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 both single- and cross-modality 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

P ERSON 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 [3]. The query person can be represented by an image [4], [5], a video sequence [6], [7], and even a text description [8], [9]. Due to the urgent demand of public safety and increasing number of surveillance cameras in university campuses, theme parks, streets, etc, person Re-ID is imperative in intelligent video surveillance system designs. Given its research impact and practical importance, Re-ID is a fast growing vision community.

Person Re-ID is a challenging task due to the presence of different viewpoints [10], [11], varying low-image resolutions [12], [13], illumination changes [14], unconstrained poses [15], [16], [17], occlusions [18], [19], heterogeneous modalities [9], [20], etc. Early research efforts mainly focus on the hand-crafted feature construction with body structures [4], [21], [22], [23], [24] or distance metric learning [25], [26], [27], [28], [29], [30]. A comprehensive survey on person Re-ID before the deep learning era is presented in [2], [31]. With the advancement of deep learning [32], person Re-

ID has achieved inspiring performance on the widely used benchmarks [5], [33], [34], [35]. However, there is still a large gap between the research-oriented scenarios and practical applications [36]. This motivates us to conduct a comprehensive survey, and discuss several future directions.

Several surveys on person Re-ID already exist, most of which are concentrated in hand-crafted systems [1], [2], [31], though some surveys have also summarized the deep learning techniques [37], [38]. 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, rather than a simple overview. This provides insights for future algorithm design and new topic exploration. 2) We design a new powerful AGW baseline and a new evaluation metric (mINP) for future developments. AGW achieves state-of-the-art performance on both single- and cross-modality 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.

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 environ-

- *M. Ye, 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*
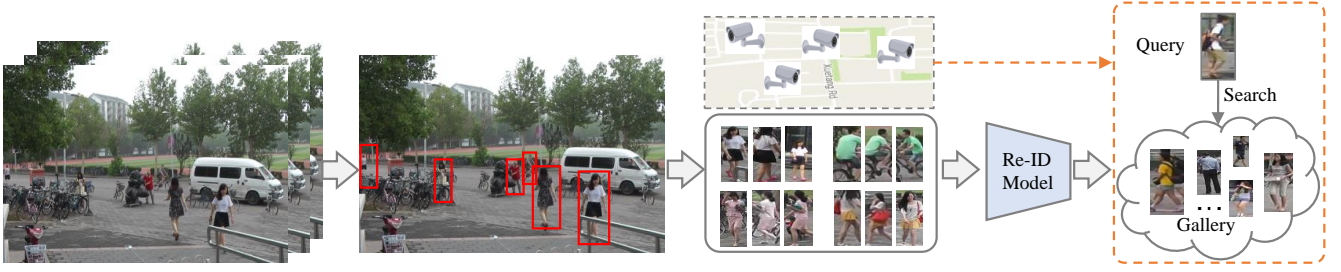- *Corresponding author: Jianbing Shen*

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)

ments [39]. 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 [40], [41] or tracking algorithms [42], [43].

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 [44], 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 [45], [46], distance metric learning [47], [48] 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 [49], [50].

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 [4], [5], [7], [33], [34], [35]. However, in practical open-world applications, the data might be heterogeneous, *e.g.*, the person images are captured by cameras with different light spectrums [20], [51], sketches [52] or depth images

[53], and even text descriptions [54]. This motivates the open-world heterogeneous Re-ID, discussed 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 [46], [55]. 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) [56] or even without any label information [57]. This inspires the discussion of the unsupervised and semi-supervised Re-ID in § 3.3.

4) *Correct Annotation vs. Noisy Annotation*: For the model training in 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). 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 [58] and mAP [5]. However, in many scenarios, the query person may not appear in the gallery set [59], [60], or we need to perform the verification rather than retrieval [61]. 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, including feature representation learning in § 2.1, deep metric learning in § 2.2 and ranking optimization in § 2.3. 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 also 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.
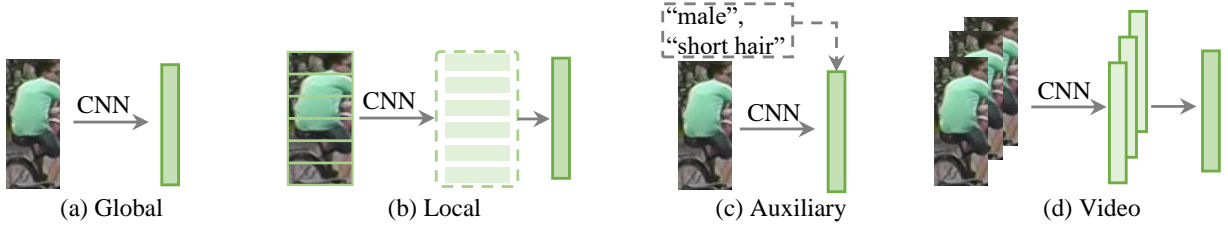
Fig. 2: Four different feature learning strategies. a) Global Feature, learning a global representation for each person image in § 2.1.1; b) Local Feature, learning part-aggregated local features in § 2.1.2; c) Auxiliary Feature, learning the feature representation using auxiliary information, *e.g.*, attributes [62], [63] in § 2.1.3 and d) Video Feature , learning the video representation using multiple image frames and temporal information [64], [65] in § 2.1.4.

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 [46]; b) Local Feature (§ 2.1.2), it aggregates part-level local features to formulate a combined representation for each person image [66], [67], [68]; c) Auxiliary Feature (§ 2.1.3), it improves the feature representation learning using auxiliary information, *e.g.*, attributes [62], [63], [69], GAN generated images [33], etc. d) Video Feature (§ 2.1.4), it learns video representation for video-based Re-ID [6] using multiple image frames and temporal information [64], [65]. 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 [32], [70], [71], [72], 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, Wu *et al.* [73] design a "PersonNet" using small-sized convolutional filters to capture the fine-grained cues in the person image. A joint learning framework consisting of a single-image representation (SIR) and cross-image representation (CIR) is developed in [74], trained with triplet loss using specific sub-networks. An ID-discriminative Embedding (IDE) model is presented in [46], which views the training process of person Re-ID as a multi-class classification problem by treating each identity as a distinct class. It is now widely used in Re-ID community [33], [49], [68], [75], [76]. Qian *et al.* [77] develop a multi-scale deep representation learning model to capture discriminative cues at different scales and adaptively mine the appropriate scale for person retrieval. Human semantic parsing technique is also applied in [78] to capture the pixel-level discriminative cues and enhance the robustness against pose variations.

Attention Information. Attention schemes have been widely studied in literature to enhance representation learning. 1) *For attention within the person image:* Harmonious Attention CNN (HA-CNN) model [79] jointly learns the soft pixel attention and hard regional attention to enhance the robustness of feature representation against misalignment. A Fully Attentional Block (FAB) is designed in [80] to adjust the channel-wise feature response. FAB can be seamlessly applied in different CNN architectures to improve the discriminability. A Kronecker Product Matching (KPM) module [81] aligns the feature maps with a soft warping scheme in an end-to-end trainable manner. The multi-scale feature learning and self residual attention are employed to improve the performance. The BraidNet [82] also tries to align the feature maps with a well-designed convolutional structure. A self-critical reinforcement learning with spatial- and channel-wise attention is introduced in [83]. A Mask-Guided Contrastive Attention Model (MGCAM) [84] is designed to eliminate the impact of background clutter, trained with a region-level triplet loss. 2) *For attention across multiple person images:* A context-aware attentive feature learning method is proposed in [85], incorporating both an intra-sequence and inter-sequence attention for pair-wise feature alignment and refinement. Since the learning strategy does not rely on temporal information, multiple images in image-based Re-ID can also be formulated as a sequence. A Siamese network with attention consistency is designed in [86] to facilitate the feature learning. The attention consistency across multiple images is also considered in [87]. Group similarity [88] is another popular approach to leverage the cross-image attention. It involves multiple images to model the local and global similarities at image-level under a unified Conditional Random Field (CRF) framework for feature learning. The spectral feature transformation [89] also adopts the group similarity for training.

