


Article

Instance Hard Triplet Loss for In-video Person Re-identification

Xing Fan ¹, Wei Jiang ^{1,*} , Hao Luo ¹, Weijie Mao ¹ and Hongyan Yu ²

¹ The State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China; xfanplus@zju.edu.cn (X.F.); haoluocsc@zju.edu.cn (H.L.); wjmao@zju.edu.cn (W.M.)

² Beijing Electro-mechanical Engineering Institute, Beijing 100074, China; yuhongyan09@163.com

* Correspondence: jiangwei_zju@zju.edu.cn

Received: 7 March 2020; Accepted: 20 March 2020; Published: 24 March 2020



Abstract: Traditional Person Re-identification (ReID) methods mainly focus on cross-camera scenarios, while identifying a person in the same video/camera from adjacent subsequent frames is also an important question, for example, in human tracking and pose tracking. We try to address this unexplored in-video ReID problem with a new large-scale video-based ReID dataset called PoseTrack-ReID with full images available and a new network structure called ReID-Head, which can extract multi-person features efficiently in real time and can be integrated with both one-stage and two-stage human or pose detectors. A new loss function is also required to solve this new in-video problem. Hence, a triplet-based loss function with an online hard example mining designed to distinguish persons in the same video/group is proposed, called instance hard triplet loss, which can be applied in both cross-camera ReID and in-video ReID. Compared with the widely-used batch hard triplet loss, our proposed loss achieves competitive performance and saves more than 30% of the training time. We also propose an automatic reciprocal identity association method, so we can train our model in an unsupervised way, which further extends the potential applications of in-video ReID. The PoseTrack-ReID dataset and code will be publicly released.

Keywords: person ReID; video; triplet loss; pose; unsupervised learning

1. Introduction

Given a query person image, person re-identification aims to identify persons with the same Identity (ID) in the gallery images. Most existing methods focus on the problem that query images and gallery images are from different camera views, i.e., a cross-camera problem. With the rising of deep learning, the ReID community has witnessed a huge jump of accuracy in recent years. For example, on the Market-1501 [1] dataset, a widely-used ReID dataset, a part-based model [2] has achieved a 93.8% rank-1 accuracy.

With the progress of ReID, it is natural to apply ReID in other areas. Some researchers tried to incorporate ReID with human tracking [3–5]. They utilized extra ReID datasets to train a ReID model and used the obtained model for feature extraction. Those extracted ReID features were then utilized to identify the tracking target from candidate persons, achieving a better performance. However, directly using a model, trained on cross-camera ReID datasets such as the Market-1501 [1] dataset and CUHK03 [6] dataset, usually obtains an inferior performance due to the cross-domain bias that the appearance in the source dataset is often much different from the appearance in the target dataset.

A possible solution is to collect and annotate ReID data of the target domain, like DukeMTMC-reID [7], which comes from the DukeMTMC [8] dataset for Multi-Target, Multi-Camera Tracking. Then, we can use the obtained ReID data to train a ReID model and improve the tracking performance. However, the collection and annotation are often expensive and sometimes

impossible to obtain. Furthermore, even if this cross-camera ReID dataset is available, we still cannot directly use those cross-camera ReID data for person tracking within the same video sequence, because person tracking within the same video needs to identify persons within the same video, which is different from traditional cross-camera ReID that identifies persons across multiple cameras. In this paper, we refer to person ReID within the same video as in-video ReID. As shown in Figure 1, the traditional cross-camera ReID task searches for the same person from images captured by different cameras at different times, while the in-video ReID task searches for the same person from subsequent frames of the same video.

