This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2957336, IEEE Access

A Survey on Deep Learning Based Person Re-Identification Systems

Muna O. AlMasawa[1], Lamiaa A. Elrefaei[1,2] (Senior Member, IEEE), Kawthar Moria[1]

[1]Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

[2]Electrical Engineering Department, Faculty of Engineering at Shoubra, Benha University, Cairo, Egypt

Corresponding author: Muna O. AlMasawa (e-mail: malmasawa0001@stu.kau.edu.sa).

This project was funded by Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under grant No. (DG-019-611-1440)

ABSTRACT Person re-identification systems (person Re-ID) have recently gained more attention between computer vision researchers. They are playing a key role in intelligent visual surveillance systems and have widespread applications like applications for public security. The person Re-ID systems can identify if a person has been seen by a non-overlapping camera over large camera network in an unconstrained environment. It is a challenging issue since a person appears differently under different camera views and faces many challenges such as pose variation, occlusion and illumination changes. Many methods had been introduced for generating handcrafted features aimed to handle person Re-ID problem. In recent years, many studies have started to apply deep learning methods to enhance the person Re-ID performance due the deep learning yielded significant results in computer vision issues. Therefore, this paper is a survey of the recent studies that proposed to improve the person Re-ID systems using deep learning. The public datasets that are used for evaluating these systems are discussed. Finally, the paper addresses future directions and current issues that must be considered toward improving the person Re-ID systems.

INDEX TERMS Deep Learning, Person Re-Identification, Video Surveillance.

I. INTRODUCTION

Person Re-Identification (Person Re-ID) has recently attracted academic attention in the field of computer vision. There is an increasing demand for robust intelligent video surveillance due to their importance in modern society for security purposes, such as preventing crimes and terrorist activities, forensic investigation, etc. Governments strive hard to improve surveillance technology for the safety of its citizens. Automated the monitoring and analyzing recorded videos is among the essential and important tasks in intelligent video surveillance systems. However, this monitoring by a human is time and effort consuming. Person Re-identification is a significant task in intelligent video surveillance systems. It is defined as the process of reidentifying and recognizing the same person over a set of non-overlapping cameras in multi-camera surveillance systems in disparate geographical areas [1] [2].

Person Re-ID is a challenging issue since the videos are recorded by non-overlapping cameras under different environments. As a result, using primary biometric data, such as face is not useful for this task. Researches focus on the appearance of a person but also there is large ambiguity of visual appearance caused by intra-class and inter-class problems. In practice, this means that the same person can appear differently, and different persons might look similar. The inter-class and intra-class problems occur essentially because of the changes in human body poses, illumination, scene occlusions, viewpoint, camera settings and background noise at short time and longtime [3][4]. So, extracting robust features under different conditions and mapping the features from the same groups closer than a different group to re-identify the same person across separate cameras is a fundamental and critical problem in person Re-ID system.

There are two basic modules for the traditional person Re-ID system: features learning and distance metric learning [5]. Feature learning focuses on extracting discriminative features that are robust to challenging issues such as illumination and pose. Distance metrics focus on finding a metric that computes the similarity between features of two images and reduces the distance between images of the same person as much as possible and increases the distance between images of different persons as much as possible. In so doing, the intra-class distance is minimized, and the inter-class distance is maximized [6][7]. There are two types of features which can be used for reidentification process. The first type is appearance-based features that are defined by persons' clothing and objects carried and so on. The second type is biometric-based features (primary biometric and soft biometric) [8]. These features can be extracted by hand-crafted methods or deep learning methods. Hand-crafted methods are the traditional methods used to extract low-level features, such as extracting color features by histogram and extracting texture by Local Binary Pattern (LBP) [1] [9] [10]. Hand-crafted features and metrics are not effective in the case of large intra-class and inter-class variations. Recently, Deep learning methods (like the Convolutional Neural Networks CNN) have been widely applied to solve many computer vision problems including image classification, face recognition, object recognition, etc. [11][12]. Some deep learning models can combine the feature learning and the distance metric learning into one integrated framework. So, many recent person Re-ID studies adopted Deep learning-based methods to achieve more accurate results [10]. Several studies combined deep learning methods with the hand-crafted methods, such in [13] and [14], hand-crafted features such as color, LBP and Local Maximal Occurrence (LOMO) features are combined with CNN and Long Short-Term Memory networks (LSTM).

Person Re-ID systems fall into three main categories. The first category is image-based person Re-ID (Single-shot/ Multiple shots) which depend on pairs of images only, each pair has a single shot or multi-shots of a person's appearance. Most of the existing person Re-ID approaches listed under this category. The second category is video-based person Re-ID which works by searching multiple frames representing the same person. Unlike than image-based person Re-ID, video-based person Re-ID can display a person in each frame in different poses and from different standpoints. Moreover, it contains not only the appearance features but also the spatial-temporal features, such as motion patterns and gait. The last category is the image to video person Re-ID which searches the person's image in the video sequence [1] [6].

Person Re-ID can be done in a short-period of time as well as a long-period. The short-period of time such as when a person appears in camera A, and then appears in camera B after seconds or minutes. Long-period of time, such as when a person appears in camera A, then after days he appears again in camera A or B. Appearance features, such as clothes, are commonly used to re-identify a person on short-period time, and most of the current researches focus on short-period time Re-ID. However, in real surveillance systems, the variations of visual appearance over time affect the recognition of the person. So, some researchers try to use another type of features such as soft biometric features or the combination of both appearance-based features and biometric-based features [7] [15].

The performance of person Re-ID system is measured by cumulative matching characteristics (CMC) which it used commonly for computing the accuracy at rank-k (k number of top images) which it is defined as the probability that the true identity is within the first 1,5,..k ranks of the ranked list [1] [2] [7]. Another widely used metric for measuring the accuracy is the mean average precision (mAP), considering person Re-ID as a retrieval task.

The contribution of this work consists in:

- Explaining the popular architecture of traditional and deep learning person Re-ID systems.
- Discussing the recent deep learning person Re-ID systems categorized under three main categories based on the input type: image-based, video-based and image to video person Re-ID systems.
- Comparing top-rank accuracy results achieved in common datasets using existing state-of-the-art models.
- Exploring the major challenging effects on the person Re-ID systems and guide future research directions.

This survey is organized as follows. In Section 2, we discuss the general architecture for person Re-ID System. In Section 3, we review the current deep learning methods designed for person ReID as per input type. In section 4, we present a common dataset that used for evaluating the existing deep learning models. Finally, in section 5, we highlight the current issues in person Re-ID systems and future directions.

II. General Architecture for Person Re-ID System

The traditional Person Re-ID system has two main steps as seen in figure 1. It starts by taking the input image (probe image) for a person that has been captured from a camera (Cam 1). Then, the discriminative features are extracted by special handcrafted extraction methods such as color histograms, texture descriptor and Bag-of-Words [16] and represented as a feature vector. The goal of person Re-ID now is to find a person image (probe image) in all gallery images, where the gallery has a set of images that have been captured from other cameras (Camera 2,3,4, etc.). This is achieved by learning the distance between the probe image and all gallery images using a special distance metric such as KISSME [17] and XQDA metric [18] and returning top-ranked list of gallery images with the smallest distances. Recently, deep learning methods and distance measures have been used together in one unified framework to find out whether the two images are for the same person or not. This is done either by modifying existing deep learning architectures or designing new deep neural network. The common deep model used in the existing studies and achieve good results are ResNet-50 [19], inception-v3 [20], CaffeNet [21] and AlexNet [22].

The loss function plays a critical role in training deep learning models which makes the distances between the different pairs of person images larger than those between the pairs of similar images [23]. In general, loss functions are divided into pairwise loss functions and triplet loss functions [10][24][25]. A single or a combination of loss functions can be employed in deep learning model for person Re-ID task and the proper selection for them can add significant improvement in the overall performance of person Re-ID systems.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2957336, IEEE Access

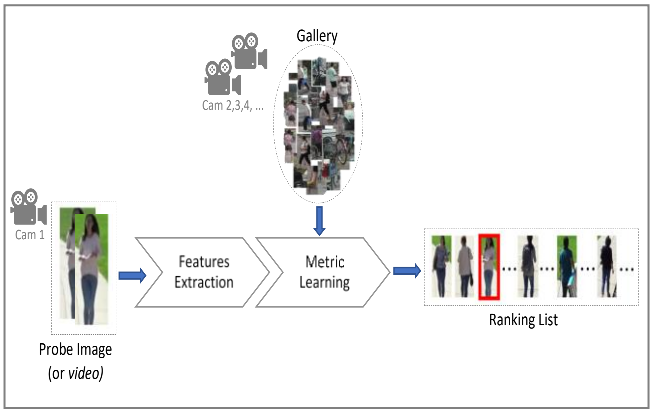

Figure 1: Traditional Person Re-ID system

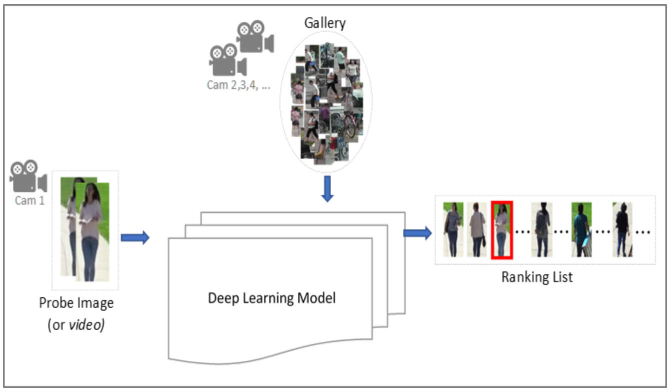

Figure 2: Deep Learning Person Re-ID system

Most of the existing studies are based on hand-crafted methods. However, such features are not discriminative enough, and they are not invariant against inter-class and intra-class variations. Therefore, in this survey, only deep learning methods are discussed because they have achieved a significant improvement in person Re-ID systems comparable with hand-crafted methods.

III. Deep Person Re-ID Systems Categories

There are three types of probe input and gallery set: 1) the probe is a static image and the gallery is a set of images. 2) The probe is a sequence of video frames and the gallery is videos frames. 3) The probe is the static image and the gallery is video frames. The deep learning methods are categorized under three categories based on these three types of probe input and gallery set. The studies in this review are divided into three categories: image-based person Re-ID, video-based person Re-ID and image to video person Re-ID, (see Figure 3). In general, these studies have gone into two directions, some of them improve person Re-ID by building a new special deep learning model to extract robust features, and the others used existing models but focus on loss functions to enhance the person Re-ID.

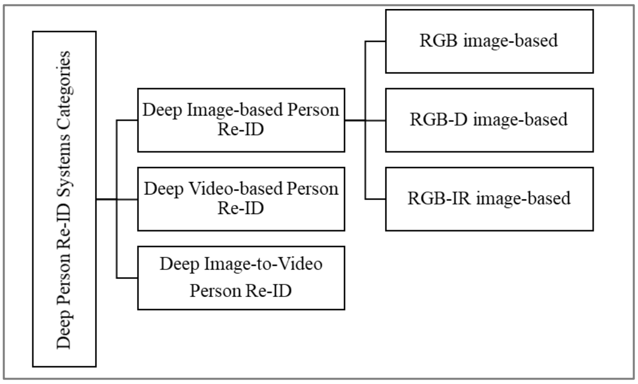
Deep Person Re-ID Systems Categories
Deep Image-based Person Re-ID
Deep Video-based Person Re-ID
Deep Image-to-Video Person Re-ID
RGB image-based
RGB-D image-based
RGB-IR image-based

Figure 3: Deep Learning Categories for Person Re-ID Systems

A. Deep Image-based person re-ID systems

Image-based person Re-ID has attracted more attention in person Re-ID research. It focuses on matching static person images across disjoint camera views. The commonly used features for this task are the appearance-based features, such as clothes. The image-based person Re-ID studied from many aspects under different scenarios to improve its results. One of these aspects is dividing the image-based person Re-ID models into three types based on the type of image: RGB, RGB-Depth and RGB-Infrared images.

1) RGB image-based person Re-ID

The first study used deep learning model for person Re-ID was proposed by authors in [26]. They introduced a deep filter pairing neural network, called DeepReID. This model contained six layers which can handle misalignment, occlusion, noisy background, photometric and geometric transformations. The first layer was a convolution layer, it was used for extracting the local features from the image. Then, max-pooling layer was adopted to make the features more robust against local misalignment. After that, a patch matching layer was added to match the filter output of local patches across different views. To increase the patch matching robustness, a maxout-grouping layer was adopted, it divided patch displacement matrices into a number of groups, and only the maximum activation value in each group was selected to form a single output and passed to the next layer. Then, another convolution layer and max-pooling layer were added to extract the local features of body parts on a larger scale. The Filter pairs and the maxout-grouping layer were used to learn photometric transforms. A patch matching layer, a convolution, max-pooling layers and fully connected layer were used for learning geometric transforms. Finally, the fully connected layer and softmax loss function were adopted to measure if the two person images were similar or not. A large dataset named CUHK03 was built which contains 13,164 images for 1,360 persons. A cumulative matching characteristic (CMC) was used for performance evaluation. Experiments were conducted using both labeled and detected persons bounding boxes. It achieved 20.65% at rank-1 in labeled dataset and 19.89% in the detected dataset.