Architecture Modification. An improved bottleneck layer utilizes the orthogonality constraint to reinforce global feature representation learning, achieved by Singular Vector Decomposition (SVDNet) [75]. Using a simple architecture with standard networks, embeddings from multiple layers are aggregated into a single embedding to enhance the final representation [13]. A Class Activation Maps (CAM) augmentation model [90] expands the activation scope to explore rich visual cues in a multi-branch network.

2.1.2 Local Feature Representation Learning

Local feature representation usually learns part/region aggregated features, making it robust against misalignment variations [68], [91]. The body parts are either generated by human pose estimation or roughly horizontal division.

The main trend is to combine the full body representation and local part features. Cheng *et al.* [92] design a multi-channel parts-aggregated deep convolutional network by integrating the local body part features and the global full-body features in a triplet training framework. Similarly, the Multi-Scale Context Aware Network (MSCAN) [93] captures the local context knowledge among the body parts by stacking multi-scale convolutions. Along this line, a multi-stage feature decomposition and selective tree-structured fusion framework [16] is proposed to capture the macro- and micro-body features. Similarly, Zhao *et al.* [94] decompose the human body into local regions (parts), where the part level matching is then performed for re-identification. The major difference is that the part-level similarity scores are aggregated rather than the feature level aggregation [68], [93]. A two-stream network [95] is proposed to extract global appearance and local body part feature maps respectively, and a bilinear-pooling layer is designed to aggregate the two streams to obtain an enhanced representation.

Some works have also studied the robustness of part-level feature learning against background clutter. A Pose-driven Deep Convolutional (PDC) model is presented in [96] to leverage the human part cues for robust representation learning, addressing the pose variations. Along this line, an attention-aware compositional network [97] develops a pose-guided part attention module to mask out undesirable background features. An attention-aware feature composition module is also designed to aggregate the part level features. Contemporarily, a person-region guided pooling deep neural network [98] employs human parsing to address the background bias. Similarly, a two-stream network with densely semantically aligned part level feature learning is introduced in [99], including one stream for full image representation learning and one stream for densely semantically-aligned part feature learning. Both the accurately detected human parts and coarse non-human parts are aligned in [100] to enhance the robustness.

For specific part attention design, a high-order polynomial predictor [101] produces scale maps that contain the high-order statistics of convolutional activations to capture the subtle discriminative features. This is different from the spatial and channel attention. Similarly, second-order non-local attention [102] is introduced to directly model long-range relationships. An Interaction-and-Aggregation (IA) [103] models the inter-dependencies between spatial features and aggregates the correlated body part features.

For horizontal region features without pose estimation, a Siamese Long Short-Term Memory (LSTM) architecture is introduced in [91] to adaptively aggregate the horizontal region features, which enables spatial dependency mining and contextual information propagating to enhance the discriminability of the aggregated region features. A strong Part-based Convolutional Baseline (PCB) [68] is designed with a uniform partition strategy to learn the part features with multiple classifiers. The performance is further improved with a refined part pooling strategy to enhance within-part consistency. It has served as a strong part feature learning baseline in the current state-of-the-art [104], [105], [106].

Discussion. The first group uses human parsing techniques to obtain semantically meaningful body parts, which provides well-align part features. However, they usually require an additional pose detector and are prone to noisy pose detections due to the large gap between the person Re-ID and human pose estimation datasets [68]. 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 [62]) or generated/augmented training samples to reinforce the feature representation [18], [33].

Semantic Attributes. A joint identity and attribute learning baseline is introduced in [63]. Su *et al.* [62] propose a deep attribute learning framework by incorporating the predicted semantic attribute information, enhancing the 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 [107]. Semantic attributes are also adopted in [108] for video Re-ID feature representation learning. They are also leveraged as the auxiliary supervision information in unsupervised learning [109].

With the language descriptions for each person image, Cheng *et al.* [110] enforce representation learning by mining global and local image-language associations, constraining the consistency between the visual and linguistic features. This also improves the visual representation learning.

Viewpoint Information. The viewpoint information is also leveraged to enhance the feature representation learning [111], [112]. Multi-Level Factorisation Net (MLFN) [111] also tries to learn the identity-discriminative and view-invariant feature representations at multiple semantic levels. Liu *et al.* [112] extract a view-invariant identity-wise representation with a view confusion feature learning, which is a combination of view-generic and view-specific learning.

Domain Information. A Domain Guided Dropout (DGD) algorithm [45] 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.* [113] 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 [17] to improve the feature representation with camera-specific information modeling.

Generation/Augmentation. This section discusses the use of augmented/GAN generated images as the auxiliary information. Zheng *et al.* [33] 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 [114] 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 [115], which enhances the robustness against pose variations. The generated images greatly reinforce the feature representation learning. Camera style information [116] is also integrated in the image generation process to address the large cross camera variations. A joint discriminative and generative learning model [117] 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 [118], [119], approximating the target distribution.

For augmented auxiliary information, adversarially occluded samples [18] are generated to augment the variation of training data. A similar but much simpler random erasing strategy is proposed in [120], adding random noise to the input images. A batch DropBlock [121] randomly drops a region block in the feature map to reinforce the attentive feature learning. Bak *et al.* [122] generate the virtual humans rendered under different illumination conditions to enrich the supervision. All of these methods enrich the supervision with the augmented training set, improving the generalizability on the unseen testing samples.

Some methods also exploit the gallery images as the auxiliary information in the training process. A deep group-shuffling random walk framework is presented in [123], which enables end-to-end joint training to utilize the affinities between all the gallery images. An improved version with a Similarity-Guided Graph Neural Network (SGGNN) [124] is introduced to model the probe-to-gallery similarity in the training process, enhancing the generalizability. A constrained clustering scheme [125], incorporating a dominant set, is developed to improve the probe-gallery measurement, further enhancing the robustness.

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.

Zheng *et al.* [7] demonstrate that the temporal information is unreliable in an unconstrained tracking sequence. Early works directly learn the frame-level representations [7], [127], [128], and then average/max pooling is applied to obtain the video level representations.

To accurately and automatically capture the temporal information, a recurrent neural network architecture is designed for video-based person Re-ID [129]. Incorporated with a Siamese network architecture, it jointly optimizes the final recurrent layer for temporal information propagation and the temporal pooling layer for video-level representation learning. A weighted scheme for spatial and temporal

streams is developed in [130]. Yan *et al.* [131] present a progressive/sequential fusion framework to aggregate the frame-level human region representations, using a Long Short Term Memory (LSTM) network, yielding a video-level feature representation. The proposed network jointly aggregates the frame-level feature and spatio-temporal appearance information to enhance the video representation learning. Zheng *et al.* [132] introduce a triplet network for cross-view person identification in temporally synchronized videos, including view-specific optical flow learning and underlying skeleton feature learning. Semantic attributes are also adopted in [108] for video Re-ID with feature disentangling and frame re-weighting.