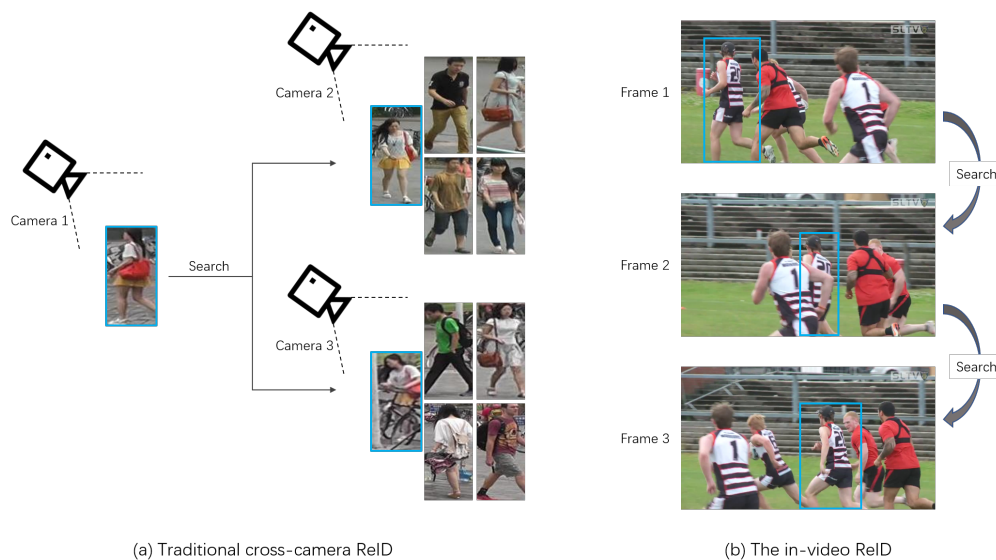


Figure 1. Illustration of (a) traditional cross-camera ReID and (b) the in-video ReID addressed in this paper. Cross-camera ReID retrieves images of the same person from different cameras, while in-video ReID finds the same person in the subsequent frames of the same video. The blue bounding boxes in each sub-figure indicate the same person.

In this paper, we try to address the aforementioned in-video ReID problem that identifies a person in the same video sequence. We argue that traditional cross-camera ReID and in-video ReID are dissimilar in the following aspects:

- (1) **Single-camera view:** Cross-camera ReID needs to identify persons appearing in multiple cameras; thus, it forces the network to extract features consistent in all camera views and drop the camera-specific features, leading to limited available features, while in-video ReID can fully utilize all features.
- (2) **Short-term:** Cross-camera ReID tries to construct a long-term association that inevitably discards transient clues. On the contrary, we will demonstrate that those transient clues are very critical for in-video ReID.
- (3) **Background:** As we mentioned above, temporary clues like the background always act as a distractor, and those features are often discarded in cross-camera ReID. For instance, MGCAM [9] learns to predict a foreground mask of the human body area to suppress background distraction. However, for in-video ReID, the background of a person usually does not change dramatically within a few subsequent frames and can help to distinguish people with a similar appearance.
- (4) **Pose:** The pose is also regarded as a distraction, and pose-unrelated features are preferred in cross-camera ReID. For example, the recent FD-GAN [10] use Generative Adversarial Nets (GAN) [11] to learn pose-unrelated representations.

Similar looking persons often exist in the same video frame, engaged in the same activity, as shown in Figure 1b. We need to fuse short-term information like background and pose with human body features to discriminate highly similar persons for in-video ReID.

Since in-video ReID is quite different from existing cross-camera ReID, why bother to investigate this new task? As mentioned in [12], in a video, people often disappear and re-appear again due to occlusions by other people or objects. A tracking method without knowing identity information cannot handle this situation properly, because it cannot distinguish well between a new person and a re-appeared existing person. That is the reason why [5] argued that the ReID module is essential for multiple person tracking. Some researchers in the pose tracking community also plan to embed a ReID model in their method [13]. Thus, a good in-video method is helpful for person tracking and pose tracking, as well as many downstream practical applications such as video surveillance and sports video analysis.

In this paper, we try to address this new in-video ReID problem, and our main contributions can be summarized in three-folds:

- (1) A new large-scale video-based in-video ReID dataset with full images available. To the best of our knowledge, no such in-video dataset has been released before. A ReID-Head network is also designed to extract features for the in-video ReID task efficiently;
- (2) A new loss function called Instance Hard Triplet (IHT) loss is proposed, which is suitable for both the cross-camera ReID task and the in-video ReID task. Compared with the widely-used Batch Hard Triplet (BHT) loss [14] for cross-camera ReID, it achieves competitive performance and saves more than 30% of the training time (see Section 4.4, Table 2);
- (3) Labeling ReID data is expensive and time consuming; thus, we also propose an unsupervised method for automatically associating persons in the same video with the same identity through reciprocal matching, so that an in-video ReID model can be trained using these associated data.

