for visible thermal person re-identification,” in *AAAI*, 2018, pp. 7501–7508.

[187] Y. Hao, N. Wang, J. Li, and X. Gao, “Hsme: Hypersphere manifold embedding for visible thermal person re-identification,” in *AAAI*, 2019, pp. 8385–8392.

[188] M. Ye, J. Shen, and L. Shao, “Visible-infrared person re-identification via homogeneous augmented tri-modal learning,” *IEEE TIFS*, 2020.

[189] Z. Wang, Z. Wang, Y. Zheng, Y.-Y. Chuang, and S. Satoh, “Learning to reduce dual-level discrepancy for infrared-visible person re-identification,” in *CVPR*, 2019, pp. 618–626.

[190] G. Wang, T. Zhang, J. Cheng, S. Liu, Y. Yang, and Z. Hou, “Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment,” in *ICCV*, 2019, pp. 3623–3632.

[191] S. Choi, S. Lee, Y. Kim, T. Kim, and C. Kim, “Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification,” in *CVPR*, 2020, pp. 10 257–10 266.

[192] M. Ye, J. Shen, D. J. Crandall, L. Shao, and J. Luo, “Dynamic dual-attentive aggregation learning for visible-infrared person re-identification,” in *ECCV*, 2020.

[193] Z. Wang, M. Ye, F. Yang, X. Bai, and S. Satoh, “Cascaded sr-gan for scale-adaptive low resolution person re-identification.” in *IJCAI*, 2018, pp. 3891–3897.

[194] Y.-J. Li, Y.-C. Chen, Y.-Y. Lin, X. Du, and Y.-C. F. Wang, “Recover and identify: A generative dual model for cross-resolution person re-identification,” in *ICCV*, 2019, pp. 8090–8099.

[195] H. Liu, J. Feng, Z. Jie, K. Jayashree, B. Zhao, M. Qi, J. Jiang, and S. Yan, “Neural person search machines,” in *ICCV*, 2017, pp. 493–501.

[196] Y. Yan, Q. Zhang, B. Ni, W. Zhang, M. Xu, and X. Yang, “Learning context graph for person search,” in *CVPR*, 2019, pp. 2158–2167.

[197] B. Munjal, S. Amin, F. Tombari, and F. Galasso, “Query-guided end-to-end person search,” in *CVPR*, 2019, pp. 811–820.

[198] C. Han, J. Ye, Y. Zhong, X. Tan, C. Zhang, C. Gao, and N. Sang, “Re-id driven localization refinement for person search,” in *ICCV*, 2019, pp. 9814–9823.

[199] X. Lan, H. Wang, S. Gong, and X. Zhu, “Deep reinforcement learning attention selection for person re-identification,” in *BMVC*, 2017.

[200] M. Yamaguchi, K. Saito, Y. Ushiku, and T. Harada, "Spatio-temporal person retrieval via natural language queries," in *ICCV*, 2017, pp. 1453–1462.

[201] S. Tang, M. Andriluka, B. Andres, and B. Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *CVPR*, 2017, pp. 3539–3548.

[202] Y. Hou, L. Zheng, Z. Wang, and S. Wang, "Locality aware appearance metric for multi-target multi-camera tracking," *arXiv preprint arXiv:1911.12037*, 2019.

[203] E. Kodirov, T. Xiang, Z. Fu, and S. Gong, "Person re-identification by unsupervised l1 graph learning," in *ECCV*, 2016, pp. 178–195.

[204] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *ICCV*, 2017, pp. 2429–2438.

[205] H. Fan, L. Zheng, and Y. Yang, "Unsupervised person re-identification: Clustering and fine-tuning," *arXiv preprint arXiv:1705.10444*, 2017.

[206] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE TIP*, vol. 28, no. 6, pp. 2976–2990, 2019.

[207] X. Wang, R. Panda, M. Liu, Y. Wang, and et al., "Exploiting global camera network constraints for unsupervised video person re-identification," *arXiv preprint arXiv:1908.10486*, 2019.

[208] K. Zeng, M. Ning, Y. Wang, and Y. Guo, "Hierarchical clustering with hard-batch triplet loss for person re-identification," in *CVPR*, 2020, pp. 13 657–13 665.

[209] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised person re-identification by soft multilabel learning," in *CVPR*, 2019, pp. 2148–2157.

[210] A. Wu, W.-S. Zheng, and J.-H. Lai, "Unsupervised person re-identification by camera-aware similarity consistency learning," in *ICCV*, 2019, pp. 6922–6931.

[211] J. Wu, Y. Yang, H. Liu, S. Liao, Z. Lei, and S. Z. Li, "Unsupervised graph association for person re-identification," in *ICCV*, 2019, pp. 8321–8330.

[212] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, and T. S. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *ICCV*, 2019, pp. 6112–6121.

[213] S. Bak and P. Carr, "One-shot metric learning for person re-identification," in *CVPR*, 2017, pp. 2990–2999.

[214] X. Wang, S. Paul, D. S. Raychaudhuri, and at al., "Learning person re-identification models from videos with weak supervision," *arXiv preprint arXiv:2007.10631*, 2020.