Since video sequences usually contain unavoidable outlier tracking frames, employing an attention scheme is also a popular approach to eliminate the effect. A temporal attention model [133] is proposed to automatically select the most discriminative frames in a given video and integrate the contextual information with a spatial recurrent model. The attention scheme is incorporated into a joint Spatial and Temporal Attention Pooling Network (ASTPN) [134] to select informative frames from the video sequence, extracting informative video representations. A co-segmentation inspired attention model [135] detects salient features across multiple video frames with mutual consensus estimation. Along this line, a diversity regularization [136] is employed to mine multiple discriminative body parts in each video sequence. The diverse property of the attention is also applied in image-based Re-ID [137]. An affine hull is adopted to handle the outlier frames within the video sequence [76].

To handle the varying lengths of video sequences, Chen *et al.* [138] divide the long video sequences into multiple short snippets, aggregating the top-ranked snippets to learn a co-attentive snippet embedding. A clip-level Spatial and Temporal Attention (STA) [139] exploits both spatial and temporal dimensional discriminative cues to produce a robust clip-level representation. Both the short- and long-term relations [140] are integrated in a self-attention scheme.

An interesting work [19] utilizes the multiple video frames to auto-complete occluded regions. Specifically, a spatio-temporal completion network is designed to generate the occluded body parts from the unoccluded parts of multiple frames, which enhances the robustness against occlusions. This provides a application-specific solution to handle challenging occlusion issues in video Re-ID.

2.1.5 Architecture Design

Framing person Re-ID as a specific pedestrian retrieval problem, most existing works adopt the network architectures [32], [70], [71], [72] 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 [72], the important modifications include changing the last convolutional stripe/size to 1 [68], employing adaptive average pooling in the last pooling layer [68], and adding bottleneck layer with batch normalization after the pooling layer [75].

For specific Re-ID network architecture design, Li *et al.* [34] start the first attempt by designing a filter pairing neural network (FPNN), which jointly handles misalignment and occlusions with part discriminative information mining.
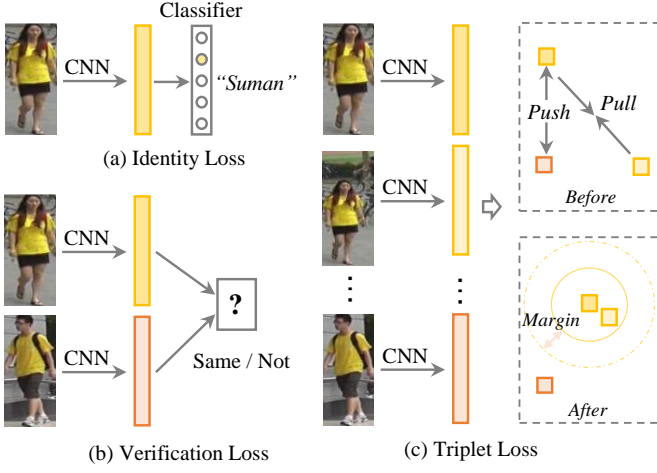
Fig. 3: Three kinds of widely used loss functions in the literature. (a) Identity Loss [33], [75], [98], [116] ; (b) Verification Loss [88], [145] and (c) Triplet Loss [13], [48], [84]. Many works employ their combinations [80], [142], [145], [146].

An improved neighborhood difference layer is proposed in [141] to capture patch feature difference and then the differences are summarized in a subsequent layer. Wang *et al.* [82] 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) [111] is designed to learn the identity-discriminative and view-invariant feature representations at multiple semantic levels. MLFN 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 [142] 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.

Recently, an efficient small scale network, namely Omni-Scale Network (OSNet) [143], is designed for person re-ID. To achieve multi-scale feature learning, a residual block composed of multiple convolutional streams is introduced. Meanwhile, point-wise and depth-wise convolutions are incorporated to guarantee efficiency.

With the increasing interest in auto-machine learning, an Auto-ReID [144] 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 [25], [26] or projection matrix [29]. 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. An overview of the distance metric learning designed for hand-crafted systems can be found in [2], [147]. 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.

Identity Loss. It treats the training process of person Re-ID as an image classification problem [46], *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

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

where n represents the number of training samples within each batch. The identity loss has been widely used in existing methods [18], [33], [75], [86], [89], [98], [105], [116], [118], [128]. Generally, it is easy to train and automatically mine the hard samples during the training process, as demonstrated in [148]. Several works have also investigated the softmax variants [149], such as the sphere loss in [150] and AM softmax in [89]. Another simple yet effective strategy, *i.e.*, label smoothing [33], [151], 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 [152].

Verification Loss. It optimizes the pairwise relationship, either with a contrastive loss [91], [118] or binary verification loss [34], [145]. The contrastive loss improves the relative pairwise distance comparison, formulated by

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

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 [74].

Binary verification [34], [145] discriminates the positive and negative of a input image pair. Generally, a differential feature f_{ij} is obtained by $f_{ij} = (f_j - f_j)^2$ [145], 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

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

The verification is often combined with the identity loss to improve the performance [88], [91], [118], [145].

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 [48]. 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

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

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 [13], [48], [84], [95]. The basic idea is to select the triplets which are hard for the model to discriminate [48], [153]. Specifically, a moderate positive mining with a weight constraint is introduced in [153] for distance metric learning, which directly optimizes the feature difference. Hermans *et al.* [48] demonstrate that the online hardest positive and hardest 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. Zhou *et al.* [154] present a point to set similarity for deep metric learning, which replace the point to point distance with the point to set metric. A hard-aware point-to-set deep metric learning [155] is designed to mine the informative hard triplets based on the point to set similarity. 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 [156], 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 [80], [83], [87], [99], [100], [106], [114], [142], [146], [157], [158]. 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 [55] 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

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

where v_i represents the corresponding stored memory feature for class y_i, and τ is a temperature parameter that controls the similarity space [159]. $v_i^T f_i$ measures the online instance matching score. The comparison with a memorized feature set of unlabelled identities is further included to calculate the denominator [55], handling the large instance number of non-targeted identities. This memory scheme is also adopted in unsupervised domain adaptive Re-ID [105].

2.2.2 Training strategy

The batch sampling strategy plays an important role in discriminative Re-ID model learning, especially for the triplet loss with hard mining. Different from general image classification, the number of annotated training images for each
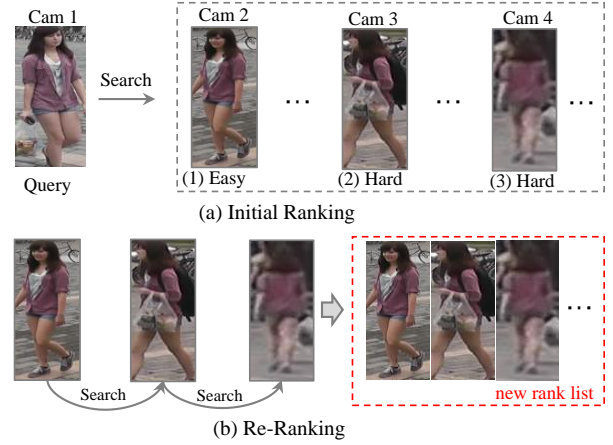

Fig. 4: An illustration of re-ranking in person Re-ID. Given a query example, an initial rank list is retrieved, where the hard matches are ranked in the bottom. Using the top-ranked easy positive match (1) as query to search in the gallery, we can get the hard match (2) and (3) with similarity propagation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

identity varies significantly [5]. Meanwhile, the severely imbalanced positive and negative sample pairs increases additional challenge for the training strategy [29].