2. Related Works

Person ReID is a popular topic in the computer vision area. Benefiting from the advances of deep neural networks, it has achieved great progress in recent years. The current person ReID studies can roughly be divided into two categories: representation learning based methods [15–18] and metric learning based methods [19–24]. In this section, we will introduce these two categories, respectively, and then introduce their applications in the tracking area.

2.1. Representation Learning Based

Representation learning based methods mainly focus on the form of input and the structure of the network for learning a better representation for the input information. Global features extracted directly by the convolutional backbone are often used [15–17,19]. Apart from learning global features, many part-based methods, such as Part-based Convolutional Baseline (PCB) [2], AlignedReID [25], and Spindle Net [26], learn discriminative local features for person re-identification. GLAD [27] combines both global and local features to further improve the performance. For video-based ReID, Li et al. [28] also combined the local short-term temporal cues and the global long-term relations to exploit the multi-scale temporal cues in video sequences.

Extra input data like attributes are also helpful for the ReID task. FT-CNN [16] and Attribute-Person Recognition (APR) [17] improve image-based person ReID by training jointly with attribute data. Zhao et al. [29] proposed an attribute-driven method for feature disentangling and frame re-weighting for video-based ReID.

2.2. Metric Learning Based

Metric learning based methods mainly focus on the loss function and the sampling scheme for learning a discriminative embedding in the feature space. Deep metric learning defines a metric among

samples to compute the loss, which focuses on maximizing inter-class similarities and inter-class differences in the feature space.

IDE (Identity) [30] network treats each person as a class and directly exploits identity information as supervision signals to learn discriminative features using cross-entropy classification loss.

Triplet loss [31] is a current commonly used metric method for person ReID. It employs a constraint that the feature distance of person images with the same identity should be smaller than the distance of ones with different identities. Based on the triplet loss, Chen et al. [20] proposed a quadruplet loss, which further forced the intra-class distance to be less than the inter-class distance between two other classes.

The performance of all the metric losses mentioned above is greatly influenced by the sampling scheme. Many works [3,14,32–34] considered that the mining of hard samples plays an essential role in the performance of deep metric learning for person ReID. Hermans et al. [14] proposed batch hard triplet loss to select the hardest positive and hardest negative of each anchor in a mini-batch to compute the triplet loss. Ristani et al. [3] developed it into Adaptive Weighted Triplet Loss (AWTL). In [34], Yu et al. discussed the robustness of the batch hard triplet to outliers and proposed a more robust loss called Hard-Aware Point-to-Set (HAP2S) loss. Xiao et al. [33] proposed Margin Sample Mining Loss (MSML), which expands the batch hard triplet from a triplet loss to a quadruplet loss. The graph-based metric is also used in the ReID task. For example, Ye et al. [35] proposed a Dynamic Graph Matching (DGM) method for label estimation and unsupervised video ReID.

2.3. ReID for Tracking

With the development of person ReID, ReID features of person images have become popular appearance features for Multiple Object Tracking (MOT) [36], especially for tracking-by-detect-based methods. A survey [37] has summarized some hand-crafted features such as color, edge, and texture features of person images for associating person bounding boxes of the same ID in the video. In [38], random forest was applied to learn robust ReID features for the tracking task. However, traditional methods cannot extract rather discriminative features to solve some difficult situations like occlusions, pose variances, illumination variances, etc. Deep learning based trackers [3,4,39] use a large-scale person ReID dataset to train the ReID model, which significantly improves the performance. Zhang et al. [4] exploited AlignedReID [25] as the feature extractor and directly connected person bounding boxes according to the Euclidean and Jaccard distances of ReID feature vectors. Such a simple method obtained great performance on the DukeMTMC tracking dataset [8] because of the strong ReID model. DeepCC [3] combines spatio-temporal information with discriminative ReID features to implement the data association. However, a systematic solution for the data, network, and training methods is missing for a good ReID model in tracking, and an efficient way to jointly perform detection and ReID is needed.