[215] Z. Zhong, L. Zheng, S. Li, and Y. Yang, "Generalizing a person retrieval model hetero-and homogeneously," in *ECCV*, 2018, pp. 172–188.

[216] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *CVPR*, 2019, pp. 7202–7211.

[217] Y. Huang, Q. Wu, J. Xu, and Y. Zhong, "Sbsgan: Suppression of inter-domain background shift for person re-identification," in *ICCV*, 2019, pp. 9527–9536.

[218] Y. Chen, X. Zhu, and S. Gong, "Instance-guided context rendering for cross-domain person re-identification," in *ICCV*, 2019, pp. 232–242.

[219] Y. Ge, D. Chen, and H. Li, "Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification," in *ICLR*, 2020.

[220] Y. Wang, S. Liao, and L. Shao, "Surpassing real-world source training data: Random 3d characters for generalizable person re-identification," in *ACM MM*, 2020, pp. 3422–3430.

[221] L. Qi, L. Wang, J. Huo, L. Zhou, Y. Shi, and Y. Gao, "A novel unsupervised camera-aware domain adaptation framework for person re-identification," in *ICCV*, 2019, pp. 8080–8089.

[222] X. Zhang, J. Cao, C. Shen, and M. You, "Self-training with progressive augmentation for unsupervised cross-domain person re-identification," in *ICCV*, 2019, pp. 8222–8231.

[223] Y. Ge, F. Zhu, D. Chen, R. Zhao, and H. Li, "Self-paced contrastive learning with hybrid memory for domain adaptive object re-id," in *NeurIPS*, 2020.

[224] J. Lv, W. Chen, Q. Li, and C. Yang, "Unsupervised cross-dataset person re-identification by transfer learning of spatial-temporal patterns," in *CVPR*, 2018, pp. 7948–7956.

[225] S. Liao and L. Shao, "Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting," in *ECCV*, 2020.

[226] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *ICCV*, 2017, pp. 994–1002.

[227] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *CVPR*, 2020, pp. 3143–3152.

[228] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji, and Y. Tian, "Multiple expert brainstorming for domain adaptive person re-identification," in *ECCV*, 2020.

[229] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.

[230] M. Ye, J. Shen, X. Zhang, P. C. Yuen, and S.-F. Chang, "Augmentation invariant and instance spreading feature for softmax embedding," *IEEE TPAMI*, 2020.

[231] W.-S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," in *ICCV*, 2015, pp. 4678–4686.

[232] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *CVPR*, 2018, pp. 7073–7082.

[233] L. He, Y. Wang, W. Liu, X. Liao, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *ICCV*, 2019, pp. 8450–8459.

[234] J. Miao, Y. Wu, P. Liu, Y. Ding, and Y. Yang, "Pose-guided feature alignment for occluded person re-identification," in *ICCV*, 2019, pp. 542–551.

[235] T. Yu, D. Li, Y. Yang, T. Hospedales, and T. Xiang, "Robust person re-identification by modelling feature uncertainty," in *ICCV*, 2019, pp. 552–561.

[236] M. Ye and P. C. Yuen, "Purifynet: A robust person re-identification model with noisy labels," *IEEE TIFS*, 2020.

[237] X. Li, A. Wu, and W.-S. Zheng, "Adversarial open-world person re-identification," in *ECCV*, 2018, pp. 280–296.

[238] M. Golfarelli, D. Maio, and D. Malton, "On the error-reject tradeoff in biometric verification systems," *IEEE TPAMI*, vol. 19, no. 7, pp. 786–796, 1997.

[239] G. Lisanti, N. Martinel, A. Del Bimbo, and G. Luca Foresti, "Group re-identification via unsupervised transfer of sparse features encoding," in *ICCV*, 2017, pp. 2449–2458.

[240] Y. Cai, V. Takala, and M. Pietikainen, "Matching groups of people by covariance descriptor," in *ICPR*, 2010, pp. 2744–2747.

[241] H. Xiao, W. Lin, B. Sheng, K. Lu, J. Yan, and et al., "Group re-identification: Leveraging and integrating multi-grain information," in *ACM MM*, 2018, pp. 192–200.

[242] Z. Huang, Z. Wang, W. Hu, C.-W. Lin, and S. Satoh, "Dot-gnn: Domain-transferred graph neural network for group re-identification," in *ACM MM*, 2019, pp. 1888–1896.

[243] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *ECCV*, 2018, pp. 486–504.

[244] R. Panda, A. Bhuiyan, V. Murino, and A. K. Roy-Chowdhury, "Unsupervised adaptive re-identification in open world dynamic camera networks," in *CVPR*, 2017, pp. 7054–7063.

[245] S. M. Assari, H. Idrees, and M. Shah, "Human re-identification in crowd videos using personal, social and environmental constraints," in *ECCV*, 2016, pp. 119–136.

[246] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018, pp. 7794–7803.

[247] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE TPAMI*, vol. 41, no. 7, pp. 1655–1668, 2018.

[248] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *CVPR*, 2019, pp. 5022–5030.

[249] L. He, Z. Sun, Y. Zhu, and Y. Wang, "Recognizing partial biometric patterns." *arXiv preprint arXiv:1810.07399*, 2018.

[250] J. Xue, Z. Meng, K. Katipally, H. Wang, and K. van Zon, "Clothing change aware person identification," in *CVPR Workshops*, 2018, pp. 2112–2120.