The most commonly used training strategy for handling the imbalanced issue is identity sampling [48], [151]. For each training batch, a certain number of identities are randomly selected, and then several images are sampled from each selected identity. The sampled images formulate a training batch. This batch sampling strategy guarantees the informative positive and negative mining.

To further handle the imbalance issue between the positive and negative, Sample Rate Learning (SRL) is proposed in [82] to adaptively adjust the contribution of positive samples and negative tries. An adaptive weighted triplet loss is introduced in [43] to balance the positive and negative triplet using the similarity difference. A focal loss scheme is also investigated in [80] to handle the imbalance problem. A curriculum sampling strategy is designed in [80] to select easy triplets first and then optimize the hard triplets. A hard sample mining method with weighted complete bipartite graph matching is introduced in [160]. An efficient reference constraint is designed in [161]. The core idea is 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 [162] 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 [49], [163] or human interaction [164], [165]. Rank/Metric fusion [166], [167] 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. A ranking optimization using top-ranked similarity pulling and bottom-ranked dissimilarity pushing is proposed in [163]. A re-ranking method with k-reciprocal encoding which mines the contextual information is introduced in [49] to improve the initial ranking list. Due to its simplicity and effectiveness, it has been widely used in current state-of-the-arts to boost the performance. Bai *et al.* [168] solve the re-ranking problem from a manifold-based affinity learning view, utilizing the geometric structure of the underlying manifold. An expanded cross neighborhood re-ranking method [17] is introduced by integrating the cross neighborhood distance. A local blurring re-ranking [89] employs the clustering structure to improve neighborhood similarity measurement, refining the ranking list.

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 [169], [170]. Andy *et al.* [169] propose a query adaptive re-ranking method using locality preserving projections. An efficient online local metric adaptation method is presented in [170], 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 [164]. This provides reliable supervision during the re-ranking process. A hybrid human-computer incremental learning model is presented in [165], 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 [50]. Zheng *et al.* [171] 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 [50]. The rank fusion process in person Re-ID is formulated as a consensus-based decision problem with graph theory [172], mapping the similarity scores obtained by multiple algorithms into a graph with path searching. An Unified Ensemble Diffusion (UED) [167] 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 [166].

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 [4], iLIDS [174], GRID [175], PRID2011 [126], CUHK01-03 [34], Market-1501 [5], DukeMTMC [33], Airport [176] and MSMT17 [35]) and 7 video datasets (PRID-2011 [126], iLIDS-VID [6], MARS [7], Duke-Video [128], Duke-Tracklet [177], LPW [178] and LS-VID [140]). 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

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.

Dataset	Time	#ID	#image	#cam.	Label	Res.	Eval.
Image datasets							
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

found in [176] and their website[1]. Several observations can be made in terms of the dataset collection over recent years:

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 mannually cropped. This simulates the real-world scenario with tracking/detection errors.

Evaluation Metrics. To evaluate a Re-ID system, Cumulative Matching Characteristics (CMC) [58] and mean Average Precision (mAP) [5] are two widely used measurements.

CMC-k (*a.k.a,* Rank-k matching accuracy) [58] 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.

The other widely used 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. We observe that the recently collected datasets in Table 2 utilize both mAP and CMC as the evaluation metrics.

Considering the efficiency and complexity of training a Re-ID model, some recent works [143], [144] 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.

---

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

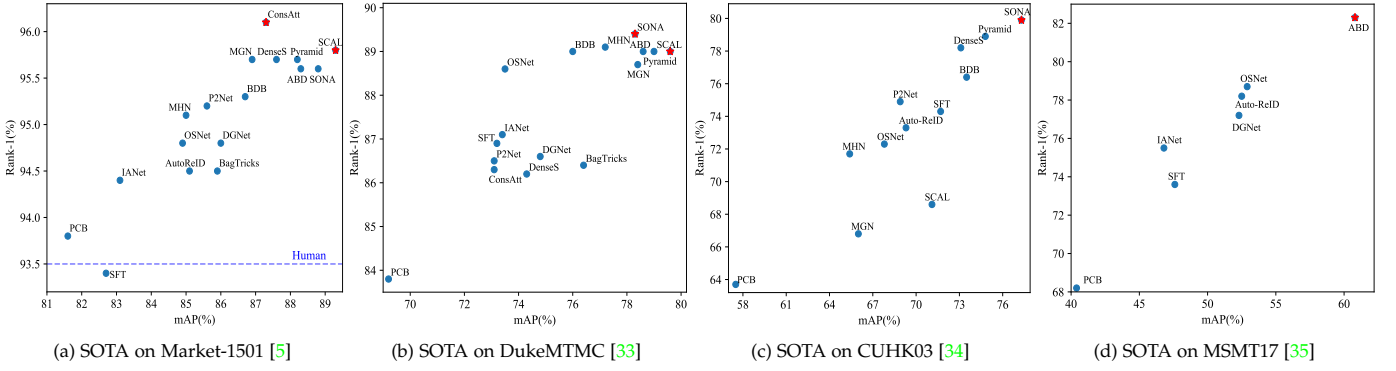(a) SOTA on Market-1501 [5] (b) SOTA on DukeMTMC [33] (c) SOTA on CUHK03 [34] (d) SOTA on MSMT17 [35]

Fig. 5: State-of-the-arts (SOTA) on four widely used image-based person Re-ID datasets. Both the Rank-1 accuracy (%) and mAP value (%) are reported. For CUHK03 [34], the detected data under the setting [49] is reported. For Market-1501, the single query setting is used, and the human-level rank-1 accuracy (93.5%) is also reported [173]. The best result is highlighted with a red star.

2.4.2 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.

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 [68], MGN [179], Auto-ReID [144], ABD-Net [137], BagTricks [151], OSNet [143], DGNet [117], SCAL [83], MHN [101], P2Net [100], BDB [121], SONA [102], SFT [89], ConsAtt [87], DenseS [99], Pyramid [162], IANet [103]. We summarize the results on four datasets, as shown in Fig. 5. From these results, we can draw five major insights, as discussed below. The analysis also motivates our proposed powerful new AGW baseline, which will be introduced § 4.2, providing an important guidance for future Re-ID model design.

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% [173]) on the widely used Market-1501 dataset. In particular, ConsAtt [87] achieves the highest rank-1 accuracy of 96.1%, and SCAL [83] obtains the best mAP of 89.3% on Market-1501 dataset. On another widely used DukeMTMC dataset, the rank-1 accuracy also reaches 89.4% by SONA [102] and the best mAP 79.6% is achieved by SCAL [83]. 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 [140] or unsupervised learning [127].

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 [151]. 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 [180]. 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 [144], [162].

Third, attention is beneficial for discriminative Re-ID model learning. We observe that all the methods (ConsAtt [87], SCAL [83], SONA [102], ABD-Net [137]) 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 [137], diverse [136], consistent [87] and high-order [102] 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 [87], ABD-Net [137] and SONA [102]. In addition, a dynamic multi-loss training strategy is designed in [162] 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 [43], [82].