3. Our Approach

3.1. ReID-Head Network

To extract features of multiple persons in an image frame, the regular solution is first to use a human detector to detect the human body area and then feed those bounding box images into another ReID network to obtain the final representation. However, in this way, detection and ReID cannot be combined properly, and this two-stage paradigm does each job without reusing features.

Efficiency is an important factor in a video-based real-time system, and the detection and ReID feature extracting can share the low-level features to speed up. An attempt of this idea was proposed in Person Search [40,41], which adopts a Faster R-CNN-like [42] structure, first generating candidate bounding boxes from anchor, then feeding the ROI-pooling [43] feature to an identification net to get the ReID features and refined detection bounding boxes. Features are reused for both detection and ReID; however, there are many anchor boxes, and each will feed into an identification net separately

without sharing features. As the detection area keeps advancing, this fixed anchor-based structure cannot be integrated with some newly proposed detection methods such as CornerNet [44] and YOLOv3 [45].

In this paper, we propose a new standalone ReID module consisting of a convolutional ReID-Head and an ROIAlign [46] layer. It begins from the mid-level feature maps of a detection/pose model and then produces its final representation for the entire input image by a light-weight convolution network, as shown in Figure 2. The detection trunk network generates final bounding boxes, and an ROIAlign [46] is applied for each bounding box to obtain the final ReID features.

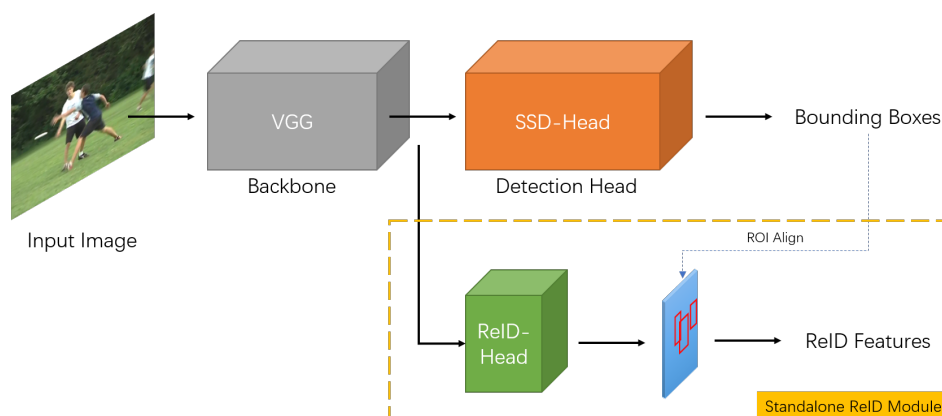


Figure 2. Our ReID-Head network structure. Given an input image, a backbone network is utilized for feature extraction. The extracted feature maps are then reused for both person detection through the Single Shot Detector (SSD)-Head and feature embedding through the ReID-Head. ROIAlign [46] is applied on detected bounding boxes (red bounding boxes in the figure) to obtain the final ReID features.

In this paper, we adopt a Single Shot Detector (SSD) network [47] as a case study. As shown in Figure 2, SSD [47] uses a VGG network [48] as the backbone, and we insert our ReID-Head module in the output of conv4_3, where the first classifier of SSD connects. As for our ReID-Head, we use three ResNet [49] bottlenecks followed by a 1×1 convolution to reduce the channel number to 10, and the ROIAlign [46] size is 5×5 , which makes the final feature vector for ReID have 250 dimensions.

Because the ReID module is a standalone part, the backbone in Figure 2 can be replaced by the more complicated ResNet-50 [49], ResNet-101 [49] or ResNeXt-101 [50], as in [46], while the detection head can be replaced by other detection heads like Feature Pyramid Networks (FPN) [51]. More complicated backbones and detection heads can further improve the performance, but that is not the focus of this paper. We choose the VGG [48] backbone and SSD [47] head as a case study in this paper.