[251] F. Wan, Y. Wu, X. Qian, and Y. Fu, "When person re-identification meets changing clothes," *arXiv preprint arXiv:2003.04070*, 2020.

[252] S. Roy, S. Paul, N. E. Young, and A. K. Roy-Chowdhury, "Exploiting transitivity for learning person re-identification models on a budget," in *CVPR*, 2018, pp. 7064–7072.

[253] Z. Zheng and Y. Yang, "Person re-identification in the 3d space," *arXiv preprint arXiv:2006.04569*, 2020.

[254] Y. Hu, D. Yi, S. Liao, Z. Lei, and S. Z. Li, "Cross dataset person re-identification," in *ACCV*, 2014, pp. 650–664.
[255] G. Wu and S. Gong, "Decentralised learning from independent multi-domain labels for person re-identification," *arXiv preprint arXiv:2006.04150*, 2020.
[256] P. Bhargava, "Incremental learning in person re-identification," *arXiv preprint arXiv:1808.06281*, 2018.
[257] F. Zhu, X. Kong, L. Zheng, H. Fu, and Q. Tian, "Part-based deep hashing for large-scale person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 10, pp. 4806–4817, 2017.
[258] J. Chen, Y. Wang, J. Qin, L. Liu, and L. Shao, "Fast person re-identification via cross-camera semantic binary transformation," in *CVPR*, 2017, pp. 3873–3882.
[259] G. Wang, S. Gong, J. Cheng, and Z. Hou, "Faster person re-identification," in *ECCV*, 2020, pp. 275–292.
[260] A. Wu, W.-S. Zheng, X. Guo, and J.-H. Lai, "Distilled person re-identification: Towards a more scalable system," in *CVPR*, 2019, pp. 1187–1196.
[261] M. Ye, X. Lan, and Q. Leng, "Modality-aware collaborative learning for visible thermal person re-identification," in *ACM MM*, 2019, pp. 347–355.
[262] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 579–590, 2020.
[263] D. Li, X. Wei, X. Hong, and Y. Gong, "Infrared-visible cross-modal person re-identification with an x modality." in *AAAI*, 2020, pp. 4610–4617.
[264] Y. Lu, Y. Wu, B. Liu, T. Zhang, B. Li, Q. Chu, and N. Yu, "Cross-modality person re-identification with shared-specific feature transfer," in *CVPR*, 2020, pp. 13 379–13 389.

Supplemental Materials:

This supplementary material accompanies our main manuscript with the implementation details and more comprehensive experiments. We first present the experiments on two single-modality closed-world Re-ID tasks, including image-based Re-ID on four datasets in Section A and video-based Re-ID on four datasets in Section B. Then we introduce the comprehensive comparison on two open-world Re-ID tasks, including visible-infrared cross-modality Re-ID on two datasets in Section C and partial Re-ID on two datasets in Section D. In addition, a structure overview for our survey is finally summarized.

A. Experiments on Single-modality Image-based Re-ID

Architecture Design. The overall structure[4] of our proposed AGW baseline for single-modality Re-ID is illustrated in § 4 (Fig. R1). We adopt ResNet50 pre-trained on ImageNet as our backbone network and change the dimension of the fully connected layer to be consistent with the number of identities in the training dataset. The stride of the last spatial down-sampling operation in the backbone network is changed from 2 to 1. Consequently, the spatial size of the output feature map is changed from 8 × 4 to 16 × 8, when feeding an image of resolution 256 × 128 as input. In our method, we replace the Global Average Pooling in the original ResNet50 with the Generalized-mean (GeM) pooling. The pooling hyper parameter p_k for generalized-mean pooling is initialized as 3.0. A BatchNorm layer, named BNNeck is plugged between the GeM pooling layer and the fully connected layer. The output of the GeM pooling layer is adopted for computing center loss and triplet loss in the training stage, while the feature after BNNeck is used for computing distance between pedestrian images during testing inference stage.

Non-local Attention. The ResNet contains 4 residual stages, i.e. $conv2\_x$, $conv3\_x$, $conv4\_x$ and $conv5\_x$, each containing stacks of bottleneck residual blocks. We inserted five non-local blocks after $conv3\_3$, $conv3\_4$, $conv4\_4$, $conv4\_5$ and $conv4\_6$ respectively. We adopt the Dot Product version of non-local block with a bottleneck of 512 channels in our experiment. For each non-local block, a BatchNorm layer is added right after the last linear layer that represents W_z. The affine parameter of this BatchNorm layer is initialized as zeros to ensure that the non-local block can be inserted into any pre-trained networks while maintaining its initial behavior.

Training Strategy. In the training stage, we randomly sample 16 identities and 4 images for each identity to form a mini-batch of size 64. Each image is resized into 256 × 128 pixels, padding 10 pixels with zero values, and then randomly cropped into 256 × 128 pixels. Random horizontally flipping and random erasing with 0.5 probability respectively are also adopted for data augmentation. Specifically, random erasing augmentation [123] randomly selects a rectangle region with area ratio r_e to the whole image, and erase its pixels with the mean value of the image. Besides, the aspect ratio of this region is randomly initialized between r_1 and r_2. In our method, we set the above hyper-parameter as $0.02 < r_e < 0.4$, $r1 = 0.3$ and $r2 = 3.33$. At last, we normalize the RGB channels of each image with mean 0.485, 0.456, 0.406 and stand deviation 0.229, 0.224, 0.225, respectively, which are the same with settings in [122].