To improve the re-identification of a person, the authors in [27] proposed an improved deep learning architecture for person Re-ID. This architecture used tied convolutions layer rather than single convolutions layer to find the local relations among the two input images. It started with double layers of tied convolution and max-pooling to extract high-level features for each of two input images. Then a novel layer was adopted that found the differences between features of the two views by comparing the features from one input image with the features extracted from adjacent locations of the other image. Then, a patch summary layer was added to summarize all the maps of the neighboring differences that produced from the previous layer by extracting their local differences into a small patch summary feature. After that, an additional convolution and max-pooling layers were added to find the spatial relationships across neighboring differences. Finally, another two fully connected layers and softmax loss function were added to determine whether both images referred to the same person or not. Experiments were conducted using three datasets: CUHK03, CUHK01 and VIPeR. The cumulative match characteristic (CMC) was applied for evaluating the model. In CHUK03 labeled dataset, it achieved 54.74% at rank-1. In CHUK03 detected dataset, it achieved 44.96% at rank-1. In CUHK01 dataset with 100 test identities, it achieved 65% at rank-1. In CUHK01 dataset with 486 test identities, it achieved 40.5% at rank-1. In VIPeR dataset, it achieved 34.81% at rank-1.

To extract better features from person image, authors in [28] introduced multi-channel parts-based CNN with improved triplet loss function to extract the features of overall body and local body parts. It contained one convolution layer as a global layer, one convolution layer for complete body, four convolution layers for body parts, five channel-wise full connected layers, and one network-wise full connected layer. A full body channel with the entire global convolution layer was used to extract global full body features and four convolution layers were used to extract local features for a person's body parts. Triplet images were used to train the model. The improved loss function was used, and it helped to bring the images of the same person closer together and push the images of the different persons faraway from each other. The four widespread person Re-ID datasets were used, i-LIDS, PRID2011, VIPeR and CUHK01. The cumulative match characteristic (CMC) was used for evaluating the model. In i-LIDS dataset, the results at rank-1 to rank-20 ranged from 60.4% to 97.8%. In PRID2011 dataset, the results at rank-1 to rank-20 ranged from 22% to 57%. In VIPeR dataset, the results at rank-1 to rank-20 ranged from 47.8% to 91.1%. In CUHK01 dataset, the results at rank-1 to rank-20 ranged from 53.7% to 96.3%. It was observed that the body parts with the person's face and shoulder achieved the best performance, while with the lower parts of the body, the performance gradually decreased, and the body parts that contained the person legs and feet achieved the lowest performance.

Since some human body parts have very important features to distinguish different persons, authors in [29] proposed a parts-based approach called DeepDiff for learning deep difference features on human body parts. It focused on dividing the body into many parts and extracting the deep features from these parts using three deep neural subnets networks. Each subnet can handle a type of intra-class variations. Using pyramid partition architecture, the image of person body was divided into 12 parts via Deep Decompositional Network. Then, the first subnet was used to extract difference features from input data. The second subnet was used for extracting difference features from features maps to decrease the impact of variations. The last subnet was applied to extract difference features from spatial variations. The similarities between those corresponding parts were evaluated using softmax to make the final decision, whether the two images belong to the same person or not. CUHK03, CUHK01 and VIPeR datasets were used for experiments. The cumulative match characteristic (CMC) was applied for evaluating the performance of the model on these datasets. In CUHK01 dataset, the results at rank-1 to rank-20 ranged from 47.9% to 86.9%. In CUHK03 labeled dataset, the results at rank-1 to rank-20 ranged from 62.4% to 96.7%. In CUHK03 detected dataset, the results at rank-1 to rank-20 ranged from 54.8% to 95.9%. In VIPeR dataset, the results at rank-1 to rank-20 ranged from 43.2% to 86.1%. This model helped to gather more information and extracted different features between the local regions.

From the concept of that the full-body and parts-body features complete each other, the authors in [30] proposed person re-identification model with human body region guided feature decomposition and fusion, called Spindle Net. In this model the input image was prepared by using Region Proposal Network (see Figure 4). The Region Proposal Network process started with detected the 14 body joints from the input image of a person by Convolutional Pose Machines with improvements to decrease the model complexity. Then, Region Proposal Network was getting 7 body sub-regions (3 macro and 4 micro sub-regions) depended on body joint locations that were previously detected. The macro sub-regions were head-shoulder, upper body and lower body, and the micro sub-regions were two arms and two legs. Then, Spindle Net model was applied; it consisted of two main components. The first component was the Feature Extraction Network, it took the person image together with the sub-regions as input and computed one global feature vector of the complete person image and seven sub-regions feature vectors for the seven body sub-regions. It contained three convolution stages and two pooling stages. The second component was Feature Fusion Network which can compute the final feature vector by merging the eight feature vectors together, the feature vector of full person image and the feature vectors of body sub-regions, by using a tree-structured fusion strategy.

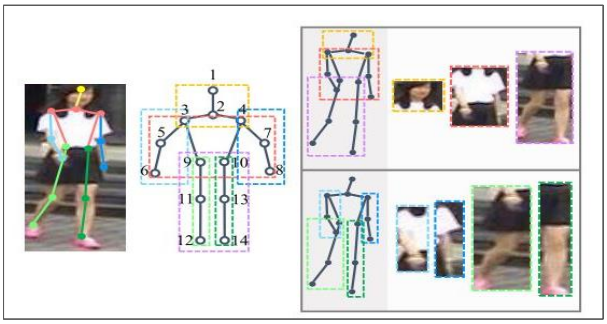

Figure 4: The Region Proposal Network [30]

The last features vector was used to decide about various persons by using softmax loss. This model was applied in seven public datasets and cumulative match characteristic (CMC) was adopted for evaluation. In CUHK03, the results at rank-1 to rank-20 ranged from 88.5% to 99.2%. In CUHK01, the results at rank-1 to rank-20 ranged from 79.9% to 98.6%. In PRID, the results at rank-1 to rank-20 ranged from 67.0% to 92.0%. In VIPeR, the results at rank-1 to rank-20 ranged from 53.8% to 92.1%. In 3DPeS, the results at rank-1 to rank-20 ranged from 62.1% to 95.7%. In LIDS, the results at rank-1 to rank-20 ranged from 66.3% to 95.3%. In Market-1501, the results at rank-1 to rank-20 ranged from 76.9% to 96.7%. In their proposed SenseReID dataset, the results at rank-1 to rank-20 ranged from 34.6% to 66.7%.

Instead of using the defined rigid parts of the body for extracting full and parts body features, authors in [31] introduced deep multi-scale context-aware model for learning the features over body and latent parts. This model consisted of a multi-scale convolutional network with Spatial Transformer Networks rather than directly using a single-scale convolution layer and pooling layer for efficient feature learning across latent body parts. It contained an initial convolution layer to extract the low-level global features maps of the body-based representation. Then for part-based representation, four multi-scale convolution layers were used to get the composited image context information. Dilated convolution was adopted for the convolution filter with different ratios to capture different scale context information. Then all dilation features maps were concatenated to get final output of the existing convolution layer. These outputs were embedded on a layer by layer convolution operation. At last, overall body features and local body parts features were integrated as the final person representation. Spatial Transformer Networks was adopted for localizing three body parts corresponding to the head-shoulder, upper body and lower body parts. It contained two elements, the spatial localization network to find the transformation parameters, and the grid generator to illustrate the probe image using a bilinear image interpolation kernel. So, this model can detect the proper latent parts automatically for feature extraction. Softmax loss function was adopted for identity prediction tasks. This model was evaluated on the main person Re-ID datasets: Market1501, CUHK03 and MARS. The cumulative match characteristic (CMC) was adopted for performance evaluation. In Market1501 with a single query, it achieved 80.31% at rank-1 and 86.79% for multiple queries. In CUHK03 detected dataset, the results at rank-1 to rank-20 ranged from 67.99% to 97.83%. In CUHK03 labeled dataset, the results at rank-1 to rank-20 ranged from 74.21% to 99.25%. In MARS, the results at rank-1 to rank-20 ranged from 71.77% to 93.08% for one query and from 83.03% to 66.43% for multiple queries.

Inspired by recent researches on Neural Architecture Search (NAS) [32], authors in [33] introduced a part-aware module for the person Re-ID called Auto-ReID and used it as a baseline for building a number of optimal Re-ID architectures in NAS search space. This was the first work for automated NAS for Re-ID task that considered body structure. This model started by dividing input feature tensor into four body parts and extracted the features from them. Then, a self-attention technique was applied on features vectors to get more distinctive body parts features at each part. After that, each part of body features vectors was repeated and joined to restore them into the same spatial form of input. The resulted global feature tensor from the previous step was fused with the original input tensor by a one-by-one convolutional layer to produce the final output. retrieval loss is adopted which consisted of cross-entropy and triplet losses. Macro structure of ResNet [17] was used as a baseline network. This module was applied in public datasets: Market-1501, CUHK03 and MSMT17. The cumulative match characteristic (CMC) and mean average precision (mAP) were adopted for evaluation. In Market1501, it achieved 94.5% (85.1% mAP) at rank-1. In CUHK03, it achieved 77.9% (73.0% mAP) at rank-1 for labeled dataset and 73.3% (69.3% mAP) for detected one. In MSMT17, the results at rank-1 to rank-10 ranged from 78.2% to 91.1% (52.5% mAP).

Since the critical impact of the loss function in person Re-ID systems, authors in [34] defended the triplet loss for person Re-ID. They trained CNN with triplet loss for proving that a good implement of triplet loss has a critical impact on the performance of person Re-ID system. Variants of triplet loss were used, and then the setting that worked best was identified for person re-identification. The selected triplet losses were tested on a pre-trained network and a network trained from scratch. The batch hard and batch all were proposed, they corresponded to the standard triplet loss. In pre-trained network, ResNet-50 architecture was used and replaced the last layer by two fully connected layers for Re-ID task. For a trained network from scratch, a network called LuNet was designed. LuNet followed the style of ResNet-v2 with some improvements in the layers. Market-1501 and MARS dataset were used for evaluating both pre-trained network and the network that was trained from scratch, and the CUHK03 dataset for pre-trained network. The cumulative match characteristic (CMC) and mean average precision score (mAP) were adopted for measuring the performance. By comparing the results for the different formulations at several margin values, the best results were achieved by the soft-margin variations of the batch hard loss. In MARS dataset, on pre-trained network, it achieved 79.80% at rank-1 and 91.36% at rank-5 (67.70% mAP); On LuNet, it achieved 75.56% at rank-1 and 89.70% at rank-5 (60.48% mAP). In Market-1501 with a single and multi-query, the results at rank-1 to rank-5 ranged from 84.92% to 96.29% (from 69.14% to 76.42% mAP) on pre-trained network; On LuNet, the results at rank-1 to rank-5 ranged from 81.38% to 95.16% (from 60.71% to 69.07% mAP). In CHUK03 with labeled dataset, it achieved 89.63% at rank-1 and 99.01% at rank-5 on the pre-trained network. In CHUK03 with the detected dataset, it achieved 87.58% at rank-1 and 98.17% at rank-5 on the pre-trained network. The experiments approved that used a triplet loss (especially batch hard) achieved advanced results with a pre-trained model and with a model that was trained from scratch.

The triplet loss has a limitation, it cannot make a complete using of batch information, therefore there is a need to manually select hard negative samples which consume a lot of time. To overcome this problem, authors in [35] adopted lifted structured loss for learning deep features embedding. The loss function of the proposed network was built according to the combination of lifted structured loss and the identification loss to examine together the relative information of the image (positive or negative) and the correct identity information. A One-branch CNN model was used, which contained 9 convolutional layers, 4 max-pooling layers, 2 fully connected layers, and a softmax loss function. The experiments were performed on publicly existing datasets, Market-1501, CUHK01, CUHK03 and VIPeR. The cumulative match characteristic (CMC) was adopted for evaluation. In Market-1501, the results at rank-1 to rank-10 ranged from 84.53% to 95.58%. In CUHK03 labeled dataset, the results at rank-1 to rank-10 ranged from 81.6% to 98.9%. In CUHK03 detected dataset, the results at rank-1 to rank-10 ranged from 79.9% to 98.7%. In CUHK01, the results at rank-1 to rank-10 ranged from 70.2% to 95.5%. In VIPeR, the results at rank-1 to rank-10 ranged from 47.3% to 88.1%.