There are several advantages of our proposed ReID-Head design:

- (1) It can jointly get detection and ReID results in a single network, and the network can be trained end-to-end.
- (2) The ReID-Head module can be integrated into both one-stage detectors like YOLO [52] and CornerNet [44] and two-stage detectors like Faster R-CNN [42] and Mask R-CNN [46], as well as keypoint detectors like OpenPose [53,54], benefiting from the progress in the detection area and improving.
- (3) It has high efficiency with feature reusing. Furthermore, the computational cost of ReID features will not linearly increase with more anchor boxes as in Person Search [40,41].

3.2. Instance Hard Triplet Loss

Person re-identification is a zero-shot problem, i.e., the IDs in the test set will not appear in the training set, and we need to distinguish different IDs without seeing any of them before. Therefore,

regarding each ID as a class and training a classification network to distinguish every training IDs may have inferior performance in testing.

Therefore, metric learning methods are introduced to learn a good embedding in the feature space supervised by a specific metric. The most famous metric learning method is the triplet-based metric, which was first introduced in the face recognition area [55,56].

A triplet is composed of an anchor sample, a positive sample, and a negative sample. The triplet loss force distance between the anchor and positive samples is smaller than the distance between the anchor and negative samples by a predefined margin, which can be formulated as follows:

$$L_{tri} = [D(x_a, x'_a) - D(x_a, x_p) + \alpha]_+ \quad (1)$$

where x_a and x'_a are the features of two different samples of person a , x_p is a feature of another person p , α is a margin constant, $[z]_+$ is $\max(0, z)$ to guarantee the loss is non-negative, and $D(\cdot)$ is a distance function between two features like the Euclidean distance.

Triplet loss can pull positive pairs together and push negative pairs away in the feature space. However, how to generate a proper training triplet is still a problem because there is a large number of potential combinations.

Batch hard triplet loss [14] generates triplets from the training batch by choosing the hardest negative and positive samples for each anchor sample as follows:

$$L_{BHT} = \sum_{p=1}^P \sum_{k=1}^K \left[\max_{i=1 \dots K} D(x_p^k, x_p^i) - \min_{\substack{n=1 \dots P \\ n \neq p}} D(x_p^k, x_n^j) + \alpha \right]_+ \quad (2)$$

where x_p^k is the p^{th} person's k^{th} sample; there are P persons in the batch, and each person has K samples. In this way, easy triplets that benefit the training a little are discarded, and hard triplets are mined. A visualization can be seen in Figure 3a.

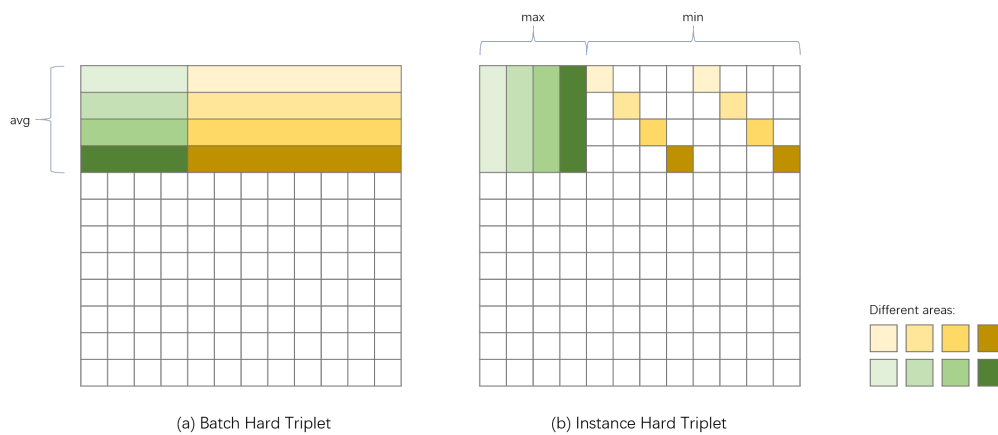


Figure 3. Visualization of the distance matrix between every two samples in a mini-batch for (a) the Batch Hard Triplet (BHT) and (b) Instance Hard Triplet (IHT) loss. Here, $P = 3$ and $K = 4$. The maximum in each green colored area and the minimum in each yellow colored area are computed, then K triplets are generated by BHT loss, and one triplet is generated by IHT loss for each person. As we can see, IHT needs less computation.