Training Loss. In the training stage, three types of loss are combined for optimization, including identity classification loss ($\mathcal{L}_{id}$), center loss ($\mathcal{L}_{ct}$) and our proposed weighted regularization triplet loss ($\mathcal{L}_{wrt}$).

$$\mathcal{L}_{total} = \mathcal{L}_{id} + \beta_1 \mathcal{L}_{ct} + \beta_2 \mathcal{L}_{wrt}. \tag{R1}$$

The balanced weight of the center loss (β_1) is set to 0.0005 and the one (β_2) of the weighted regularized triplet loss is set to 1.0. Label smoothing is adopted to improve the original identity classification loss, which encourages the model to be less confident during training and prevent overfitting for classification task. Concretely, it changes the one-hot label as follow:

$$q_i = \begin{cases} 1 - \frac{N-1}{N}\varepsilon & \text{if } i = y \\ \varepsilon / N & \text{otherwise} \end{cases}, \tag{R2}$$

where N is the total number of identities, ϵ is a small constant to reduce the confidence for the true identity label y and q_i is treated as a new classification target for training. In our method, we set ϵ to be 0.1.

Optimizer Setting. Adam optimizer with a weight decay 0.0005 is adopted to train our model. The initial learning rate is set as 0.00035 and is decreased by 0.1 at the 40th epoch and 70th epoch, respectively. The model is trained for 120 epochs in total. Besides, a warm-up learning rate scheme is also employed to improve the stability of training process and bootstrap the network for better performance. Specifically, in the first 10 epochs, the learning rate is linearly

[4] https://github.com/mangye16/ReID-Survey

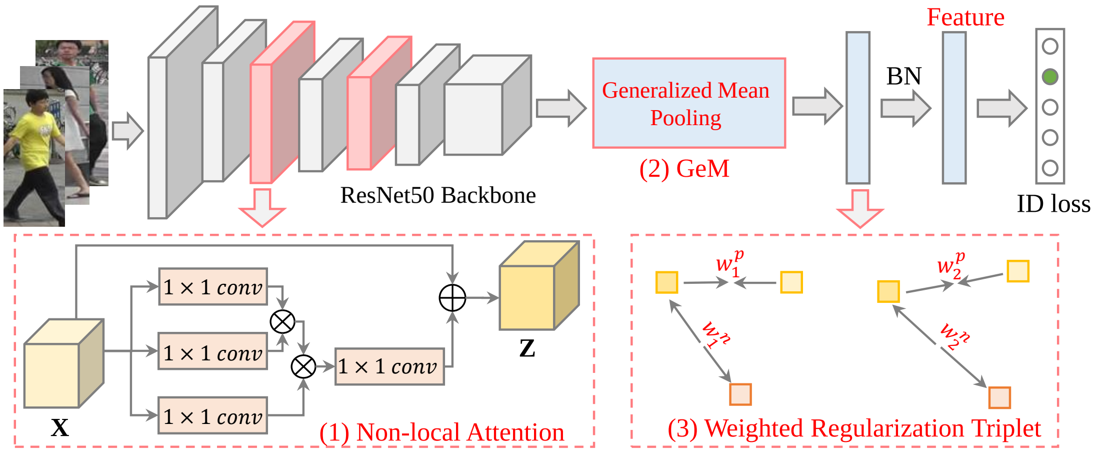


Fig. R1: The framework of the proposed AGW baseline for single-modality image-based Re-ID.

TABLE R1: Comparison with the state-of-the-arts on four video-based Re-ID datasets, including MARS, DukeVideo, PRID2011 and iLIDS-VID. Rank-1, -5, -10 accuracy (%), mAP (%) and mINP (%) are reported.

Method	Venue	MARS				DukeVideo				PRID2011				iLIDS-VID			
		R1	R5	mAP	mINP	R1	R5	mAP	mINP	R1	R5	R20	mINP	R1	R5	R20	mINP
ETAP [144]	CVPR18	80.7	92.0	67.3	-	83.6	94.5	78.3	-	-	-	-	-	-	-	-	-
DRSA [133]	CVPR18	82.3	-	65.8	-	-	-	-	-	93.2	-	-	-	80.2	-	-	-
Snippet [134]	CVPR18	86.3	94.7	76.1	-	-	-	-	-	93.0	99.3	-	-	85.4	96.7	-	-
STA [135]	AAAI18	86.3	95.7	80.8	-	96.2	99.3	94.9	-	-	-	-	-	-	-	-	-
VRSTC [20]	CVPR19	88.5	96.5	82.3	-	95.0	99.1	93.5	-	-	-	-	-	83.4	95.5	99.5	-
ADFD [110]	CVPR19	87.0	95.4	78.2	-	-	-	-	-	93.9	99.5	100	-	86.3	97.4	99.7	-
GLTR [136]	ICCV19	87.0	95.7	78.4	-	96.2	99.3	93.7	-	95.5	100.0	-	-	86.0	98.0	-	-
CoSeg [132]	ICCV19	84.9	95.5	79.9	57.8	95.4	99.3	94.1	89.8	-	-	-	-	79.6	95.3	99.3	-
BagTricks [122]	CVPR19W	85.8	95.2	81.6	62.0	92.6	98.9	92.4	88.3	84.3	93.3	98.0	88.5	74.0	93.3	99.1	82.2
AGW	-	87.0	95.7	82.2	62.8	94.6	99.1	93.4	89.2	87.8	96.6	98.9	91.7	78.0	97.0	99.5	85.5
AGW$_+$	-	87.6	85.8	83.0	63.9	95.4	99.3	94.9	91.9	94.4	98.4	100	95.4	83.2	98.3	99.7	89.0