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 [35] are much lower than that on Market-1501 and DukeMTMC. Meanwhile, on some other challenging datasets with limited training samples (*e.g.*, GRID [175] and VIPeR [4]), the performance is still very low. In addition, Re-ID models usually suffers significantly on cross-dataset evaluation [45], [106], and the performance drops dramatically under adversarial attack [181]. 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 [135], GLTR [140], STA [139], ADFD [108], STC [19], DRSA [136], Snippet [138], ETAP [128], DuATM [85], SDM [182], TwoS [130], ASTPN [134], RQEN [178], Forest [133], RNN [129]
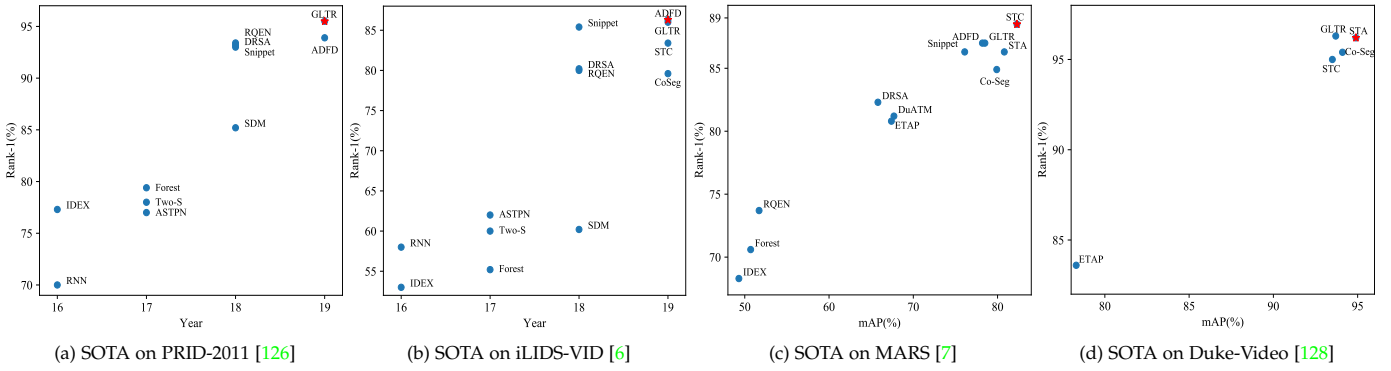
Fig. 6: State-of-the-arts (SOTA) on four widely used video-based person Re-ID datasets. The Rank-1 accuracies (%) over years are reported. mAP values (%) on MARS [7] and Duke-Video [128] are reported. For Duke-Video, we refer to the settings in [128]. The best result is highlighted with a red star.

and IDEX [7]. 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 [129] in 2016) to 95.5% (GLTR [140] in 2019) on PRID-2011 dataset, and from 58% (RNN [129]) to 86.3% (ADFD [108]) on iLIDS-VID dataset. On the large-scale MARS dataset, the Rank-1 accuracy/mAP increase from 68.3%/49.3% (IDEX [7]) to 88.5%/82.3% (STC [19]). On the Duke-Video dataset [128], STA [139] 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 [139], STC [19], GLTR [140]) design spatial-temporal aggregation strategies to improve the video Re-ID performance. Similar to image-based Re-ID, the attention scheme across multiple frames [108], [139] also greatly enhances the discriminability. Another interesting observation in [19] 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 [140], the Rank-1 accuracy/mAP of GLTR [140] are only 63.1%/44.43%, while GLTR [140] can achieve state-of-the-art or at least comparable performance on the other four daatsets. LS-VID [140] 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 [183] 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. [184] 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 [53], [185] 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 [54]. 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 [54] is built on top of a recurrent neural network to learn 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. [110] 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 is also presented in [186] to learn a shared space with image-to-text matching. A deep adversarial graph attention convolution network is designed in [187] with graph relation mining.

For text attribute queries, a global visual-textual embedding is learned in a zero-shot learning framework for local attribute recognition [188].

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 [20], [51], [189].

Wu *et al.* [20] start the first attempt to address this issue, by proposing a deep zero-padding framework [20] to adaptively learn the modality sharable features. A two stream network is introduced in [146], [190] to model the modality-sharable and -specific information, addressing the intra- and cross-modality variations simultaneously. Besides the cross-modality shared embedding learning [191], the classifier-level discrepancy is also investigated in [192], [193]. Recent methods [194], [195] adopt the GAN technique to generate cross-modality person images to reduce the cross-modality discrepancy at both image and feature level.

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 [12], [13]. A cascaded SR-GAN [196] generates the high-resolution person images in a cascaded manner, incorporating the identity information. Li *et al.* [197] adopt the adversarial learning technique to obtain resolution-invariant image representations.

Regarding the illumination factor, matching person images across multiple lighting conditions is studied in [14].

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 [46], [55]. It is challenging due to the different focuses of two major components.

Zheng *et al.* [46] 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.* [55] 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) [198] 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 [199] is learned in a graph learning framework to improve the end-to-end person search. A query-guided end-to-end person search system [200] 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 [201] to generate more reliable bounding boxes.

Yamaguchi *et al.* [202] 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 [43]. A graph-based formulation to link person hypotheses is proposed for multi-person tracking [203], where the holistic features of the full human body and body pose layout are combined as the representation for each person. Ristani *et al.* [43] 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) [204] with both intra- and inter-camera relation modeling is proposed.

3.3 Semi-supervised and Unsupervised Re-ID

3.3.1 Unsupervised Re-ID

Without cross-camera labels, early unsupervised Re-ID mainly learns invariant components, *i.e.*, dictionary [205], metric [206], [207] or saliency [57], which leads to limited discriminability or scalability.

For deeply unsupervised methods, cross-camera label estimation is one the popular approaches [127], [208]. Dynamic graph matching (DGM) [209] formulates the label estimation as a bipartite graph matching problem. Liu *et al.* progressively mine the labels with step-wise metric promotion [207]. A robust anchor embedding method [76] 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 [208]. Soft multi-label learning [210] mines the soft label information from a reference set for unsupervised learning. A Tracklet Association Unsupervised Deep Learning (TAUDL) framework is introduced in [177]. The main idea is to jointly conduct the within-camera tracklet association and model the cross-camera tracklet correlation. Similarly, an unsupervised camera-aware similarity consistency mining method [211] 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 [212].

The semantic attributes are also adopted in unsupervised Re-ID. A Transferable Joint Attribute-Identity Deep Learning (TJ-AIDL) framework is introduced in [109], which incorporates identity-discriminative and semantic attribute feature learning in a two-branch network.

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 [157] is designed to learn discriminative patch features by mining patch level similarity. A Self-similarity Grouping (SSG) approach [213] 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-supervised Re-ID. With limited label information, a one-shot metric learning method is proposed in [214], which incorporates a deep texture representation learned from intensity image and learns a color metric using a single pair of ColorChecker images. A stepwise one-shot learning method (EUG) is proposed in [128] for video-based person re-identification, gradually selecting a few candidates from unlabeled tracklets to enrich the labeled tracklet set.