Most of the previous studies trained the deep model with pairwise or triplet loss using Euclidean distance metric to compute the similarity. However, Euclidean is suboptimal for complicated features in images with large variations. So, authors in [36] introduced deep adaptive feature embedding model with local sample distributions. In this model, the distance metric was adapted into local range and found the appropriate positive samples toward getting a robust deep embedding in the situation of large intra-class variations. The Convolutional Restricted Boltzmann Machines [37] was used to extract the appearance features from person images. Then, the similarity was computed, and the one hard quadruplet was chosen from the local range of positive and negative pairs. After that, the features were extracted from each sample in hard quadruplet by Convolutional Restricted Boltzmann Machines. The specific local loss was adopted which it was an extension of triplet loss. It offered more triplet pairs to compute similarity based on a large margin criterion. For improving the efficiency of training, a neighborhood based gradient memorization method was applied to reuse previous gradients. The experiments were done on Market-1501, CUHK01, CUHK03 and VIPeR. The cumulative match characteristic (CMC) was adopted for evaluation. In Market-1501, the results at rank-1 to rank-20 ranged from 84.14% to 98.07%. In CUHK03, the results at rank-1 to rank-20 ranged from 73.02% to 98.58%. In CUHK01, the results at rank-1 to rank-20 ranged from 71.60% to 97.25%. In VIPeR, the results at rank-1 to rank-20 ranged from 49.04% to 96.20%.

While hard triplet loss consumed a lot of time and picking more hard triplets affected in the training process, the authors in [38] went beyond the triplet loss and proposed a deep quadruplet network. In this work, a quadruplet ranking loss to improve the lack of triplet loss by considering two aspects in one quadruplet. The first aspect is finding correct orders for pairs and the second aspect was focusing on push negative pairs faraway from positive pairs. This resulted in higher inter-class changes and smaller intra-class changes; therefore, the performance was improved. To handle the weakness of normalization, a fully connected layer with a 2-dimension output was added. Then, the softmax loss function was adopted to normalize the 2-dimension output and send one-dimension to triplet loss. Given the fact that the triplet loss is training a model just according to a relative distance between positive and negative pairs, the quadruplet loss presented a new constraint that considered the orders of positive and negative pairs with various input images. By using this new constraint, the lowest inter-class distance has to become larger than the highest intra-class distance even if the pairs contained the same input images. The adaptive margin threshold was adopted to describe the average distance of the two distributions, it must have a positive relation with the average distance and avoidance over or under the sampling problems. A pre-trained AlexNet model was used. The experiment was done with three datasets including CUHK03, CUHK01, and VIPeR. The cumulative matching characteristic (CMC) was adopted for evaluation. The experiments were conducted with various loss functions and provide numerous baselines to explain the efficiency of each component in the proposed method. When comparing the performance between improved triplet without softmax and improved triplet with softmax, the research concluded that applying the softmax added a little enhancement in the performance of the triplet loss. By applying the new constraint on all datasets, the quadruplet had better performance than the improved triplet losses. Also, it was obvious that when used the network with the proposed margin-based online hard negative mining, the results of the quadruplet were improved. In CUHK03, the results at rank-1 to rank-10 ranged from 75.53% to 99.16%. In CUHK01(486 person), the results at rank-1 to rank-10 ranged from 62.55% to 89.71%. In CUHK01(100 person), the results at rank-1 to rank-10 ranged from 81% to 98%. In VIPeR, the results at rank-1 to rank-20 ranged from 56.11% to 96.97%.

To overcomes the generalization limitations of the triplet loss, the authors in [39] introduced A Discriminatively Learned CNN Embedding and a similarity metric model which integrated identification loss and verification loss. This model was built on Siamese network architecture. It consisted of two pretrained CNN models, three extra convolutional layers, one square layer, and three losses. The pretrained model ResNet-50 was used as baseline network but the fully connected layer was replaced by convolution layer and softmax. A cross-entropy loss was used for identity prediction. A nonparametric layer called the square layer was used to compare the high-level features for computing the similarity. For experiments, Market1501, CUHK03 and Market1501+500k dataset were used. Also, this model was tested on Oxford Buildings which it is a popular images retrieval dataset. The cumulative match characteristic (CMC) was used for the performance evaluation. In Market1501, the results at rank-1 was 79.51% for single query and 85.84% for multiple query. In CUHK03, the results at rank-1 to rank-10 ranged from 83.4% to 98.7% by Single-Shot. In CUHK03, the results at rank-1 to rank-10 ranged from 88.3% to 97.8% by multi shot. In Market1501+500k, it achieved 68.26% at rank-1. In Oxford5k, the accuracy results by CaffeNet or VGG16 ranged from 66.2% to 76.4%. In [40], authors proposed four stream Siamese convolution neural network with joint verification and identification loss. A quartet loss function was introduced for training the model. The four images were used as input, two images are matched, and the other two images are mismatched. The components of the verification and identification models were integrated to improve person re-identification accuracy, which inter-class variations increased by identification model, and the intra-class variations decreased by verification model. A pre-trained network AlexNet was used as baseline network. For verification, the features were extracted from low-level layers; whereas, for identification, the features were extracted from higher-layers. The identification loss was determined using a softmax layer for classification. The cross-entropy loss was used for predicting the true identity. In the experiments, four person Re-ID datasets were used: VIPeR, CUHK03, CUHK01 and PRID2011. The cumulative match characteristic (CMC) was used for the performance evaluation of the proposed model. In VIPeR, the results at rank-1 to rank-10 ranged from 68.7% to 94.6%. In CUHK03, the results at rank-1 to rank-10 ranged from 85.5% to 99.8%. In CUHK01, the results at rank-1 to rank-10 ranged from 83.95% to 98.97%. In PRID2011, the results at rank-1 to rank-10 ranged from 75% to 97%. Person Re-ID performance was improved when considered it as a verification task and identification task instead of considering it as each one independently.

To integrate spatial relationship into feature learning, several studies employed RNN in their model. Authors in [41] introduced Deep Spatially Multiplicative Integration Networks. This model consisted of two-stream CNNs, M-Net [42] and D-Net [43], to extract the features. Then, the output from each CNN were fused by multiplicative integration gate. The result from previous step was sent to a stacked four-directional recurrent layers to find the spatial relationships. To compute the similarity, the cosine function and binomial deviance loss function were adopted. Three common datasets CUHK03, Market-1501, and VIPeR were used for experiments. The cumulative matching characteristic (CMC) was adopted for evaluation. In Market-1501, the results at rank-1 to rank-20 ranged from 67.15% to 97.08%. In CUHK03, the results at rank-1 to rank-20 ranged from 73.23% to 97.52%. In VIPeR, the results at rank-1 to rank-20 ranged from 49.11% to 93.47%. In [44], authors presented a deep visual attention model with efficient spatially recursive encoding structure for fine-grained visual recognition. This model can detect distinctive parts of image and convert them into spatial representation. Each image features were extracted by two-stream Convolution neural networks CNN, M-Net and D-Net. Then, the outputs from the previous step were integrated by bilinear pooling at each location. The spatial LSTMs with visual attention convert recursively the bilinear pooling outputs into spatial representations and hidden states were generated as feature representation. This feature representation was sent to a softmax layer which cross-entropy loss was used for classification. For experiments, three datasets including CUHK03, Market-1501, and VIPeR were used. The cumulative matching characteristic (CMC) was adopted for evaluation. In Market-1501, the results at rank-1 were 64.23%. In CUHK03, the results at rank-1 to rank-20 ranged from 65.23% to 98.52%. In VIPeR, the results at rank-1 to rank-20 ranged from 56.11% to 96.97%.

Authors in [45] proposed Cross-Entropy Adversarial View Adaptation Framework. In this framework, two asymmetric mappings, probe mapping and galley mapping, were learned using M-Net and D-Net which each one dependent on the other. To minimize the distance between probe and gallery mapping and make view-invariant feature space, cross-entropy adversarial mapping loss was adopted. Then, a similarity discriminator network with a margin-based separability was adopted on the Euclidean distance of positive and negative pairs to find the effective similarity metrics. Four datasets were used for experiments: VIPeR, CUHK03, Market-1501, and DukeMTMC-reID. The cumulative matching characteristic (CMC) was adopted for evaluation. In VIPeR, the results at rank-1 to rank-20 ranged from 55.9% to 97.7%. In CUHK03, the results at rank-1 to rank-20 ranged from 88.9% to 99.9%. In Market-1501, the results at rank-1 to rank-20 ranged from 89.1% to 99.7%. In DukeMTMC- reID, the results at rank-1 to rank-20 ranged from 80.1% to 96.9%.

Pedestrian misalignment is one of the critical problems in person Re-ID (see Figure 5). Most existing studies focused on extracting the feature vectors from high-level convolution layers to overcome affine transformations, but this method cannot handle misalignment problems such as pose variations. Authors in [46] proposed a new Adaptive Alignment Network to handle the misalignment challenge to get the accurate and

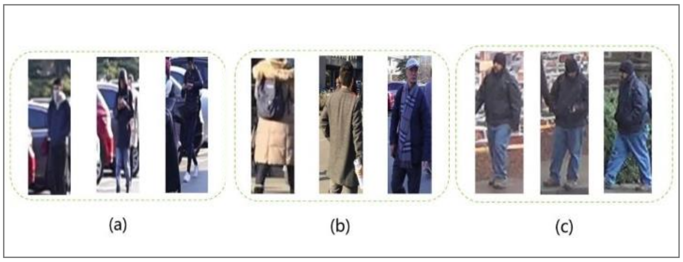

Figure 5: Misalignment challenge examples: (a) noisy background, (b) partial person body, (c) person pose changes [46].

robust person Re-ID by learning both patch-wise and pixel-wise alignments from coarse to fine. Adaptive Alignment Network consisted of a base network that was built on ResNet-50, a patch alignment module and a pixel alignment module. The patch alignment module contained two convolution layers followed by two fully connected layers. It partitioned the features map from the base network into multiple patches and estimated the alignment offset of each patch. It then conducted a patch-wise alignment based on the learned offsets. The pixel alignment module was used for fine-grained pixel-wise alignment. It also contained two convolution layers followed by two fully connected layers. It learned the local offset for each pixel within a patch and produced an accurately aligned feature map. Cross-entropy loss was used to predict a number of probabilities. By aligning pedestrian images, Adaptive Alignment Network can learn more discriminative and robust pedestrian features and enhanced the performance of person Re-ID. Extensive experiments were done on common datasets: Market1501, DukeMTMC-reID and MSMT17. The cumulative match characteristic (CMC) was used for model evaluation. The additional evaluation metric was the mean average precision (mAP). In MSMT17 dataset, it achieved 70.5% at rank-1 and 82.8% at rank-5 and 86.9% at rank-10 (40.9% mAP). In the Market1501 and DukeMTMC-reID, it achieved 92.0% (78.2%mAP), 84.1% (66.4%mAP) at rank-1 respectively.

2) RGB-depth image person Re-ID

To improve the performance of image-based person Re-ID models, the number of recent studies proposed hand-crafted methods to extract new types of features from depth images. These types of features were invariant against many variations such as illumination changes. Depth images captured by using RGB-D sensors like Kinect sensors. Then these features were combined with the appearance features to enhance the Re-ID accuracy. The features that were extracted from depth image such as skeleton data and human body shape as in [47] where the authors proposed robust depth-based person Re-ID model, and in [48] where authors introduced online Re-ID model which pre-trained the metric model offline and then updated it online. In [49] authors introduced the two types of histograms as features that were extracted from the depth images.

Many researchers started to use deep learning with depth images, but there are still few studies in this area. The first work used deep learning with depth images for person Re-ID task was proposed by the authors in [50]. They introduced multi-modal uniform deep learning for RGB-D person Re-ID. They used a deep learning model for extracting appearance and anthropometric features from RGB-D images. Depth image and color image are taken as input, appearance features were extracted by one CNN and anthropometric features were extracted by another CNN. Then, to combine both types of features (appearance and anthropometric), a multi-modal fusion layer was used with a uniform latent variable that was noise resistant. The hinge loss function was adopted for classification. The two existing dataset, Kinect-REID and RGBD-ID were used to present how this model was effective and robust. The cumulative match characteristic (CMC) was used for performance evaluation. It is observed that the depth features were more effective than appearance features in some specific conditions such as clothes changing. In comparison with other state-of-the-art models, In Kinect-REID dataset, it achieved 97.0% at rank-1, 100% at rank-5 and 100% at rank-10. In the RGBD-ID dataset, it achieved 76.7% at rank-1, 87.5% at rank-5 and 96.1% at rank-10.