increased from 3.5×10^{-5} to 3.5×10^{-4}. The learning rate $lr(t)$ at epoch t can be computed as:

$$\mathrm{lr}(t) = \begin{cases} 3.5 \times 10^{-5} \times \frac{t}{10} & \text{if } t \leq 10 \\ 3.5 \times 10^{-4} & \text{if } 10 < t \leq 40 \\ 3.5 \times 10^{-5} & \text{if } 40 < t \leq 70 \\ 3.5 \times 10^{-6} & \text{if } 70 < t \leq 120. \end{cases} \tag{R3}$$

B. Experiments on Video-based Re-ID

Implementation Details. We extend our proposed AGW baseline to a video-based Re-ID model by several minor changes to the backbone structure and training strategy of single-modality image-based Re-ID model. The video-based AGW baseline takes a video sequence as input and extracts the frame-level feature vectors, which are then averaged to be a video-level feature vector before the BNNeck layer. Besides, the video-based AGW baseline is trained for 400 epochs totally to better fit the video person Re-ID datasets. The learning rate is decayed by 10 times every 100 epochs. To form an input video sequence, we adopt the constraint random sampling strategy [133] to sample 4 frames as a summary for the original pedestrian tracklet. The BagTricks [122] baseline is extended to a video-based Re-ID model in the same way as AGW baseline for fair comparison. In addition, we also develop a variant of AGW baseline, termed as AGW$_+$, to model more abundant temporal information in a pedestrian tracklet. AGW$_+$ baseline adopts the dense sampling strategy to form an input video sequence in the testing stage. Dense sampling strategy takes all the frames in a pedestrian tracklet to form input video sequence, resulting better performance but higher computational cost. To further improve the performance of AGW$_+$ baseline on video re-ID datasets, we also remove the warm-up learning rate strategy and add dropout operation before the linear classification layer.

Detailed Comparison. In this section, we conduct the performance comparison between AGW baseline and other state-of-the-art video-based person Re-ID methods, including ETAP [144], DRSA [133], STA [135] Snippet [134], VRSTC [20], ADFD [110], GLTR [136] and CoSeg [132]. The comparison results on four video person Re-ID datasets (MARS, DukeVideo, PRID2011 and iLIDS-VID) are listed in Table R1. As we can see, by simply taking video sequence as input and adopting average pooling to aggregate frame-level feature, our AGW baseline achieves competitive results on two large-scale video Re-ID dataset, MARS and DukeVideo. Besides, AGW baseline also performs significantly better than BagTricks [122] baseline under multiple evaluation metrics. By further modeling more temporal information and adjusting training strategy, AGW$_+$ baseline gains huge improvement and also achieves competitive results on both PRID2011 and iLIDS-VID datasets. AGW$_+$ baseline outperforms most state-of-the-art methods on MARS, DukeVideo and PRID2011 datasets. Most of these video-based person Re-ID methods achieve state-of-the-art performance by designing complicate temporal attention mechanism to exploit temporal dependency in pedestrian video. We believe that our AGW baseline can help video Re-ID model achieve higher performance with properly designed mechanism to further exploit spatial and temporal dependency.

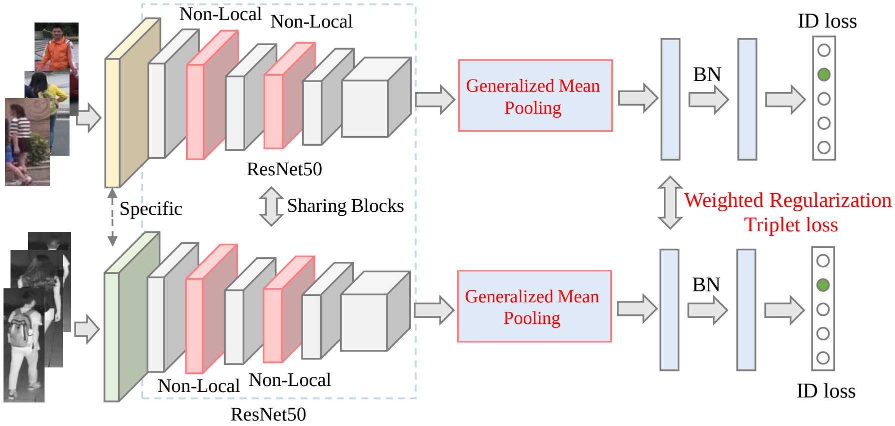


Fig. R2: The framework of the proposed AGW baseline for cross-modality visible-infrared Re-ID.