3.3.2 Unsupervised Domain Adaptation

Unsupervised domain adaptation (UDA) transfers the knowledge on a labeled source dataset to the unlabeled target dataset [44]. 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.* [35] handle the domain gap problem by proposing a Person Transfer Generative Adversarial Network (PTGAN), transferring the knowledge from one labeled source dataset to another unlabeled target dataset. An image-image domain adaptation method [118] is developed with preserved self-similarity and domain-dissimilarity, trained with a similarity preserving generative adversarial network (SPGAN). A Hetero-Homogeneous Learning (HHL) method [215] simultaneously considers the camera invariance with homogeneous learning (image pairs from the same domain) and domain connectedness with heterogeneous learning (cross-domain negative pairs). 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 in to improve the image generation process [119].

Bak *et al.* [122] generate a new 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 with a cycle-consistency translation constraint.

Target Domain Supervision Mining. Some methods directly mine the supervision on the unlabeled target dataset. An exemplar memory learning scheme is designed in [105]. Three types of invariant cues, including exemplar-invariance, camera invariance and neighborhood-invariance are adopted as the supervision signal design [105]. The Domain-Invariant Mapping Network (DIMN) [106] 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 [219] as the supervision signal to reduce the domain gap. A self-training method with progressive augmentation [220] jointly captures the local structure and global data distribution on the target dataset.

The spatio-temporal information is also utilized as the supervision in TFusion [221]. TFusion transfers the spatio-temporal patterns learned in the source domain to the target domain with a Bayesian fusion model in an incremental learning manner, improving the cross-dataset Re-ID.

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 publi-

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.

Methods	Source	Gen.	Market-1501 R1	Market-1501 mAP	DukeMTMC R1	DukeMTMC mAP
PUL [208] TOMM18	Model	No	45.5	20.5	30.0	16.4
CAMEL [206] ICCV17	Model	No	54.5	26.3	-	-
PTGAN [118] CVPR18	Data	Yes	58.1	26.9	46.9	26.4
TJ-AIDL[†] [109] CVPR18	Data	No	58.2	26.5	44.3	23.0
HHL [215] ECCV18	Data	Yes	62.2	31.4	46.9	27.2
DAS[§] [122] ECCV18	Synthesis	Yes	65.7	-	-	-
MAR[‡] [210] CVPR19	Data	No	67.7	40.0	67.1	48.0
ENC [105] 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[‡] [157] 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 [219] ICCV19	Data	No	64.3	34.5	55.4	36.7
CASC[‡] [211] ICCV19	Model	No	65.4	35.5	59.3	37.8
PDA [119] 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 [220] ICCV19	Model	No	78.4	54.6	72.4	54.3
SSG [213] ICCV19	Model	No	80.0	58.3	73.0	53.4

- [†] TJ-AIDL [109] requires additional attribute annotation.
- [§] DAS [122] generates synthesized virtual humans under vairous lightings.
- [‡] PAUL [157], MAR [210] and CASC [211] use MSMT17 as source dataset.
- UGA [212] reports a rank-1/mAP 87.2%/70.3% under a different setting.

cations in top venues, *e.g.*, more than 10 papers in 2019. We review the SOTA for unsupervised deeply learned methods on two widely-used image-based Re-ID datasets. These methods include PUL [208], CAMEL [206], PTGAN [118], TJ-AIDL [109], HHL [215], DAS [122], MAR [210], ENC [105], ATNet [216], PAUL [157], SBGAN [217], UCDA [219], CASC [211], PDA [119], CR-GAN [218], PAST [220], SSG [213]. 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 [206]) to 80.0%/58.3% (SSG [213]) on the Market-1501 dataset within two years. The performance for DukeMTMC dataset increases from 30.0%/16.4% to 73.0%/54.3%. This demonstrates the success of unsupervised Re-ID with deep learning.

Second, current unsupervised Re-ID is still under-developed 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 [220], SSG [213]). 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 [87] has achieved 96.1% on the Market-1501 dataset, while the highest accuracy of unsupervised SSG [213] is about 80.0%. Recently, He *et al.* [222] have demonstrated that unsupervised learning with large-scale unlabeled training data has the ability to outperform the supervised learning methods on various tasks. 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 [223]. A fully convolutional network [224] 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.* [180] 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 [225] also tries to learn from the unoccluded regions. The Pose-Guided Feature Alignment (PGFA) [226] exploits the pose landmarks to mine discriminative part information from occlusion noise.

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 [16], [17] or attention cues [57], [84], [227] are exploited. For video sequences, set-level feature learning [76] or frame level re-weighting [138] are the commonly used approaches. Hou *et al.* [19] also utilize multiple video frames to auto-complete occluded regions.

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 [33]. A Distribution Net (DNet) that models the feature uncertainty is proposed in [228] for robust Re-ID model learning against label noise, reducing the impact of samples with high feature uncertainty. Different from the general classification problem [229], robust Re-ID model learning suffers from limited training samples for each identity.

3.5 Open-set Re-ID

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 [59], [60]. The verification usually requires a learned condition τ, *i.e.*, $sim(query, gallery) > \tau$. Early researches design hand-crafted systems [59], [60], [61]. For deep learning methods, an Adversarial PersonNet (APN) is proposed in [230], 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 [228]. However, it remains quite challenging to achieve a high true target recognition and maintain low false target recognition rate [231].

Another challenging issue is the dynamically increased gallery [232], which may contain both seen and unseen identities. Enrolling the new identities to existing gallery provides enriched supervision to enhance the Re-ID model, but it also introduces additional challenge to identify it.
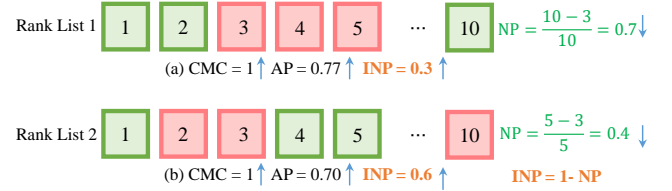

Fig. 7: Illustration of the difference between the widely used CMC, AP and the negative penalty (NP) measurements. True matching and false matching are bounded in green and red boxes, respectively. Assume that only three correct matches exist in the gallery, rank list 1 gets better AP, but gets much 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 convenience, we compute the inverse negative penalty (INP), *e.g.*, INP = 1- NP. Larger INP means better performance.

4 AN OUTLOOK: RE-ID IN NEXT ERA

This section discusses some future directions with a new evaluation metric, a strong baseline for person Re-ID. It provides an important guidance for future Re-ID research. We first introduce a new evaluation metric (mINP) for practical Re-ID in § 4.1, and then propose a simple yet effective AGW baseline in § 4.2. Finally, we discuss some under-investigated open issues in § 4.3.

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. In practical applications, the algorithm usually returns a retrieved ranking list for further manual investigation. Considering that the target person should not be neglected in the top-ranked list retrieved from multiple cameras, the rank position of the hardest correct match determines the workload of the inspectors. 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

$$\text{NP}_i = \frac{R_i^{hard} - |G_i|}{R_i^{hard}}, \qquad (6)$$

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

$$\text{mINP} = \frac{1}{n} \sum_i (1 - \text{NP}_i) = \frac{1}{n} \sum_i \frac{|G_i|}{R_i^{hard}}. \qquad (7)$$

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. This metric evaluates the ability to retrieve the hardest correct match, providing a supplement for measuring the Re-ID performance.

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.