3) RGB-Infrared image person Re-ID

The RGB cameras that captured RGB images cannot capture clear appearance features under low illumination environments such as at night. This limitation led the researchers to start using thermal images (infrared images) to improve the person Re-ID performance (see figure 6). Some studies proposed hand-crafted models such as in [51], where the authors proposed a tri-modal Re-ID system based on RGB, depth, and thermal features.

Using deep learning, the authors in [52] proposed the first model for handling RGB-Infrared cross-modality person Re-ID problem and introduced a new multi-modal Re-ID dataset called SYSU-MM01. It used three common neural network structures including one-stream, two-stream and asymmetric fully connected layer. Also, it proposed a new model called a deep zero-padding with one-stream. It achieved the best performance using a deep zero-padding method. The performance was measured by the cumulative match characteristic (CMC) and the mean average precision (mAP). In single-shot SYSU-MM01 dataset, it achieved 14.80% at rank-1 and 54.12% at rank-10 (15.95% mAP). With multi-shot, it achieved 19.13% at rank-1 and 61.40% at rank-10 (10.89% mAP). In [53], authors proposed end-to-end dual-path network with a novel bi-directional dual-constrained top-ranking loss to extract the features from the path of visible image and the path of the thermal image. This was one of the earliest works for end-to-end visible-thermal person Re-ID. The dual-path feature learning network contained two parts: the first one was feature extractor and the other one was feature embedding. For features extraction, AlexNet was adopted as the baseline network for both paths to capture features from different image modalities.

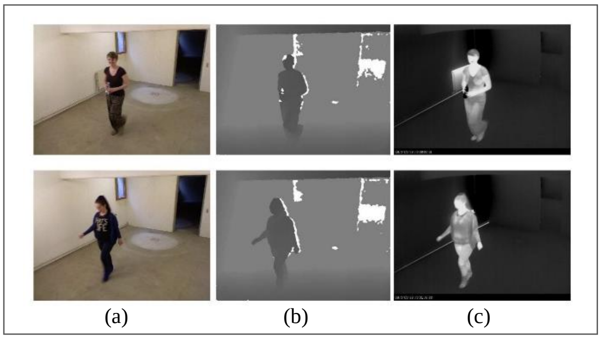

Figure 6: Examples of a) RGB, b) depth, and c) thermal images[51].

For feature embedding, a fully connected layer was embedded as a shared layer on top of both paths to bridge the gap between two varied modalities. Then, a novel bi-directional dual-constrained top-ranking loss was adopted to guide the feature learning objectives. It had cross-modality and intra-modality constraints. Cross-modality constraints compared the distance of a positive visible-thermal pair and the lowest distance of all negative visible-thermal pairs, rather than each of the negative pairs. Then intra-modality constraints were added to ensure that the hardest cross-modality negative sample should also be far away from its corresponding cross-modality positive samples. Finally, the softmax loss was adopted for classification. The two public datasets were used to evaluate the model, RegDB and SYSU-MM01 datasets. The performance was measured by cumulative match characteristic (CMC) metric and the mean average precision (mAP). By comparing this model with other existing models, In RegDB dataset, the results at rank-1 to rank-20 ranged from 33.47 % to 67.52% (31.83% mAP). In the SYSU-MM01-ID dataset, the results at rank-1 to rank-20 ranged from 17.01% to 71.96% (19.66% mAP).

B. Deep Video-based person Re-ID systems

The video is more realistic to perform person re-identification. Video-based person Re-ID matches a sequence of video of a person in one camera with a gallery of video sequences recorded by other non-overlapping cameras. The re-identification of a person in videos has not received the same attention of image-based Re-ID. The usage of video for reidentifying persons has many advantages over static images; it provides continuous images that contain appearance information about the same person, and this is effective for decreasing the impact of some ambiguous situations. Also, not only appearance features are extracted, the temporal features associated with the motion of person across frames are also extracted, such as captured their gait. This can enhance the performance of person Re-ID task. Despite these advantages, there are new challenges appearing such as the lack of large video datasets for Re-ID purposes, the different length of video sequences and unstable person tracking. In general, Conventional Neural Network and Recurrent Neural Network (CNN and RNN) architecture are adopted. The appearance features are extracted by CNN and then the temporal features are extracted by RNN. Video-based person Re-ID models with RGB input will be discussed in this survey.

The first video-based person Re-ID model using deep learning was proposed by authors in [54]. They introduced the recurrent convolutional network that used both color and optical flow features, the color described the appearance of the person, and optical flow described short-term motion, containing the person gait and other motion features. By using the combination of color and optical flow, the model must be better able to use the short-term temporal features to enhance the accuracy of re-identification. At each timestep, a CNN was used for extracting a high-level representation from the image and send the output of CNN to RNN layer. RNN output depended on both the current input and information from the earlier timesteps frames, RNN can remember information over time. Then, temporal pooling was used to summarize the appearance representation of the complete sequences with different video length and frame rates into a single feature vector and avoiding bias towards later time steps that happen by RNN. They applied two approaches of temporal pooling to produce one feature vector; average-pooling and max-pooling to select the average or maximum activation of each element of the feature vector. For training, a Siamese network was used to map image sequences of the same person to closer feature vectors. The cross-entropy function was used for identity prediction. The jointly training for Siamese and identification cost had a critical impact in overall training result. Two different datasets were used for experiments: iLIDS-VID and PRID-2011. The cumulative match characteristic (CMC) was used for performance evaluation. By Comparing this model with the other recent models, in iLIDS-VID, the results at rank-1 to rank-20 ranged from 58% to 96%. In PRID-2011, the results at rank-1 to rank-20 ranged from 70% to 97%. The best performance occurred when recurrent connections were allowed, and when the combination of optical flow and appearance features were used. Also, average-pooling achieved better results than max-pooling.

The success of the fusion between the full-body features and part-body features in image-based Re-ID guided the authors in [55] to propose a deep end-to-end spatial and temporal fusion network. It captured full-body and part-body features to learn the fusion between spatial features and temporal features for video-based person Re-ID. The input was raw images and the optical flow data, the person image was divided into the upper part and lower part. Then, the extracted features from the upper and lower part were fused to get part-based features. Next, the part-based features were fused with the global features, where the global features were extracted from the full image sequences, to get the final feature vector. The CNN layer extracted high-level representation, and then the output of CNN was sent to the RNN layer to get the final temporal features. The temporal pooling was used to combine information across all time steps and avoid bias towards the latest time steps. The Siamese was adopted with a softmax loss layer for classification. For the experiments, three common datasets on the video-based person Re-ID were used:

PRID-2011, i-LIDS-VID and MARS. The cumulative match characteristic (CMC) was used for measuring the performance. By Comparing this model with the other recent models, in PRID-2011, the results at rank-1 to rank-20 ranged from 77% to 98%. In iLIDS-VID, the results at rank-1 to rank-20 ranged from 61% to 97%. In MARS, the results at rank-1 to rank-20 ranged from 71% to 96%. The performance of the combination between CNN and RNN was better than using only CNN or RNN. The performance of fusion between global and local features was better than using only global or local features which proving the complementarities between global features and local features improved the re-identification accuracy. In [56], the authors introduced end-to-end accumulative motion context network. It was a two-stream convolution architecture that consisted of a spatial network in the first stream and temporal network with another spatial network in the second stream. The spatial appearance features were extracted by first stream whereas the temporal features between two sequential frames were extracted by second stream. After that, both stream features were combined in a recurrent way to learn the distinctive accumulative motion contexts. Since the length of the input sequence was random, the motion context was also inconstant for each person. So, RNN was adopted to handle a randomly long time series. The output of RNN may be influenced by later time-steps more than earlier ones which making RNN efficiency decreased when needing to gather information on longer time steps. To solve this problem, a temporal pooling layer was added after RNN to capture long term information presented in the entire sequence that merged with the motion context information accumulated through RNN. Average pooling or max pooling was used across the temporal dimension. Finally, both streams of sub networks of two sequences from two different cameras were built as the Siamese network architecture. Multi-task loss functions were adopted consisting of contrastive loss and softmax loss. Three public benchmarks were used to evaluate the model, iLIDS-VID, PRID-2011 and MARS. The optical flow maps were extracted by EpicFlow algorithm that increases the performance of this model. This model achieved the best performance by the fusion between spatial and temporal features. The average-pooling was better than the max-pooling in the temporal pooling step. The cumulative match characteristic (CMC) was used for performance evaluation. By comparing the proposed model with the recent existing models, In PRID-2011, the results at rank-1 to rank-20 ranged from 83.7% to 100%. In iLIDS-VID, the results at rank-1 to rank-20 ranged from 68.7% to 99.3%. In MARS, the results at rank-1 to rank-20 ranged from 68.3% to 90.6%.

However, the previous aforementioned approaches did not study the case that the same person sometimes could walk in variable speed across different camera views. To handle these issues, the authors in [57] introduced two stream multi-rate recurrent neural networks for capturing the spatial features and temporal features and handling the motion speed change. To extract features, VGG and CNN_M networks were adopted, the spatial stream was built for extracting the appearance features existing in video frames, and the motion stream was built on the optical flows and extracted between all pairs of sequential video frames. Then these features were inserted into a multi-rate gated recurrent unit (MR GRUs) for temporal modeling, which can deal with speed variance. GRU is a type of RNN that make each recurrent unit to capture dependencies of different time scales. Average pooling was used to fuse these two types of features and get the video-level representations. Two real-world datasets were used for evaluation, iLIDS-VID and PRID-2011 dataset. For all the datasets, the performance was evaluated by the cumulative match characteristic (CMC). In comparison their model with the recent existing models, in PRID-2011, the results at rank-1 to rank-20 ranged from 78.7% to 99.2%. In iLIDS-VID, the results at rank-1 to rank-20 ranged from 59.4% to 99.1%.

By observation, we find that not all images are informative, and there are useful features may be abandoned. Therefore, in [58], the authors proposed joint spatial and temporal recurrent neural networks. It was an end-to-end deep neural network architecture that measured the importance of each video frame and selected only the more informative frames. Its input was a triplet of image sequences. At each stream, a CNN was used for extracting the features of each image, CaffeNet was adopted for CNN. Then, an attention technique was implemented to find the temporal structure of the given image sequence which it had two parts: the attention unit to find the importance of each frame, and the RNN unit where LSTM network was implemented to learn feature representations. After these two parts, temporal average pooling was used. Then, Spatial Recurrent Model was adopted as metric learning. It had 6 spatial RNNs, each one of them moved the features map over a specified direction, and the output of each spatial RNN was placed together. In the end, convolution layers and fully connected layers were added to extract higher-order spatial relations within the contextual features. Thus, each location in the features map of the convolution layer was a combination of its six-surrounding information. A triplet loss was applied to decrease the distance between similar pairs and increase the distance between different pairs. Three public datasets were used to evaluate the model, iLIDS-VID, PRID-2011 and MARS. The performance was evaluated by the cumulative match characteristic (CMC). In comparison the proposed model with the recent models, In PRID-2011, the results at rank-1 to rank-20 ranged from 79.4% to 99.3%. In iLIDS-VID, the results at rank-1 to rank-20 ranged from 55.2% to 97.0%. In MARS, the results at rank-1 to rank-20 ranged from 70.6% to 97.6%.

To boost the local spatial features and make them more discriminative and focusing on more related features, authors in [59] proposed deep Siamese attention networks for improving the spatial representation. This model started by extracting the features using CNN. GoogLeNet and VGG-16 were used as baseline network in this step. Then, to find the spatial relationships between extracted features, these features were weighted by soft attention and then the spatial relationships were captured using a probability distribution on spatial dimensions. Gated Recurrent Unit (GRU) was adopted as the recurrent units which the output from the previous step was fed into them to find the most important features over time. Average temporal pooling was adopted to integrate the features from the selected frames at all levels to produce the final complete video representation. Cross-entropy loss was adopted for the identification task. Three available datasets were used for experiments, iLIDS-VID, PRID-2011 and MARS. The performance was evaluated by the cumulative match characteristic (CMC). By Comparing this model with the other current models, in PRID-2011, the results at rank-1 to rank-20 ranged from 77% to 99.4%. In iLIDS-VID, the results at rank-1 to rank-20 ranged from 61.9% to 98.6%. In MARS, the results at rank-1 to rank-20 ranged from 73.5% to 97.5%. To fully leverage the temporal features and dealing with the low spatial alignment, authors in [60] proposed model for person Re-ID by temporal residual learning. There were two types of stream in this model: the main stream and the alignment stream to process the main input sequences and the alignment sequences. This model consisted of two main components, the temporal residual learning module and the extended versions of spatial transformer network. The temporal residual learning module which used for extracting the generic and specific features of sequence frames together using two bi-directional LSTMs (BiLSTMs) followed by an average temporal pooling layer. When using BiLSTMs, the joint features provided full information for person descriptions. Also, the information has flexibility in moving forward and backwards to enhance the temporal features. The extended versions of a spatial transformer network used to improve input noises and poor alignments in videos. It contained the localization network part, the grid generator part and the bilinear sampler part, all parts worked to find the optimum parameters of spatial transformation automatically in sequential frames. The localization network part expected the transformation parameters, then these parameters used by grid generator to build sampling grid, finally, the sampling grid go into bilinear sampler to create feature maps of transformations. The GoogLeNet was adopted as the base network and the cross-entropy loss was used to manage the feature learning of both main and alignment streams. The experiments were done on four generally datasets, MARS, PRID-2011, iLIDS-VID and SDU-VID. For all these datasets, the performance was evaluated by the cumulative match characteristic (CMC). In comparison with state-of-the-art models, in PRID-2011, the results at rank-1 to rank-20 ranged from 87.8% to 99.3%. In iLIDS-VID, the results at rank-1 to rank-20 ranged from 57.7% to 94.1%. In SDU-VID, the results at rank-1 to rank-20 ranged from 97.7% to 100.0%. In MARS, the results at rank-1 to rank-20 ranged from 80.5% to 96.0%.