TABLE R2: Comparison with the state-of-the-arts on SYSU-MM01 dataset. Rank at r accuracy (%), mAP (%) and mINP (%) are reported. (*Single-shot query setting [21] for experiments*). "*" represents methods published after the paper submission.

Settings		*All Search*					*Indoor Search*				
Method	Venue	$r = 1$	$r = 10$	$r = 20$	mAP	mINP	$r = 1$	$r = 10$	$r = 20$	mAP	mINP
One-stream [21]	ICCV17	12.04	49.68	66.74	13.67	-	16.94	63.55	82.10	22.95	-
Two-stream [21]	ICCV17	11.65	47.99	65.50	12.85	-	15.60	61.18	81.02	21.49	-
Zero-Pad [21]	ICCV17	14.80	54.12	71.33	15.95	-	20.58	68.38	85.79	26.92	-
TONE [186]	AAAI18	12.52	50.72	68.60	14.42	-	20.82	68.86	84.46	26.38	-
HCML [186]	AAAI18	14.32	53.16	69.17	16.16	-	24.52	73.25	86.73	30.08	-
BDTR [142]	IJCAI18	27.32	66.96	81.07	27.32	-	31.92	77.18	89.28	41.86	-
eBDTR [142]	TIFS19	27.82	67.34	81.34	28.42	-	32.46	77.42	89.62	42.46	-
HSME [187]	AAAI19	20.68	32.74	77.95	23.12	-	-	-	-	-	-
D^{2}RL [189]	CVPR19	28.9	70.6	82.4	29.2	-	-	-	-	-	-
MAC [261]	MM19	33.26	79.04	90.09	36.22	-	36.43	62.36	71.63	37.03	-
MSR [262]	TIP19	37.35	83.40	93.34	38.11	-	39.64	89.29	97.66	50.88	-
AlignGAN [190]	ICCV19	42.4	85.0	93.7	40.7	-	45.9	87.6	94.4	54.3	-
X-Modal* [263]	AAAI-20	49.9	89.8	96.0	50.7	-	-	-	-	-	-
Hi-CMD* [191]	CVPR20	34.9	77.6	-	35.9	-	-	-	-	-	-
cm-SSFT* [264]	CVPR20	47.7	-	-	54.1	-	-	-	-	-	-
DDAG* [192]	ECCV20	54.75	90.39	95.81	53.02	39.62	61.02	94.06	98.41	67.98	62.61
HAT* [188]	TIFS20	55.29	92.14	97.36	53.89	-	62.10	95.75	99.20	69.37	-
AGW	-	47.50	84.39	92.14	47.65	35.30	54.17	91.14	95.98	62.97	59.23

C. Experiments on Cross-modality Re-ID

Architecture Design. We adopt a two-stream network structure as the backbone for cross-modality visible-infrared Re-ID[5]. Compared to the one-stream architecture in single-modality person Re-ID (Fig. 8), the major difference is that, *i.e.*, the first block is specific for two modalities in order to capture modality-specific information, while the remaining blocks are shared to learn modality sharable features. Compared to the two-stream structure widely used in [142], [261], which only has one shared embedding layer, our design captures more sharable components. An illustration for cross-modality visible-infrared Re-ID is shown in Fig. R2.

Training Strategy. At each training step, we random sample 8 identities from the whole dataset. Then 4 visible and 4 infrared images are randomly selected for each identity. Totally, each training batch contains 32 visible and 32 infrared images. This guarantees the informative hard triplet mining from both modalities, *i.e.*, we directly select the hard positive and negative from both intra- and inter-modalities. This approximates the idea of bi-directional center-constrained top-ranking loss, handling the inter- and intra-modality variations simultaneously.

For fair comparison, we follow the settings in [142] exactly to conduct the image processing and data augmentation. For infrared images, we keep the original three channels, just like the visible RGB images. All the input images from both modalities are first resized to 288 × 144, and random crop with zero-padding together with random horizontal flipping are adopted for data argumentation. The cropped image sizes are 256 × 128 for both modality. The image normalization are exactly following the single-modality setting.

Training Loss. In the training stage, we combine with the identity classification loss ($\mathcal{L}_{id}$) and our proposed weighted regularization triplet loss ($\mathcal{L}_{wrt}$). The weight of combining the identity loss and weighted regularized triplet loss is set to 1, the same as the single-modality setting. The pooling parameter p_k is set to 3. For stable training, we adopt the same identity classifier for two heterogeneous modalities, mining sharable information.

Optimizer Setting. We set the initial learning rate as 0.1 on both datasets, and decay it by 0.1 and 0.01 at 20 and 50 epochs, respectively. The total number of training epoch is 60. We also adopt a warm-up learning rate scheme. We adopt the stochastic gradient descent (SGD) optimizer for optimization, and the momentum parameter is set to 0.9. We have tried the same Adam optimizer (used in single-modality Re-ID) on cross-modality Re-ID task, but the performance is much lower than that of SGD optimizer by using a large learning rate. This is crucial since ImageNet

[5] https://github.com/mangye16/Cross-Modal-Re-ID-baseline

TABLE R3: Comparison with the state-of-the-arts on RegDB dataset on both query settings. Rank at r accuracy (%), mAP (%) and mINP (%) are reported. (*Both the visible to thermal and thermal to visible query settings are evaluated.*) "*" represents methods published after the paper submission.