However, there are still two problems remain:

- (1) **Imbalance:** As shown in Equation (2) and Figure 3a, there is an imbalance between positive and negative samples that only K positive pairs, but $P \times (K - 1)$ negative pairs are compared, resulting in a harder negative sample mining.

- (2) Computation: This requires a huge computation overhead. For $P \times K$ samples in a batch, it needs $P \times K$ comparisons to generate triplets, leading to a computational complexity of $\mathcal{O}(P^2 K^2)$.

Unlike the motivation for batch hard triplet loss [14] to find the hardest positive and negative pairs in a batch, our motivation of instance hard triplet loss derives from the observation of persons in a video sequence. As shown in Figure 4, there are multiple persons in an image frame. If we want to identify each person in the subsequent frames, we should guarantee that the feature representations are similar among those frames across time. Meanwhile, to distinguish all persons correctly, we should also force the representation of persons in the same image to be different. Based on the above observation, we come up with a new loss function with the following form:

$$L_{IHT} = \sum_{p=1}^P \left[\max_{k=1 \dots K} \max_{i=1 \dots K} D(x_p^k, x_p^i) - \min_{j=1 \dots K} \min_{\substack{n \in \mathbb{P}_j \\ n \neq p}} D(x_p^j, x_n^j) + \alpha \right]_+ \quad (3)$$

where x_p^k is the feature of the p^{th} person in the k^{th} image, K is the number of images, P is the number of persons that appear in all K images, and \mathbb{P}_j is a set of all persons that show up in the j^{th} image.

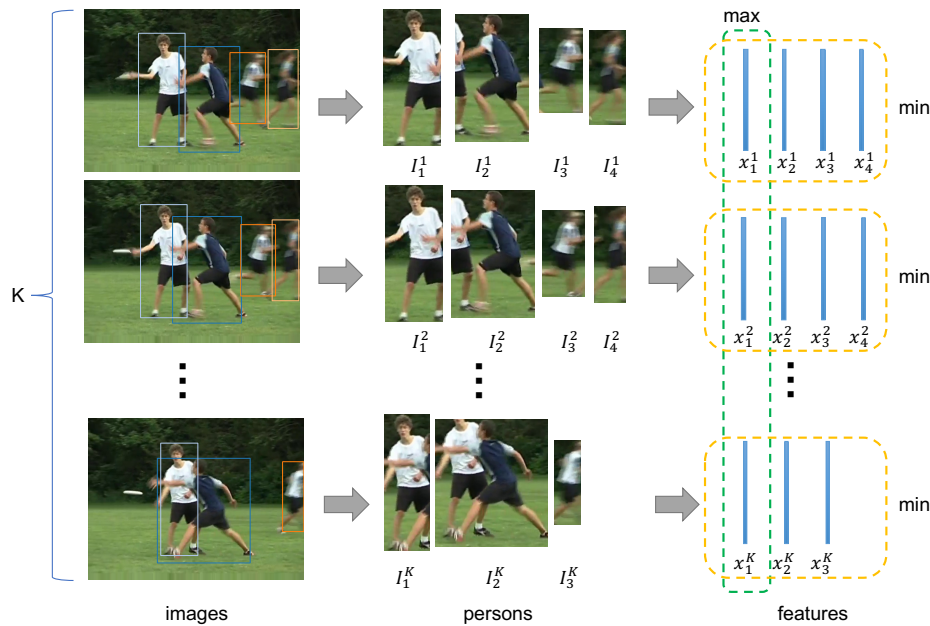


Figure 4. The motivation of our instance hard triplet loss is to keep features consistent across frames and distinguishable with other persons in the same frame. I_p^k is the p^{th} person in the k^{th} image and x_p^k is the corresponding feature. Note that a variable number of persons shows up in every frame, because persons often appear and disappear in a video.

Recalling Figure 4, the proposed instance triplet hard loss in Equation (3) selects P persons appearing in all K images, and three constraints are required: (1) features of each person should be consistent across all images; (2) features of all persons in the same image should be distinguishable; and (3) even the hardest positive sample should still be more similar to the anchor than the closest negative sample in every image by a predefined margin α .