Authors in [61] introduced spatial and temporal mutual promotion model that used temporal information to recover spatial information in one frame by information in the same location of the other frames. The proposed model had two main modules: refining recurrent unit module and spatial-temporal clues integration module. After extracting features via Inception-v3 from each video frames, feature vectors of all video frames were inserted into refining recurrent unit for refinement. This module was used to eliminate noise and refer to previous video frames for using their extracted appearance and motion features to retrieve the missing information. Refining recurrent unit was different from the other recurrent models (RNN or LSTM), it was implemented to improve the features of frame-level by referring the spatial and temporal features as an alternative to extract recent features from temporal feature vectors. The refining recurrent unit had an update gate model which it was used to decide how to update the refined feature. The refined feature maps were then inserted into spatial-temporal clues integration to allow this model to extract simultaneously the appearance features and motion context to generate the final video-level feature representation. The spatial-temporal clues integration was extracted better spatial and temporal features from high-level appearance feature maps produced by refining recurrent unit. With cross-entropy losses, multi-level training objective was implemented to focus on different local parts and improve the ability of both modules. It contained two constraints, the video-level ranking and part-level ranking constraint which both based on batch hard triplet loss. This model was evaluated on public video benchmarks including iLIDS-VID, PRID-2011 and MARS. The performance was measured by the cumulative match characteristic (CMC). In comparison this model with state-of-the-art models, In PRID-2011, the results at rank-1 to rank-20 ranged from 92.7% to 99.8%. In iLIDS-VID, the results at rank-1 to rank-20 ranged from 84.3% to 99.5%. In MARS, the results at rank-1 to rank-20 ranged from 84.4% to 96.3%.

2-Dimensions Convolutions missed temporal features immediately after the convolutions, to expand the interested area of visual representation in spatial and temporal dimensions for getting distinctive appearance and motion features by using 3-Dimensions convolutional network, authors in [62] proposed a Dense 3-Dimensions convolutional network which consisted of a sequence of a 3-dimensions convolution layers and a 3-Dimensions max-pooling layers, four 3-Dimensions dense blocks, four 3-Dimensions transition layers, a fully connected layer and a softmax classification layer. Each 3-Dimensions dense block contained multiple mixed function layers where every layer had a direct connection with each other layers in a feed-forward manner to maximize the flow of the information within the network. The 3-Dimensions dense blocks had a layer of batch normalization, a 3-Dimensions convolution, and a rectified linear unit. The transition layer contained a 3-Dimensions convolution layer and a 3-Dimensions pooling operator. Finally, the loss function was added, it consisted of identification loss and center loss to decrease intra-class variations and increase inter-class variations. This model was trained over the two video datasets: MARS and iLIDS-VID. By comparing this model with state-of-the-art models, In MARS, the results at rank-1 to rank-20 ranged from 76% to 94.1%. In iLIDS-VID, the results at rank-1 to rank-20 ranged from 65.4% to 98.3%. In [63], authors proposed 3D PersonVLAD aggregation model. It was an end-to-end deep 3D convolution model built based on C3D architecture [64]. It contained five layers of convolution and pooling, a layer of 3-D body parts alignment, a layer of spatial-temporal aggregation, and a loss function layer. The 3D part alignment module was adopted, which depended on the attention technique, to handle misalignments and extracted distinctive local features from part maps of 3D convolutions. Online instance matching loss function was adopted to identify the person. For experiments, MARS, iLIDS-VID and PRID2011 were used. For all the datasets, the cumulative match characteristic (CMC) metric was used for evaluating the performance of model. In comparison the proposed model with state-of-the-art models, In MARS, the results at rank-1 to rank-20 ranged from 80.8% to 99%. In iLIDS-VID, the results at rank-1 to rank-20 ranged from 70.7% to 99.2%. In PRID2011, the results at rank-1 to rank-20 ranged from 88% to 99.7%.

One of the robust features in the field of person identification is biometrics. It is divided into two main categories: primary biometrics, such as fingerprint and face, and soft biometrics, such as gait. The first category is not compatible with solving the person Re-ID problem whereas the soft biometrics especially the gait biometrics, gets significant attention in video surveillance systems because they have several advantages. soft biometric features can be extracted from low-resolution videos and the soft biometrics like a gait can be recorded by a camera from a long distance [65]. Also, gait features can be used for both short-term and long-term person Re-ID. One of the earliest studies that used soft biometrics was done by the authors in [66]. They proposed handcrafted method for enhancing person re-identification by integrating gait features with appearance features, also authors in [67] proposed handcrafted method for long-term person Re-ID using true motion features. The motion patterns were extracted by encoding trajectory-aligned descriptors with Fisher vectors in a spatial-aligned pyramid. This work supports long-term person Re-ID since it didn't depend on appearance features. None of the existing long-term person Re-ID works was built based on the deep learning methods and this is one of the future directions in person Re-ID task. Also, since anthropometric features that extracted from RGB-Depth images improved the performance of image-based Re-ID, authors in [68] proposed handcrafted context-aware ensemble fusion system that extracted the anthropometric features with the gait features from KinectTM based dataset and fused between these two types of features, but the accuracy results still need improvements and maybe adopting the deep learning methods into RGB-Depth videos and RGB-Infrared will improve the video-based person Re-ID.

C. Deep Image-to-Video person Re-ID systems

There is a special case between image-based person Re-ID and video-based person Re-ID when the probe image is a static image of a person, but the gallery contains a video sequence of many persons. This case is common in the smart surveillance systems of public places. This is one of challenging task, because there are large differences between different frames inside each video, which will increase the difficulty of matching between image and video. Also, only the appearance features are shared between both image and video and it cannot use the spatial-temporal features directly that it has existed in the video. So, recent studies try to solve this problem and find the best methods of fusion between features that are extracted from image and the features that are extracted from video. In [69], authors introduced a general cross-modality model with temporally memorized similarity learning for image-to-video person Re-ID. It used CNN for extracting appearance features then the output of CNN inserted into LSTM network for extracting the spatial-temporal features of video encoding. Similarity sub-network was used to measure the similarity between the features of image and the features of video. This model was compared with the video-based person Re-ID approaches. It achieved in PRID-2011, 68.5% at rank-1 and 84.7% at rank-5. In iLIDS-VID, it achieved 39.5% at rank-1 and 66.9% at rank-5. In MARS, it achieved 56.5% at rank-1 and 70.6% at rank-5. In [70], the authors proposed a deep architecture called Point-to-Set Network that integrated point-to-set distance metric learning with convolution feature representation learning. A k-Nearest Neighbor triplet (kNN-triplet) module was adopted to make this network focused on the important frames while disregarding the other useless frames in a video. The Point-to-Set Network was evaluated on three new image-to-video person Re-ID datasets: iLIDS-VID-P2S, PRID2011-P2S and MARS-P2S. In iLIDS-VID-P2S, it achieved 40% at rank-1 and 68.54% at rank-5. In PRID2011-P2S, it achieved 73.31% at rank-1 and 90.45% at rank-5. In MARS-P2S, it achieved 55.25% at rank-1 and 72.88% at rank-5. Good results were achieved if the correct fusion between image and video features was done. Image-to-Video person Re-ID systems still have a few studies and are considered to be one of the current person Re-ID issues that require further research.

The summary of these studies is presented in table 3.

V. Person Re-ID Benchmark Datasets

To evaluate the robustness of person Re-ID systems, it is critical to have available person Re-ID datasets with a large volume of data for training deep learning models and with the characteristics of inter-class and intra-class variations. These datasets differ from each other in the number of images, identities, cameras and image types.

A. Image-based dataset

Several datasets for image-based person Re-ID have existed, and the most used datasets are summarized in table1.

VIPeR [71]*:* it consists of 1,264 images for 632 person which they are recorded from 2 non-overlapping camera.

CUHK01 [72]*:* it consists of 3,884 images for 971 person which they are recorded from 2 non-overlapping camera in a college campus.

CUHK03 [26]*:* it is one of the largest image-based person Re-ID datasets. It contains 13,164 images of 1360 person. They are recorded from 5 non-overlapping cameras.

Market-1501 [73]*:* it consists of 32,643 images for 1,501 person that they are recorded from 2 to 6 non-overlapping cameras of the front of a supermarket.

DukeMTMC-ReID [74]*:* it consists of 46,261 images for 1852 persons which they are recorded from 8 non-overlapping camera located at the Duke University campus.
Kinect-REID [75]: It contains sequences of 71 person which they are recorded from the authors' department.
RGBD-ID [76]: it consists of four groups with different views, each group contains the same 80 person. It is created in different days with different appearance variations.
RegDB [77]*:* it consists of 4120 RGB images and 4120 thermal images of 412 persons. They are taken by two kinds of camera.
SYSU-MM01 [52]: it consists 15,792 infrared images and 287,628 RGB images of 491 person. They are recorded from the authors' department. by six cameras, including four RGB cameras and two infrared cameras.

B. Video-based dataset

There are also several datasets that are used for evaluating video-based person Re-ID models. The mostly used datasets are listed in table 2.
PRID2011 [78]: it consists of 24541 images for 934 persons from 600 videos that they are recorded from 2 non-overlapping cameras of a multi-camera network in an airport.
iLIDSVID [79]: it consists of 42495 images for 300 person from 600 videos that they are recorded from 2 non-overlapping cameras in an airport.
MARS [80]: it is the largest video-based person Re-ID dataset. It consists of about 1191003 images for 1261 person from 200 videos that they are recorded from 2 to 6 non-overlapping cameras.
RPIfield [81]: it consists of 601,581 images for 112 person that they are recorded from 2 non-overlapping cameras in an outdoor field on the campus.
DukeMTMC-VideoReID [82]*:* it consists of 369,656 images for 702 identities for training and 445,764 images for 702 identities for testing, there are 408 identities as the distractors recorded from 8 camera.

TABLE I
IMAGE-BASED DATASET

Dataset	Release year	Identities	Camera	Images	Image type
VIPeR	2007	632	2	1,264	RGB
CUHK01	2012	971	2	3,884	RGB
CUHK03	2014	1360	2	13,164	RGB
Market-1501	2015	1,501	6	32,668	RGB
DukeMTM-ReID	2016	1812	8	36,411	RGB
Kinect-REID	2016	71	1 RGB-D	-	RGB-D
RGBD-ID	2012	80	1 RGB-D	-	RGB-D
RegDB	2017	412	1 RGB + I IR	4120 RGB 4120 IR	RGB-IR
SYSU-MM01	2017	491	4 RGB + 2 IR	289,145 RGB 16,579 IR	RGB-IR

TABLE 2
VIDEO-BASED DATASET

Dataset	Release year	Identities	Cameras	Images	Sequence
PRID2011	2011	934	2	24541	Yes
iLIDSVID	2014	300	2	42495	Yes
MARS	2016	1261	6	1191003	Yes
RPIfield	2018	112	12	601581	Yes
DukeMTMC-VideoReID	2019	702/train 702/test	8	369,656/train 445,764/test	Yes

VI. Challenges and Open Issues

At the end of this review, there are several challenges and issues can be observed, and they will direct the incoming studies in the field of person Re-ID.
Long-term person Re-ID: Most existing person Re-ID models support short-term Re-ID in which person always move in a small space for a short period of time and assume the probe image of person and its matching gallery images have the similar appearance such as clothing color, but the most real-world scenario usually need a long-term person Re-ID systems in which people probably appear after various days [83]. The first question can be asked is what is the type of features can be extracted for long-term person Re-ID where the appearance features are not appropriate in this case? Soft biometrics such as anthropometric and gait are suggestions to be used in this task but must consider that there are some types of features such as the anthropometric need a special type of camera which can be captured the depth data [84].