Settings		*Visible to Thermal*					*Thermal to Visible*				
Method	Venue	$r = 1$	$r = 10$	$r = 20$	mAP	mINP	$r = 1$	$r = 10$	$r = 20$	mAP	mINP
HCML [186]	AAAI18	24.44	47.53	56.78	20.08	-	21.70	45.02	55.58	22.24	-
Zero-Pad [21]	ICCV17	17.75	34.21	44.35	18.90	-	16.63	34.68	44.25	17.82	-
BDTR [142]	IJCAI18	33.56	58.61	67.43	32.76	-	32.92	58.46	68.43	31.96	-
eBDTR [142]	TIFS19	34.62	58.96	68.72	33.46	-	34.21	58.74	68.64	32.49	-
HSME [187]	AAAI19	50.85	73.36	81.66	47.00	-	50.15	72.40	81.07	46.16	-
D^{2}RL [189]	CVPR19	43.4	66.1	76.3	44.1	-	-	-	-	-	-
MAC [261]	MM19	36.43	62.36	71.63	37.03	-	36.20	61.68	70.99	36.63	-
MSR [262]	TIP19	48.43	70.32	79.95	48.67	-	-	-	-	-	-
AlignGAN [190]	ICCV19	57.9	-	-	53.6	-	56.3	-	-	53.4	-
XModal* [263]	AAAI20	62.21	83.13	91.72	60.18	-	-	-	-	-	-
Hi-CMD* [191]	CVPR20	70.93	86.39	-	66.04	-	-	-	-	-	-
cm-SSFT* [264]	CVPR20	72.3	-	-	72.9	-	71.0	-	-	71.7	-
DDAG* [192]	ECCV20	69.34	86.19	91.49	63.46	49.24	68.06	85.15	90.31	61.80	48.62
HAT* [188]	TIFS20	71.83	87.16	92.16	67.56	-	70.02	86.45	91.61	66.30	-
AGW	-	70.05	86.21	91.55	66.37	50.19	70.49	87.12	91.84	65.90	51.24

initialization is adopted for the infrared images.

Detailed Comparison This section conducts the comparison with the state-of-the-art cross-modality VI-ReID methods, including eBDTR [142], HSME [187], D^{2}RL [189], MAC [261], MSR [262] and AlignGAN [190]. These methods are published in the past two years. AlignGAN [190], published in ICCV 2019, achieves the state-of-the-art performance by aligning the cross-modality representation at both the feature level and pixel level with GAN generated images. The results on two datasets are shown in Tables R2 and R3. We observe that the proposed AGW consistently outperforms the current state-of-the-art, without the time-consuming image generation process. For different query settings on RegDB dataset, our proposed baseline generally keeps the same performance. Our proposed baseline has been widely used in many recently developed methods. We believe our new baseline will provide a good guidance to boost the cross-modality Re-ID.

D. Experiments on Partial Re-ID

Implementation Details. We also evaluate the performance of our proposed AGW baseline on two commonly-used partial Re-ID datasets, Partial-REID and Partial-iLIDS. The overall backbone structure and training strategy for partial Re-ID AGW baseline model are the same as the one for single-modality image-based Re-ID model. Both Partial-REID and Partial-iLIDS datasets offer only query image set and gallery image set. So we train AGW baseline model on the training set of Market-1501 dataset, then evaluate its performance on the testing set of two partial Re-ID datasets. We adopt the same way to evaluate the performance of BagTricks [122] baseline on these two partial Re-ID datasets for better comparison and analysis.

Detailed Comparison. We compare the performance of AGW baseline with other state-of-the-art partial Re-ID methods, including DSR [232], SFR [249] and VPM [67]. All these methods are published in recent years. The comparison results on both Partial-REID and Partial-iLIDS datasets are shown in Table R4. The VPM [67] achieves a very high performance by perceiving the visibility of regions through self-supervision and extracting region-level features. Considering only global features, our proposed AGW baseline still achieves competitive results compared to the current state-of-the-arts on both datasets. Besides, AGW baseline brings significant improvement comparing to BagTricks [122] under multiple evaluation metrics, demonstrating its effectiveness for partial Re-ID problem.

TABLE R4: Comparison with the state-of-the-arts on two partial Re-ID datasets, including Partial-REID and Partial-iLIDS. Rank-1, -3 accuracy (%) and mINP (%) are reported.

Method	Partial-REID			Partial-iLIDS		
	R1	R3	mINP	R1	R3	mINP
DSR [232] CVPR18	50.7	70.0	-	58.8	67.2	-
SFR [249] ArXiv18	56.9	78.5	-	63.9	74.8	-
VPM [67] CVPR19	67.7	**81.9**	-	**67.2**	76.5	-
BagTricks [122] CVPR19W	62.0	74.0	45.4	58.8	73.9	68.7
AGW	**69.7**	80.0	**56.7**	64.7	**79.8**	**73.3**

E. Overview of This Survey

The overview figure of this survey is shown in Fig. R3. According to the five steps in developing a person Re-ID system, we conduct the survey from both closed-world and open-world settings. The closed-world setting is detailed in three different aspects: feature representation learning, deep metric learning and ranking optimization. We then summarize the datasets and SOTAs from both image- and video-based perspectives. For open-world person Re-ID, we summarize it into five aspects: including heterogeneous data, Re-ID from raw images/videos, unavailable/limited labels, noisy annotation and open-set Re-ID.