Method	Market-1501 [5]			DukeMTMC [33]		
	R1	mAP	mINP	R1	mAP	mINP
BagTricks [151] CVPR19W	94.5	85.9	59.4	86.4	76.4	40.7
ABD-Net [137] 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 [233]	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 [234]	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

4.2 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- and cross-modality Re-ID tasks. Specifically, our new baseline is designed on top of BagTricks [151], and AGW contains the following three major improved components :

(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 [233] to obtain the weighted sum of the features at all positions, represented by

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

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 [233]. We adopt the default setting from [233] 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* [234], formulated by

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

where p_k is a pooling hyperparameter, which can be learned in the back-propagation process [234]. 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,

$$\mathcal{L}_{wrt}(i, j, k) = \log(1 + \exp(w_i^p d_{ij}^p - w_i^n d_{ik}^n)). \qquad (10)$$

$$w_i^p = \frac{\exp(d_{ij}^p)}{\sum_{d^p \in \mathcal{P}} \exp(d^p)}, w_i^n = \frac{\exp(-d_{ik}^n)}{\sum_{d^n \in \mathcal{N}} \exp(-d^n)}, \quad (11)$$

where (i, j, k) represents the mined hard triplet within each training batch. $\mathcal{P}$ is the corresponding positive set, and $\mathcal{N}$ is the negative set. The above weighted regularization inherits the advantage of relative distance optimization between positive and negative pairs, but avoids introducing any additional margin parameters. Our weighting strategy is similar to [235], but our solution is entirely parameter free, which makes it more flexible and adaptable.

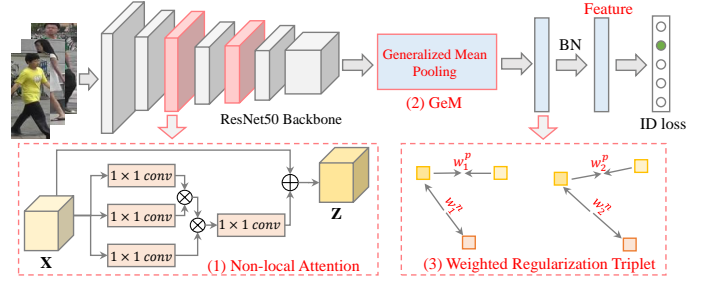3. Details in https://github.com/mangye16/ReID-Survey

Fig. 8: The framework of the proposed AGW baseline using the widely used ResNet50 [72] as the backbone network.

TABLE 5: Comparison with the state-of-the-arts on CUHK03 [34] and MSMT17 [35]. Rank-1 accuracy (%), mAP (%) and mINP (%) are reported on two public datasets.

Method	CUHK03 [34]			MSMT17 [35]		
	R1	mAP	mINP	R1	mAP	mINP
BagTricks [151] 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

Overall Framework. The overall framework of AGW is shown in Fig 8. Other components are exactly the same as [151]. In the testing phase, the output of BN layer is adopted as the feature representation for Re-ID. The implementation details can be found in the supplementary material.

Results on Single-modality Re-ID. We first evaluate each component on two widely used single modality image-based datasets (Market-1501 and DukeMTMC), as shown in Table 4. We also compare two state-of-the-art methods, BagTricks [151] and ABD-Net [137]. We report the results on CUHK03 and MSMT17 datasets in Table 5. We obtain the following two observations:

1) All the components consistently improve the performance, 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 [68], 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 [137], 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 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 6. We follow the settings in AlignG [195] to perform the experiments. Results show that AGW achieves much better performance than existing cross-modality Re-ID models. However, there is still a large gap between the single- and cross-modality Re-ID, which needs further study.

4.3 Under-Investigated Open Issues

4.3.1 Scalable Re-ID

With the significantly increasing scale of data in the Re-ID community and real large scale camera network, attempts to address the scalability issue should be made in the future. Probably, from the following three perspectives.

TABLE 6: 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 [51] | | SYSU-MM01 [20] | | | |
| | Visible-Thermal | | All Search | | Indoor Search | |
Method	R1	mAP	R1	mAP	R1	mAP
Zero-Pad [20] ICCV17	17.75	18.90	14.8	15.95	20.58	26.92
HCML [190] AAAI18	24.44	20.08	14.32	16.16	24.52	30.08
eBDTR [146] TIFS19	34.62	33.46	27.82	28.42	32.46	42.46
HSME [191] AAAI19	50.85	47.00	20.68	23.12	-	-
D^{2}RL [194] CVPR19	43.4	44.1	28.9	29.2	-	-
MAC [193] MM19	36.43	37.03	33.26	36.22	36.43	37.03
MSR [192] TIP19	48.43	48.67	37.35	38.11	39.64	50.88
AlignG [195] ICCV19	57.9	53.6	42.4	40.7	45.9	54.3
AGW	**70.05**	**66.37**	**47.50**	**47.65**	**54.17**	**62.97**
	mINP = 50.19		mINP =35.30		mINP = 59.23	

Fast Re-ID. For fast retrieval, hashing has been extensively studied to boost the searching speed, approximating the nearest neighbor search. Cross-camera Semantic Binary Transformation (CSBT) [236] transforms the original high-dimensional feature representations into compact low-dimensional identity-preserving binary codes. Zhu *et al.* [237] present a part-based deep hashing method, while the structure information is incorporated in [238].

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 [79], [143], [144]. Model distillation is another approach, *e.g.*, a multi-teacher adaptive similarity distillation framework is proposed in [239], 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) [13] employs a simple distance based routing strategy to adaptively adjust the model, fitting to hardware devices with different computational resources.

4.3.2 Domain Generalized Re-ID

Multi-Domain Data. It is well recognized that there is a large domain gap between different datsets [35]. 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 [106]. The Domain-Invariant Mapping Network (DIMN) [106] starts the first attempt by designing a meta-learning pipeline, learning a mapping between a person image and its identity classifier.

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 [196], both the query and gallery sets may contain different modalities (visible, thermal [20], depth [53] or text description [9]). 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 along this direction is expected to promote the applicability of Re-ID in real scenarios.

4.3.3 Minimizing Human Annotation

Besides the unsupervised learning, incorporated with active learning or human interaction [158] provides another possible solution to alleviate the reliance on human annotation.

Minimizing Human Annotation. Incorporating human interaction, a pairwise subset selection framework is proposed in [240] to minimize 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 [158] 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 [182] is designed. A weakly supervised framework with multi-instance multi-label learning is proposed in [241] to mine potential intra-class person labels with the weakly labelled video data.

Learning for Virtual Data. This provides an alternative for minimizing the human annotation. A synthetic dataset is collected in [242] for training and testing, and they achieve competitive performance on real-world datasets when trained on this synthesized dataset. Bak *et al.* [122] 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 [104] to systematically study the effect of viewpoint for a person Re-ID system. How to bridge the gap between synthesized images and real-world datasets remains challenging.

4.3.4 Dynamic Camera Network.

A new camera may be temporarily inserted into an existing surveillance system [243] in practical applications, termed as *dynamic camera network*. To adapt a learned model to a new camera, a transitive inference algorithm [243] 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 a open-world person Re-ID system [244]. The model adaptation and environmental factors of cameras are crucial in practical dynamic camera networks.

In a dynamically updated network, it is very likely to contain a large number of persons with changing clothes. Yang *et al.* [245] present a spatial polar transformation (SPT) to learn cross-cloth invariant representation. It would be interesting to further explore the possibility of addressing the cloth-changing issue in future Re-ID research.