Soft biometrics are used with deep learning in the field of human identification such as in [85] [86] [87] [88] [89]. Future person Re-ID dataset should consider the long-term dataset that recorded over several days [2] [83]. Long-term deep person Re-ID become one of the main research directions.

Person Re-ID datasets: Datasets are very important for validating the new models. None of these datasets is large enough in terms of the number of people, period of time and number of cameras which the most dataset recorded by only two cameras. In particular, Deep learning model needs large-scale data for training task. One of the early realistic and large-scale dataset is Market-1501 (+500k) [73]. So, build new datasets would help quick progress in person Re-ID. Also, there is a gap between different person Re-ID datasets when training model on one dataset and testing on another one, the performance is dropped. To tackle this challenge, the generative adversarial network (GAN) worked as a technique of data augmentation which helps in overcome the weakness of existing person Re-ID datasets. One of the recent work was done by [90] and introduced MSMT-17 large-scale dataset.

Also, authors in [91] [92] using GAN for expanding the person Re-ID samples. This is a promising direction for both image-based dataset and video-based dataset.

Multimodal person Re-ID: Depending only on the appearance features such as clothing that extracted from RGB images are not robust enough, authors try to combine other modalities, like depth data and thermal data where they are robust against many environments variations for improving the accuracy. So, using deep learning on multimodal data is one of new directions in person Re-ID especially with the appearing of cheap RGB-Depth sensors devices and there are only a few studies pay attention to this issue. Also, one of the challenges in the multimodal need to consider in new studies is developing a framework that handles missing features or modalities that mostly occur by occlusions or pose variations or other variations [75].

Unsupervised person Re-ID: Most of current deep learning based person Re-ID systems depend on supervised learning which train labeled data in the same environment. So, training no annotation data in new and real-world environments will led us to high annotation cost since the deep learning models need a huge data for training. To overcome this problem, some recent studies such as in [93][94][95] proposed unsupervised models to extract discriminative features from unlabeled dataset. Unsupervised deep person Re-ID is one of the significant research directions.

Attribute learning: The attributes of a person are a semantic high-level representation like hair, gender, age, etc. which is robust against many environment transformations, (see Figure 7). Several studies adopted these attributes for deep learning based person Re-ID systems such as in [96][97][98] to bridge the gap between the images and high level semantic features. Attribute learning achieved promised results which are make it one of the future directions.

Language based person Re-ID: Searching persons in a large-scale image dataset using natural language description rather than image-based, video-based or attributes-based person Re-ID and retrieve the most similar person images from gallery is useful task in many situations such as when probe image is not found. One of these work was done by [99]. Language based person Re-ID and large-scale person Re-ID description dataset most be considered in new person Re-ID studies.

Automated person Re-ID architecture: Building the architectures of deep learning model manually by a human is consuming time and effort and exposed to mistake. Recently, the neural architecture search (NAS) which is the process of automating architecture engineering [32] is used to tackle this challenge. Currently, the research on NAS is getting more attention. So, adopting NAS for person Re-ID task is one of the important directions that must be considered in future studies since the most NAS methods do not guarantee that the suggested CNN is appropriate for person Re-ID tasks.

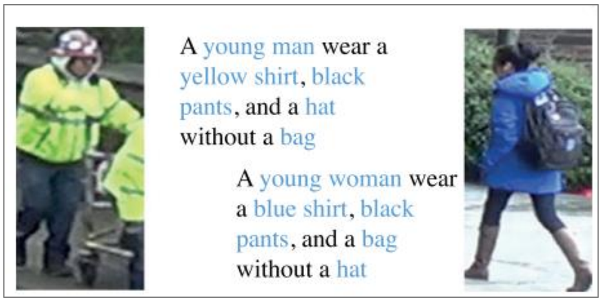
A young man wear a yellow shirt, black pants, and a hat without a bag
A young woman wear a blue shirt, black pants, and a bag without a hat

Figure 7: Examples of describe a person using attributes [98].

Efficiency vs Accuracy: the best accuracy is achieved mostly by large models, but these large models may consume a lot of time and memory size which effects in the efficiency of these models especially when applying into real video surveillance systems. Most existing models did not consider the processing time and memory size for the goal of achieving higher accuracy. The trade-off between ranking accuracy and the processing time is required and should be considered by authors that working in these fields.

VII. CONCLUSION

Person Re-ID is one of the important and critical tasks in intelligent video surveillance systems and it is still a challenging task. This survey discussed deep learning person Re-ID systems. We described a general architecture for traditional systems and deep learning systems. Many recent studies go toward adopted deep learning to avoid the limitations of handcrafted methods. We listed these studies based on three categories: image-based person Re-ID where the probe and gallery are images, video-based person Re-ID where the probe and gallery are video frames, or image-to-video person Re-ID where the probe is an image and the gallery is video frames. Different results have been achieved and have been affected by many factors such as the type of extracted features, the architecture of models (pre-trained model or trained from scratch), loss functions, and so on.

However, despite the improvements that have been made using deep learning methods comparable with handcrafted methods, there are still many issues and limitations that need to be considered in future research directions to improve the person Re-ID systems and achieve promising results. Further research should be done to investigate the long-term person Re-ID whereas the most existing studies built for only short-term. Also, using multi modalities person Re-ID system to support different types of features is important for improving the system performance. Building new person Re-ID datasets is essential for all the improvement directions.

ACKNOWLEDGMENT

This project was funded by Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under grant No. (DG-019-611-1440). The authors, therefore, gratefully acknowledge the DSR technical and financial support.

REFERENCES

[1] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person Re-identification: Past, Present and Future," vol. 14, no. 8, pp. 1–20, Oct. 2016.

[2] D. Wu *et al.*, "Deep learning-based methods for person re-identification: A comprehensive review," *Neurocomputing*, vol. 337, no. xxxx, pp. 354–371, Apr. 2019.

[3] H. B. Zaman, M. H. M. Saad, M. A. Saghafi, and A. Hussain, "Review of person re-identification techniques," *IET Comput. Vis.*, vol. 8, no. 6, pp. 455–474, Dec. 2014.

[4] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, "The re-identification challenge," in *Advances in Computer Vision and Pattern Recognition*, vol. 56, London: Springer London, 2014, pp. 1–20.

[5] K. Wang, H. Wang, M. Liu, X. Xing, and T. Han, "Survey on person re-identification based on deep learning," *CAAI Trans. Intell. Technol.*, vol. 3, no. 4, pp. 219–227, Dec. 2018.

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

[7] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 523–536, Mar. 2019.

[8] I. Bouchrika, "On Using Gait Biometrics for Re-Identification in Automated Visual Surveillance," in *Computer Vision*, no. January, IGI Global, 2018, pp. 2363–2386.

[9] K. Jungling, C. Bodensteiner, and M. Arens, "Person re-identification in multi-camera networks," in *CVPR 2011 WORKSHOPS*, 2011, pp. 55–61.

[10] B. Lavi, M. F. Serj, and I. Ullah, "Survey on Deep Learning Techniques for Person Re-Identification Task," Jul. 2018.

[11] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep Learning for Computer Vision: A Brief Review," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–13, 2018.

[12] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[13] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M. S. Nixon, "Soft Biometrics and Their Application in Person Recognition at a Distance," *IEEE Trans. Inf. Forensics Secur.*, vol. 9, no. 3, pp. 464–475, Mar. 2014.

[14] G. Wang, Y. Fang, J. Wang, and J. Sun, "Extensive Comparison of Visual Features for Person Re-identification," in *Proceedings of the International Conference on Internet Multimedia Computing and Service - ICIMCS'16*, 2016, pp. 192–196.

[15] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, no. Ldml, pp. 2288–2295.

[16] S. Liao, Y. Hu, Xiangyu Zhu, and S. Z. Li, "Person re-identification by Local Maximal Occurrence representation and metric learning," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, vol. 07-12-June, pp. 2197–2206.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, vol. 2016-Decem, pp. 770–778.

[18] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, vol. 07-12-June, no. 8, pp. 1–9.

[19] Y. Jia *et al.*, "Jia - Caffe - Convolutional Method for Fast Feature Embedding," 2014.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[21] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang, "Person Re-identification via Recurrent Feature Aggregation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9910 LNCS, 2016, pp. 701–716.

[22] Sutong Zheng, Xiaoyu Li, Aidong Men, Xiaoqiang Guo, and Bo Yang, "Integration of deep features and hand-crafted features for person re-identification," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2017, vol. II, no. July, pp. 674–679.

[23] D. Wu, S.-J. Zheng, W.-Z. Bao, X.-P. Zhang, C.-A. Yuan, and D.-S. Huang, "A novel deep model with multi-loss and efficient training for person re-identification," *Neurocomputing*, vol. 324, pp. 69–75, Jan. 2019.

[24] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun, "A Guide to Convolutional Neural Networks for Computer Vision," *Synth. Lect. Comput. Vis.*, vol. 8, no. 1, pp. 1–207, Feb. 2018.

[25] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, vol. 07-12-June, pp. 815–823.

[26] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep Filter Pairing Neural Network for Person Re-identification," in *2014 IEEE Conference on*

Computer Vision and Pattern Recognition, 2014, pp. 152–159.

[27] E. Ahmed, M. Jones, and T. K. Marks, “An improved deep learning architecture for person re-identification,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3908–3916.

[28] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1335–1344.

[29] Y. Huang, H. Sheng, Y. Zheng, and Z. Xiong, “DeepDiff: Learning deep difference features on human body parts for person re-identification,” *Neurocomputing*, vol. 241, pp. 191–203, Jun. 2017.

[30] H. Zhao *et al.*, “Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 1, pp. 907–915.

[31] D. Li, X. Chen, Z. Zhang, and K. Huang, “Learning Deep Context-Aware Features over Body and Latent Parts for Person Re-identification,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7398–7407.

[32] T. Elsken, J. H. Metzen, and F. Hutter, “Correction to: Neural Architecture Search,” in *Automated Machine Learning: Methods, Systems, Challenges*, F. Hutter, L. Kotthoff, and J. Vanschoren, Eds. Cham: Springer International Publishing, 2019, pp. C1--C1.

[33] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, “Auto-ReID: Searching for a Part-aware ConvNet for Person Re-Identification,” 2019.

[34] A. Hermans, L. Beyer, and B. Leibe, “In Defense of the Triplet Loss for Person Re-Identification,” Mar. 2017.

[35] Z. He, C. Jung, Q. Fu, and Z. Zhang, “Deep feature embedding learning for person re-identification based on lifted structured loss,” *Multimed. Tools Appl.*, vol. 78, no. 5, pp. 5863–5880, Mar. 2019.

[36] L. Wu, Y. Wang, J. Gao, and X. Li, “Deep adaptive feature embedding with local sample distributions for person re-identification,” *Pattern Recognit.*, vol. 73, pp. 275–288, 2018.

[37] A. Y. Ng, H. Lee, R. Grosse, and R. Ranganath, “Unsupervised learning of hierarchical representations with convolutional deep belief networks,” *Commun. ACM*, vol. 54, no. 10, pp. 95–103, 2011.

[38] W. Chen, X. Chen, J. Zhang, and K. Huang, “Beyond Triplet Loss: A Deep Quadruplet Network for Person Re-identification,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1320–1329.

[39] Z. Zheng, L. Zheng, and Y. Yang, “A Discriminatively Learned CNN Embedding for Person Reidentification,” *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 14, no. 1, pp. 1–20, Dec. 2017.

[40] A. Khatun, S. Denman, S. Sridharan, and C. Fookes, “A Deep Four-Stream Siamese Convolutional Neural Network with Joint Verification and Identification Loss for Person Re-Detection,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, vol. 2018-Janua, pp. 1292–1301.

[41] L. Wu, Y. Wang, X. Li, and J. Gao, “What-and-where to match: Deep spatially multiplicative integration networks for person re-identification,” *Pattern Recognit.*, vol. 76, pp. 727–738, Apr. 2018.

[42] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the Devil in the Details: Delving Deep into Convolutional Nets,” *BMVC 2014 - Proc. Br. Mach. Vis. Conf. 2014*, pp. 1–11, May 2014.

[43] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” pp. 1–14, Sep. 2014.

[44] L. Wu, Y. Wang, X. Li, and J. Gao, “Deep Attention-Based Spatially Recursive Networks for Fine-Grained Visual Recognition,” *IEEE Trans. Cybern.*, vol. 49, no. 5, pp. 1791–1802, May 2019.

[45] L. Wu, R. Hong, Y. Wang, and M. Wang, “Cross-Entropy Adversarial View Adaptation for Person Re-identification,” pp. 1–12, 2019.