Following the summary, we present an outlook for future person Re-ID. We design a new evaluation metric (mINP) to evaluate the difficulty to find all the correct matches. By analyzing the advantages of existing Re-ID methods, we develop a strong AGW baseline for future developments, which achieves competitive performance on four Re-ID tasks. Finally, some under-investigated open issues are discussed. Our survey provides a comprehensive summarization of existing state-of-the-art in different subtasks. Meanwhile, the analysis of future directions is also presented for further development guidance.

Acknowledgement. The authors would like to thank the anonymous reviewers for providing valuable feedbacks to improve the quality of this survey. The authors also would like to thank the pioneer researchers in person re-identification and other related fields. This work is sponsored by CAAI-Huawei MindSpore Open Fund.

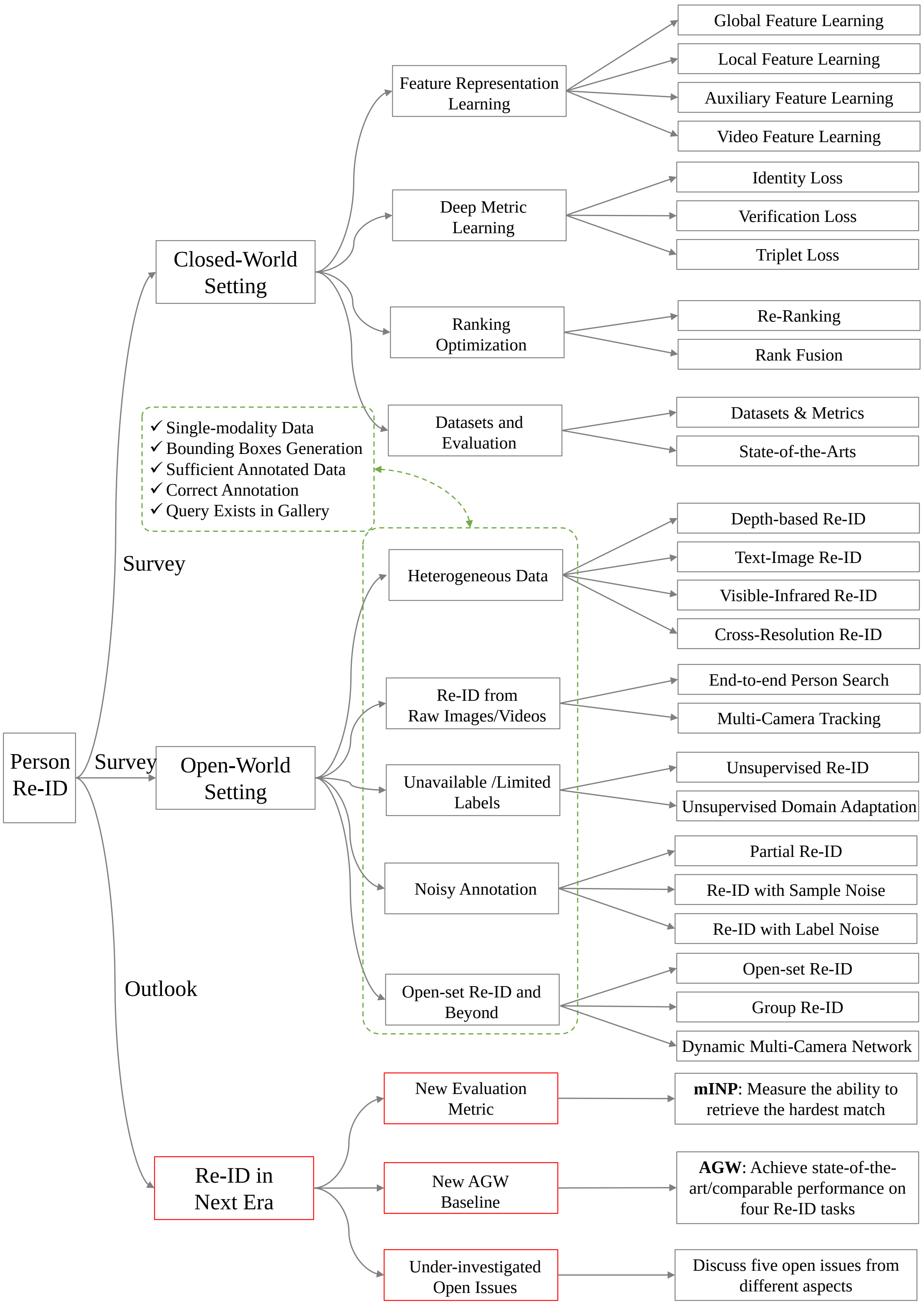


Fig. R3: An overview of this survey. It contains three main components, including Closed-World Setting in Section 2, Open-World Setting in Section 3 and an outlook of Re-ID in Next Era in Section 4.