4.3.5 Domain-Specific Architecture Design

Existing deep Re-ID methods usually adopt the architectures designed for image classification as the backbone network. Some methods also slightly modify the architecture to achieve better Re-ID features [75], [151]. Very recently, researchers have started to design domain specific architectures, *e.g.*, OSNet with omni-scale feature learning [143]. It detects the small-scale discriminative features at a certain scale. OSNet is extremely lightweight and achieves competitive performance when training from scratch.

With the advancement of automatic neural architecture search (*e.g.*, Auto-ReID [144]), more domain-specific powerful architectures are expected in the future.

5 CONCLUDING REMARKS

With the significantly increasing interest in person re-identification, 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 in terms of three aspects: feature representation learning, deep metric learning and ranking optimization. With powerful deep learning techniques, 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. By summarizing the advantages of existing methods on various benchmarks, we design a new AGW baseline, which achieves state-of-the-art performance on both single- and cross-modality 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 the new baseline and new evaluation metric will provide important guidance for future Re-ID research. Finally, we discuss some under-investigated open issues.

REFERENCES

[1] S. Gong, M. Cristani, S. Yan, and C. C. Loy, Eds., *Person Re-Identification*. Springer, 2014.

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

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

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

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

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

[16] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017, pp. 1077–1085.

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

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

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

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

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

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

[30] H.-X. Yu, A. Wu, and W.-S. Zheng, "Unsupervised person re-identification by deep asymmetric metric embedding," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018.

[31] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, 2014.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

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

[40] D. Geronimo, A. M. Lopez, A. D. Sappa, and T. Graf, "Survey of pedestrian detection for advanced driver assistance systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, no. 7, pp. 1239–1258, 2009.

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

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

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

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

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

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

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

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

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

[60] 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.

[61] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 38, no. 3, pp. 591–606, 2015.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

[67] 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.

[68] 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.

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

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

[71] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

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

[73] L. Wu, C. Shen, and A. v. d. Hengel, "Personnet: Person re-identification with deep convolutional neural networks," *arXiv preprint arXiv:1601.07255*, 2016.

[74] 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.

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

[76] 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.

[77] 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.

[78] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *CVPR*, 2018, pp. 1062–1071.

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

[80] 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.

[81] 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.

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

[83] 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.

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

[85] 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.

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

[87] 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.

[88] 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.

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

[90] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *CVPR*, 2019, pp. 1389–1398.

[91] 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.

[92] 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.

[93] 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.

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

[95] 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.

[96] 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.

[97] 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.

[98] 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.

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

[100] 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.

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

[102] 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.

[103] 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.

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

[105] 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.

[106] 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.

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

[108] 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.

[109] 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.

[110] 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.

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

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

[113] 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.

[114] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *CVPR*, 2018, pp. 4099–4108.

[115] 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.

[116] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *CVPR*, 2018, pp. 5157–5166.

[117] 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.

[118] 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.

[119] 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.

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

[121] 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.

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

[123] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang, "Deep group-shuffling random walk for person re-identification," in *CVPR*, 2018, pp. 2265–2274.

[124] 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.

[125] L. T. Alemu, M. Pelillo, and M. Shah, "Deep constrained dominant sets for person re-identification," in *ICCV*, 2019, pp. 9855–9864.

[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] 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.

[128] 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.

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

[130] 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.

[131] 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.

[132] K. Zheng, X. Fan, Y. Lin, H. Guo, H. Yu, D. Guo, and S. Wang, "Learning view-invariant features for person identification in temporally synchronized videos taken by wearable cameras," in *ICCV*, 2017, pp. 2858–2866.

[133] 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.

[134] 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.

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

[136] 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.

[137] 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.

[138] 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.

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

[140] 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.

[141] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *CVPR*, 2015, pp. 3908–3916.

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

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

[144] 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.

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

[146] 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.

[147] 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.

[148] X. Zhang, F. X. Yu, S. Karaman, W. Zhang, and S.-F. Chang, "Heated-up softmax embedding," *arXiv preprint arXiv:1809.04157*, 2018.

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

[150] X. Fan, W. Jiang, H. Luo, and M. Fei, "Spherereid: Deep hypersphere manifold embedding for person re-identification," *Journal of Visual Communication and Image Representation*, vol. 60, pp. 51–58, 2019.

[151] 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.

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

[153] 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.

[154] 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.

[155] 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.

[156] 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.

[157] 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.

[158] 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.

[159] 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.

[160] H. Sun, Z. Chen, S. Yan, and L. Xu, "Mvp matching: A maximum-value perfect matching for mining hard samples, with application to person re-identification," in *ICCV*, 2019, pp. 6737–6747.

[161] 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.

[162] 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.

[163] 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.

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

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

[166] 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.

[167] 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.

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

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

[170] 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.

[171] 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.

[172] 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.

[173] 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.

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

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

[176] 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 Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 3, pp. 523–536, 2018.

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

[178] 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.

[179] 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.

[180] 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.

[181] 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.

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

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

[184] 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.

[185] 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.

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

[187] 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.

[188] Q. Dong, S. Gong, and X. Zhu, "Person search by text attribute query as zero-shot learning," in *ICCV*, 2019, pp. 3652–3661.

[189] 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.

[190] 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.

[191] 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.

[192] 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.

[193] M. Ye, X. Lan, and Q. Leng, "Modality-aware collaborative learning for visible thermal person re-identification," in *ACM MM*, 2019, pp. 347–355.

[194] 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.

[195] 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.

[196] 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.

[197] 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.

[198] 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.

[199] 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.

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

[201] 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.

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

[203] 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.

[204] 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.

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

[206] 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.

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

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

[209] 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 Transactions on Image Processing (TIP)*, vol. 28, no. 6, pp. 2976–2990, 2019.

[210] 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.

[211] 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.

[212] 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.

[213] 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.

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

[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] 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.

[220] 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.

[221] 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.

[222] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv preprint arXiv:1911.05722*, 2019.

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

[224] 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.

[225] 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.

[226] 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.

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

[228] 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.

[229] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *NeurIPS*, 2018, pp. 8536–8546.

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

[231] M. Golfarelli, D. Maio, and D. Malton, "On the error-reject trade-off in biometric verification systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 19, no. 7, pp. 786–796, 1997.

[232] M. Zheng, S. Karanam, and R. J. Radke, "Rpifield: A new dataset for temporally evaluating person re-identification," in *CVPR Workshops*, 2018, pp. 1893–1895.

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

[234] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 7, pp. 1655–1668, 2018.

[235] 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.

[236] 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.

[237] 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.

[238] L. Wu, Y. Wang, Z. Ge, Q. Hu, and X. Li, "Structured deep hashing with convolutional neural networks for fast person re-identification," *Computer Vision and Image Understanding*, vol. 167, pp. 63–73, 2018.

[239] 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.

[240] 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.

[241] J. Meng, S. Wu, and W.-S. Zheng, "Weakly supervised person re-identification," in *CVPR*, 2019, pp. 760–769.

[242] I. B. Barbosa, M. Cristani, B. Caputo, A. Rognhaugen, and T. Theoharis, "Looking beyond appearances: Synthetic training data for deep cnns in re-identification," *Computer Vision and Image Understanding*, vol. 167, pp. 50–62, 2018.

[243] 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.

[244] 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.

[245] Q. Yang, A. Wu, and W. Zheng, "Person re-identification by contour sketch under moderate clothing change." *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019.