[46] X. Zhu, J. Liu, H. Xie, and Z.-J. Zha, “Adaptive Alignment Network for Person Re-identification,” in *MultiMedia Modeling*, vol. 8936, X. He, S. Luo, D. Tao, C. Xu, J. Yang, and M. A. Hasan, Eds. Cham: Springer International Publishing, 2019, pp. 16–27.

[47] A. Wu, W.-S. Zheng, and J.-H. Lai, “Robust Depth-Based Person Re-Identification,” *IEEE Trans. Image Process.*, vol. 26, no. 6, pp. 2588–2603, Jun. 2017.

[48] H. Liu, L. Hu, and L. Ma, “Online RGB-D person re-identification based on metric model update,” *CAAI Trans. Intell. Technol.*, vol. 2, no. 1, pp. 48–55, Mar. 2017.

[49] Z. Imani and H. Soltanizadeh, “Histogram of the node strength and histogram of the edge weight: two new features for RGB-D person re-identification,” *Sci. China Inf. Sci.*, vol. 61, no. 9, p. 092108, Sep. 2018.

[50] L. Ren, J. Lu, J. Feng, and J. Zhou, “Multi-modal uniform deep learning for RGB-D person re-identification,” *Pattern Recognit.*, vol. 72, pp. 446–457, Dec. 2017.

[51] A. Mogelmose, C. Bahnsen, T. B. Moeslund, A. Clapes, and S. Escalera, “Tri-modal Person Re-identification with RGB, Depth and Thermal Features,” in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 301–307.

[52] A. Wu, W. Zheng, H. Yu, S. Gong, and J. Lai, “RGB-Infrared Cross-Modality Person Re-identification,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5390–5399.

[53] M. Ye, Z. Wang, X. Lan, and P. C. Yuen, “Visible Thermal Person Re-Identification via Dual-Constrained Top-Ranking,” in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, pp. 1092–1099.

[54] N. McLaughlin, J. M. del Rincon, and P. Miller, “Recurrent Convolutional Network for Video-Based Person Re-identification,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1325–1334.

[55] L. Chen, H. Yang, J. Zhu, Q. Zhou, S. Wu, and Z. Gao, “Deep Spatial-Temporal Fusion Network for Video-Based Person Re-identification,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, vol. 2017-July, pp. 1478–1485.

[56] H. Liu *et al.*, “Video-Based Person Re-Identification With Accumulative Motion Context,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2788–2802, Oct. 2018.

[57] Z. Zeng, Z. Li, D. Cheng, H. Zhang, K. Zhan, and Y. Yang, “Two-Stream Multirate Recurrent Neural Network for Video-Based Pedestrian Reidentification,” *IEEE Trans. Ind. Informatics*, vol. 14, no. 7, pp. 3179–3186, Jul. 2018.

[58] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6776–6785.

[59] L. Wu, Y. Wang, J. Gao, and X. Li, “Where-and-When to Look: Deep Siamese Attention Networks for Video-Based Person Re-Identification,” *IEEE Trans. Multimed.*, vol. 21, no. 6, pp. 1412–1424, Jun. 2019.

[60] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, “Video Person Re-Identification by Temporal Residual Learning,” *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1366–1377, Mar. 2019.

[61] Y. Liu, Z. Yuan, W. Zhou, and H. Li, “Spatial and Temporal Mutual Promotion for Video-Based Person Re-Identification,” *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 8786–8793, 2019.

[62] J. Liu, Z. Zha, X. Chen, Z. Wang, and Y. Zhang, “Dense 3D-Convolutional Neural Network for Person Re-Identification in Videos,” *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 15, no. 1s, pp. 1–19, Jan. 2019.

[63] L. Wu, Y. Wang, L. Shao, and M. Wang, “3-D PersonVLAD: Learning Deep Global Representations for Video-Based Person Reidentification,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. PP, pp. 1–13, 2019.

[64] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning Spatiotemporal Features with 3D Convolutional Networks,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, vol. 2015 Inter, pp. 4489–4497.

[65] I. Bouchrika, “A Survey of Using Biometrics for Smart Visual Surveillance: Gait Recognition,” in *Surveillance in Action*, 2018, pp. 3–23.

[66] Z. Liu, Z. Zhang, Q. Wu, and Y. Wang, “Enhancing person re-identification by integrating gait biometric,” *Neurocomputing*, vol. 168, pp. 1144–1156, Nov. 2015.

[67] P. Zhang, Q. Wu, J. Xu, and J. Zhang, “Long-Term Person Re-identification Using True Motion from Videos,” in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 494–502.

[68] A. Nambiar, A. Bernardino, J. C. Nascimento, and A. Fred, “Context-Aware Person Re-Identification in the Wild Via Fusion of Gait and Anthropometric Features,” in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 2017, pp. 973–980.

[69] D. Zhang, W. Wu, H. Cheng, R. Zhang, Z. Dong, and Z. Cai, “Image-to-Video Person Re-Identification With Temporally Memorized Similarity Learning,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2622–2632, Oct. 2018.

[70] G. Wang, J. Lai, and X. Xie, “P2SNet: Can an Image Match a Video for Person Re-Identification in an End-to-End Way?,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2777–2787, Oct. 2018.

[71] D. Gray, S. Brennan, and H. Tao, “Evaluating appearance models for recognition, reacquisition, and tracking,” *10th Int. Work. Perform. Eval. Track. Surveill. (PETS)*, vol. 3, pp. 41–47, 2007.

[72] W. Li, R. Zhao, and X. Wang, “Human Reidentification with Transferred Metric Learning,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7724 LNCS, no. PART 1, 2013, pp. 31–44.

[73] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable Person Re-identification: A Benchmark,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, vol. 2015 Inter, no. December 2015, pp. 1116–1124.

[74] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance Measures and a Data Set for Multi-target, Multi-camera Tracking,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9914 LNCS, no. c, 2016, pp. 17–35.

[75] F. Pala, R. Satta, G. Fumera, and F. Roli, "Multimodal Person Reidentification Using RGB-D Cameras," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 4, pp. 788–799, Apr. 2016.

[76] I. B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with RGB-D Sensors," in *Computer Vision -- ECCV 2012. Workshops and Demonstrations*, A. Fusiello, V. Murino, and R. Cucchiara, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 433–442.

[77] D. Nguyen, H. Hong, K. Kim, and K. Park, "Person Recognition System Based on a Combination of Body Images from Visible Light and Thermal Cameras," *Sensors*, vol. 17, no. 3, p. 605, Mar. 2017.

[78] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person Re-identification by Descriptive and Discriminative Classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6688 LNCS, 2011, pp. 91–102.

[79] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person Re-identification by Video Ranking," in *Computer Vision -- ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 688–703.

[80] L. Zheng *et al.*, "MARS: A Video Benchmark for Large-Scale Person Re-Identification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9910 LNCS, 2016, pp. 868–884.

[81] M. Zheng, S. Karanam, and R. J. Radke, "RPIfield: A New Dataset for Temporally Evaluating Person Re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, no. 2013, pp. 1974–19742.

[82] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive Learning for Person Re-Identification with One Example," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2872–2881, 2019.

[83] Q. Leng, M. Ye, and Q. Tian, "A Survey of Open-World Person Re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. PP, no. c, pp. 1–1, 2019.

[84] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," *Comput. Vis. Image Underst.*, vol. 171, no. October 2017, pp. 118–139, Jun. 2018.

[85] F. M. Castro, M. J. Marín-Jiménez, N. Guil, and N. Pérez de la Blanca, "Automatic Learning of Gait Signatures for People Identification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10306 LNCS, 2017, pp. 257–270.

[86] W. Liu, C. Zhang, H. Ma, and S. Li, "Learning Efficient Spatial-Temporal Gait Features with Deep Learning for Human Identification," *Neuroinformatics*, vol. 16, no. 3–4, pp. 457–471, Oct. 2018.

[87] S. Arseev, A. Konushin, and V. Liutov, "Human Recognition by Appearance and Gait," *Program. Comput. Softw.*, vol. 44, no. 4, pp. 258–265, Jul. 2018.

[88] M. Babaee, L. Li, and G. Rigoll, "Person identification from partial gait cycle using fully convolutional neural networks," *Neurocomputing*, vol. 338, pp. 116–125, Apr. 2019.

[89] L. Yao, W. Kusakunniran, Q. Wu, J. Zhang, and Z. Tang, "Robust CNN-based Gait Verification and Identification using Skeleton Gait Energy Image," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*, 2018, pp. 1–7.

[90] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person Transfer GAN to Bridge Domain Gap for Person Re-identification," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88.

[91] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled Samples Generated by GAN Improve the Person Re-identification Baseline in Vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-Octob, pp. 3774–3782.

[92] X. Qian *et al.*, "Pose-Normalized Image Generation for Person Re-identification," in *Computer Vision -- ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 661–678.

[93] H. Fan, L. Zheng, C. Yan, and Y. Yang, "Unsupervised Person Re-identification: Clustering and Fine-tuning," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 14, no. 4, pp. 1–18, Oct. 2018.

[94] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A Bottom-Up Clustering Approach to Unsupervised Person Re-Identification," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 8738–8745, 2019.

[95] Y. Ding, H. Fan, M. Xu, and Y. Yang, "Adaptive Exploration for Unsupervised Person Re-Identification," Jul. 2019.

[96] A. Schumann and R. Stiefelhagen, "Person Re-identification by Deep Learning Attribute-Complementary Information," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, vol. 2017-July, pp. 1435–1443, 2017.

[97] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Multi-type attributes driven multi-camera person re-identification," *Pattern Recognit.*, vol. 75, pp. 77–89, Mar. 2018.

[98] Y. Lin *et al.*, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, 2019.

[99] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang,

"Person search with natural language description," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 5187–5196, 2017.

[100] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, vol. 2016-Decem, pp. 1288–1296.

[101] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning Deep Feature Representations with Domain Guided Dropout for Person Re-identification," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1249–1258.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2957336, IEEE Access

TABLE 3
SUMMARY OF STATE-OF-THE-ARTS STUDIES

Category	Reference	Methods Name	Suggestion for improve the Person Re-ID performance	Input Type	Features Type	Deep Learning Model			Outcomes	
						(1) Model Type	(2) Features Extraction Method	(3) Loss function	(1) Dataset: Accuracy (Rank-1%)	(2) Period Time
Image- based person Re-ID	Li *et al.*, 2014 [26]	Deep Filter Pairing Neural Network (DeepReID)	First work used deep learning models rather than handcrafted models to improve the performance of person Re-ID. It can handle misalignment, photometric and geometric transformations, occlusions and background noise under one integrated network.	RGB	Appearance	Trained from scratch	CNN with six layers: convolutional and max pooling layer, patch matching layer, maxout-grouping layer, additional convolution and max-pooling layer, fully connected layer and softmax layer	Softmax loss	CUHK03 (labeled): 20.65% CUHK03 (detected): 19.89%	Short
	Ahmed, Jones and Marks, 2015 [27]	Improved deep learning architecture	Using tied convolution layers help in captured local relations between the two images by comparing the features from one input image with those extracted from neighboring locations of the other image and then summarize these differences as high-level summary	RGB	Appearance	Trained from scratch	CNN contains two layers of tied convolution with max pooling layer, cross-input neighborhood differences layer, patch summary features layer, across-patch features layer, higher-order relationships layer, and a softmax function	Softmax loss	CUHK03 (labeled): 54.74% CUHK03(detected): 44.96% CUHK01(486 ids): 40.5% CUHK01(100 ids): 65% VIPeR: 34.81%	Short
	Cheng *et al.*, 2016 [28]	Multi-channel parts-based model	Using multi-channel CNN model to extract the full body features (global features) and body-parts features (local features) and combined them into one feature vector.	RGB	Appearance	Trained from scratch	CNN with a single network that contains multiple channels: one global convolution layer, 1 full-body convolution layer, 4 body-parts convolution layers, 5 channel-wise full connection layers, and 1 network-wise full connection layer.	Triplet loss	I-LIDS: 60.4% PRID2011: 22% VIPeR: 47.8% CUHK01: 53.7%	Short

	Huang *et al.*, 2017 [29]	Part-based DeepDiff model	Dividing the body into many parts using pyramid partition architecture and extracted the distinctive deep features from these parts can handle the type of intra-class variations	RGB	Appearance	Trained from scratch	CNN consists of three subnets; each one has a number of convolution layer followed by max pooling and a fully connected layer, Finally, a fully connected layer with 2 softmax units was added.	Softmax loss	CUHK01: 47.9% CUHK03 (labeled): 62.4% CUHK03 (detected): 54.8% VIPeR: 43.2 %	Short
	Zhao *et al.*, 2017 [30]	Body region guided Spindle Net model	Extracting one global features vector of full image and 7 sub regions features vector matching 7 suggested body sub regions, then using tree-structured feature fusion instead of directly combining features together. This can represent a lot of details, then identify person with slight differences.	RGB	Appearance	Trained from scratch	CNN which contains three convolution stages and two ROI pooling stages	Softmax loss	CUHK03: 88.5% CUHK01: 79.9% PRID: 67.0% VIPeR: 53.8% 3DPES: 62.1% LIDS: 66.3% MARKET1501: 76.9% SENSEREID: 34.6%	Short
	Li *et al.*, 2017 [31]	Multi-scale context-aware network for learning the features over body and latent parts	Using multi-scale CNN help to extract robust features of complete body and body parts rather than single-scale CNN, and instead of using rigid parts, prior constraints were used to learn and localize latent human parts since rigid predefined grids were not robust enough for effective part-based feature learning	RGB	Appearance	Trained from scratch	CNN consists of initial convolution layer to capture global features, four multi-scale convolution layers for part-based features, dilated convolution.	Softmax loss	MARKET1501: 80.31% CUHK03 (detected): 67.99% CUHK03 (labeled): 74.21% MARS: 71.77%	Short
	Cheng *et al.*, 2016 [33]	Auto-ReID: a part-aware module	The first work for automated Neural Architecture Search (NAS) for Re-ID task.	RGB	Appearance	Trained from scratch	Macro structure of ResNet	Retrieval loss (Cross-entropy + Triple loss)	MARKET1501: 94.5% CUHK03 (detected): 73.3% % CUHK03 (labeled): 77.9% % MSMT17:78.2%	Short

	Hermans, Beyer and Leibe, 2017 [34]	Variant triplet losses	Using plain CNN with triplet loss can improve the Re-ID results both with a pre-trained CNN and with a model trained from scratch.	RGB	Appearance	Pretrained model / Trained from scratch model	CNN: Resnet-50 architecture and New network called LUNET (it is as Resnet-v2, but it used nonlinear leaky ReLU, multiple 3×3 max pooling with stride 2 and omits the final average pooling of feature maps)	Triple loss	Resnet -50 / LUNET: MARS: 79.80% / 75.56% MARKET1501: 84.92% / 81.83% For Resnet -50: CHUK03 (labeled): 89.63% CHUK03(detected): 87.58%	Short
	He *et al.*, 2019 [35]	Lifted structured loss	Using lifted structured loss rather than triple loss makes complete use of the batch, this minimizes the influence of sample distribution on training and minimizes the time consuming	RGB	Appearance	Trained from scratch	CNN: One branch of CNN has nine convolution layers, four max pooling layers, two fully connected layers, and a softmax layer for classification	Lifted structured loss + identification loss	MARKET1501: 84.53% CUHK03 (labeled): 81.6% CUHK03(detected): 79.9% CUHK01: 70.2% VIPeR: 47.3%	Short
	Wu *et al.*, 2018 [36]	Deep adaptive feature embedding model with local sample distributions	Adapting the distance metric into local range and finding the appropriate positive samples help to get a robust deep embedding model that improving the person Re-ID in the situation of large intra-class variations	RGB	Appearance	Trained from scratch	Convolutional Restricted Boltzmann Machines (CRBMs)	extension of triplet loss	MARKET1501: 68.32% VIPeR: 49.04% CUHK03: 73.02% CUHK01: 71.60%	Short
	W. Chen *et al.*, 2017 [38]	Deep quadruplet network	Extending the triple loss to quadrable loss, help in get an improved generalization ability, so the model can achieve a higher person Re-ID performance	RGB	Appearance	Pretrained model	CNN: AlexNet	Quadruplet loss + metric learning as in [100]	CUHK03: 75.53% CUHK01(486 persons): 62.55% CUHK01(100 person): 81% VIPeR: 49.05%	Short

	Z. Zheng, Zheng, and Yang 2017b [39]	Discriminatively Learned CNN Embedding and a similarity metric	Integrating the identification loss and verification loss to overcomes the weakness of the typical triplet loss.	RGB	Appearance	Pretrained model	CNN: ResNet-50	Cross-entropy loss + Square layer	MARKET1501(single query): 79.51% MARKET1501(multi query): 85.84% CUHK03(single shot): 83.4% CUHK03(multi shot): 88.3% Market1501+500k: 68.26% Oxford5k (CaffeNet + VGG16): 66.2%	Short
	Khatun *et al.*, 2018 [40]	Four stream Siamese CNN with joint verification and identification loss	To overcomes the weakness of the typical triplet loss, four stream networks are used with consider the person Re-ID as verification and identification task	RGB	Appearance	Pretrained model	CNN: AlexNet	Quarter loss	VIPeR: 68.7% CUHK03: 85.5% CUHK01: 83.95% PRID2011: 75%	Short
	L. Wu, Wang, Li, et al. 2018 [41]	Deep Spatially Multiplicative Integration Networks	Integrating spatial relationship into feature learning can improve person Re-ID performance.	RGB	Appearance	Pretrained model	CNN: M-Net and D-Net	Cosine function + binomial deviance loss function	MARKET1501:67.15% CUHK03: 73.23% VIPeR: 49.11%	Short
	L. Wu, Yang Wang, Li and Gao, 2019 [44]	Deep visual attention model with efficient spatially recursive encoding structure	considering the spatial relationship of a fine-grained object with a recursive structure can improve person Re-ID performance	RGB	Appearance	Pretrained model	CNN: M-Net and D-Net	Cross-Entropy loss	MARKET1501:64.23%. VIPeR: 65.11% CUHK03: 65.23%	Short

	L. Wu, Hong, Wang, et al. 2019 [45]	Cross-Entropy Adversarial View Adaptation framework	Using an adversarial view adaptation approach to improve the person Re-ID system.	RGB	Appearance	Pretrained model	CNN: M-Net and D-Net	Cross-Entropy loss	VIPeR: 55.9% CUHK03: 88.9% MARKET1501: 89.1% DukeMTMC- reID: 80.1%	Short
	Zhu *et al.*, 2019 [46]	Adaptive Alignment Network	Considering the pedestrian misalignment challenge when built the model can help in improving the person Re-ID	RGB	Appearance	Pretrained model	CNN: ResNet-50, in addition: patch alignment module which contains two convolution layers followed by two fully connected layers, and pixel alignment module: consists of two convolution layers followed by two fully connected layers	Cross-Entropy loss	MSMT17: 70.5% MARKET1501: 92.0% DUKEMTMC-REID: 84.1%	Short
	Ren *et al.*, 2017 [50]	Multi-modal uniform deep learning model for RGB-D person Re-ID	This is the first RGB-D work for handling person Re-ID problem, which Extracting anthropometric features from depth images and appearance features from RGB images can improve person Re-ID task	RGB-D	Appearance + anthropometric	Pretrained model	CNN in [101] which contains four convolution layers followed by a pooling layer, a series of 6 inception units, and fully connected layer.	Hing loss	KINECT-REID: 97.0% RGBD-ID: 76.7%	Short
	Wu *et al.*, 2017 [52]	RGB-IR cross-modality person Re-ID problem model	This is the first RGB-IR cross-modality work for addressing person Re-ID problem which considering both the two modalities: RGB and infrared images and extracting features from them to improve the person Re-ID system	RGB-IR	Appearance	Pretrained model	CNN: ResNet-6	Softmax loss	SYSU-MM01 (single-shot): 14.80% SYSU-MM01 (multi-shot): 14.80%	Short

	Ye *et al.*, 2018 [53]	End-to-end dual-path network with a novel bi-directional dual-constrained top-ranking loss	Extracting the features from visible and thermal images can improve the person Re-ID task by assuming the low-level visual patterns such as texture and corner of thermal images are similar to general visible images	RGB-IR	Appearance	Pretrained model	CNN: AlexNet	Softmax + ranking loss contains cross-modality and intra-modality constraints	REGDB: 33.47 % SYSU-MM01 (single-shot): 17.01%	Short
Video-based person Re-ID	McLaughlin, Rincon and Miller, 2016 [54]	Recurrent convolutional network	This is the first video-based person Re-ID that extracts features from video frames, a rich temporal information can be captured and improve the person re-id task.	RGB	Spatial-Temporal Features (Appearance)	-	CNN + RNN *(no details about structure)*	Cross-entropy loss or softmax function, and Siamese	ILIDS-VID: 58% PRID-2011: 70%	Short
	L. Chen *et al.*, 2017 [55]	Deep end-to-end spatial and temporal fusion network	Overcoming the problem of missing important local features, full-body and part-body features for person from video frames are captured to get more accurate spatial-temporal features and improve person Re-ID accuracy	RGB	Spatial-Temporal Features (Appearance)	Trained from scratch	CNN + RNN CNN: three repetitive convolution layers, max-pooling layers and ReLU activation layers	Siamese loss + softmax loss layer	PRID-2011: 77% ILIDS-VID: 61% MARS: 71%	Short
	H. Liu *et al.*, 2018 [56]	End-to-end accumulative motion context network	Using two stream convolution architecture to extract appearance and motion features, and adding two spatial networks in each of two streams to find the spatial features from raw video frames	RGB	Spatial-Temporal Features (Appearance)	Trained from scratch	CNN+RNN CNN Spatial: three convolution layers and three max pooling layers with a non-linear layer (tanh), a fully connected layer at the top of last max pooling layer. CNN Motion: six convolution layers, a non-linear layer (tanh), several deconvolutional layers.	Contrastive loss + softmax loss	PRID-2011: 83.7% ILIDS-VID: 68.7% MARS: 68.3%	Short

	Zeng *et al.*, 2018 [57]	Two stream multi-rate recurrent neural networks	Processing the motion speed variance when extracting spatial-temporal features can improve the Re-ID task	RGB	Spatial-Temporal Features (Appearance)	Pretrained model/ Trained from scratch	CNN: VGG and CNN_M RNN: MR GRUS	-	PRID-2011: 78.7% ILIDS-VID: 59.4%	Short
	Zhou *et al.*, 2017 [58]	Joint spatial and temporal recurrent neural networks	Select only the more informative frames from videos for extracting spatial-temporal features can improve the Re-ID task	RGB	Spatial-Temporal Features (Appearance)	Pretrained model	CNN: CaffeNet RNN: LSTM	Triplet loss	PRID-2011: 79.4% ILIDS-VID: 55.2% MARS: 70.6%	Short
	L. Wu, Yang Wang, Gao and Li, 2019 [59]	Deep Siamese attention networks for improving the spatial representation	Improving the local spatial features and make them more discriminative and focusing on more related features improve the Re-ID task	RGB	Spatial-Temporal Features (Appearance)	Pretrained model	CNN: GoogLeNet RNN: GRU	Cross-entropy loss	PRID-2011: 77% ILIDS-VID: 61.9% MARS: 73.5%	Short
	Dai *et al.*, 2019 [60]	Temporal residual learning	Processing the poor spatial alignment and describing person in different aspects to fully leverage temporal information can improve the Re-ID task	RGB	Spatial-Temporal Features (Appearance)	Pretrained model	CNN: GoogLeNet RNN: Bi LSTM	Softmax + cross-entropy loss	PRID-2011: 87.8% ILIDS-VID: 57.7% SDU-VID: 97.7% MARS: 80.5%	Short
	Y. Liu *et al.*, 2019 [61]	Spatial and temporal mutual promotion model	Using temporal information to recover the spatial information in one frame by area of the same location in other frames	RGB	Spatial-Temporal Features (Appearance)	Pretrained model	CNN: Inception- v3 RNN: RRU	Cross-entropy loss + batch hard triplet	PRID-2011: 92.7% ILIDS-VID: 84.3% MARS: 84.4%	Short

	J. Liu *et al.*, 2019 [62]	Dense 3-D convolutional network	Using 3D Convolutional Network can expand the visual representation in spatial and temporal dimensions for getting discriminative appearance and motion features	RGB	Spatial-Temporal Features (Appearance)	Trained from scratch	CNN: a 3D convolutional layer and a 3D max pooling layer, four dense blocks, four 3D transition layers, a fully connected layer and a softmax classifier layer	Identification loss + center loss	MARS: 76% ILIDS-VID: 65.4%	Short
	L. Wu, Yang Wang, Shao and Wang, 2019 [63]	3-D PersonVLAD aggregation model	Using 3D Convolutional Network with 3D part alignment module that extracts features over 3D human body parts can improve the person Re-Id accuracy	RGB	Spatial-Temporal Features (Appearance)	Trained from scratch	CNN: five convolution layers and five pooling layers, the 3D body part alignment layer, the spatial-temporal aggregation layer, and a loss function layer	Online instance matching (OIM) loss	MARS: 80.8% ILIDS-VID: 70.7% PRID2011: 88%	Short