The above motivation is clear, and the formulation is in line with our intuition. This loss functions also show much flexibility and can be generalized to an image-based cross-camera problem, where we can group a $P \times K$ batch into K groups with P different persons in each group. The variant can be formulated as follows:

$$L'_{IHT} = \sum_{p=1}^P \left[\max_{k=1 \dots K} \max_{i=1 \dots K} D(x_p^k, x_p^i) - \min_{j=1 \dots K} \min_{\substack{n=1 \dots P \\ n \neq p}} D(x_p^j, x_n^j) + \alpha \right]_+ \quad (4)$$

A visualization is shown in Figure 3b, where we only compare negative samples in the same group.

Different from batch hard triplet loss, instance hard triplet loss only compares the anchor person with persons in the same image rather than all negative samples. For K images with average P persons in each image, we compare the anchor sample with K positive samples and P negative samples, rather than $P \times (K - 1)$ negative sample in batch hard triplet loss. Thus, the proposed loss is more balanced between positive and negative pairs. On the other hand, we generate only one triplet for each person with comparison times of K^2 and $P \times K$ for positive and negative pairs, respectively. Therefore, for a batch with P person, the computational complexity is $\mathcal{O}(PK(P + K))$. Since we exploit all instances of each person and find the hardest triplet, we call it instance hard triplet loss.

In summary, compared with the widely-used triplet batch hard loss, our proposed loss is:

- (1) more balanced between positive and negative pairs;
- (2) faster with a smaller computational complexity of $\mathcal{O}(PK(P + K))$ compared to $\mathcal{O}(P^2K^2)$;
- (3) competitive in performance, which can be seen in the experimental part of this paper;
- (4) more general for both the cross-camera and in-video ReID problem, the image-based and video-based ReID problem, as well as training samples with an indefinite quantity persons.

3.3. Unsupervised In-video ReID

In practice, labeling cross-camera ReID data is expensive and time consuming, because it involves identifying a person from a camera network. Although our in-video ReID only requires associating a person from continuous frames, which is much easier for labeling training data, annotating for each new scenario still needs much manpower, and this deficiency limits the potential large-scale practical applications of in-video ReID.

Can we train an in-video ReID model in an unsupervised way? We observe the fact that a person can only appear once in a frame; in other words, we will not see the same person twice at the same time. Therefore, we can guarantee that the detected persons within the same frame are different persons, and images of other persons in the same frame can be used as negative samples. Now, the problem is how to get positive samples of a person with the same identity. In this paper, we propose a Reciprocal Identity Association (RIA) method to associate the same person across frames automatically for in-video ReID.

For a person p in a frame, we denote the k -nearest neighbors as:

$$\mathcal{N}(p, k) = \{g_1, g_2, \dots, g_k\}, |\mathcal{N}(p, k)| = k \quad (5)$$

where g is the persons in the next frame, g_i is the i^{th} sample in the top k ranking list of p , and $|\cdot|$ is the number of samples. If g_i is in the k -nearest neighbors of p , in turn, p should also be in the k -nearest neighbors of g_i if they are the same person, and the k -reciprocal neighbors [57] can be defined as:

$$\mathcal{R}(p, k) = \{g_i | (g_i \in \mathcal{N}(p, k)) \wedge (p \in \mathcal{N}(g_i, k))\} \quad (6)$$

Persons in $\mathcal{R}(p, k)$ are potential positive samples. With a larger k , it is more likely that \mathcal{R} contains the correct match, but it also brings more ambiguity into determining the corresponding identity. Because there is only one correct match in the next frame, we choose $k = 1$. If $|\mathcal{R}(p, 1)| = 1$, we take it as a positive sample. In this way, some person associations will be missed ($|\mathcal{R}(p, 1)| = 0$, no sample meets the condition), but the remaining association is more reliable. In our experiments, wrong associations with false supervision signals are more likely to cause model degeneration. Although we dropped

some potentially available associations, in a real scenario, we can usually collect many unlabeled data, which is much less expensive than collecting labeled data, so plenty of reliable associations can be automatically generated, and this problem could be compensated.

4. Experiments

4.1. Datasets

For the cross-camera ReID task, we used three widely-used public datasets, Market-1501 [1], DukeMTMC-reID [7], and CUHK03 [6], while for the in-video ReID task, we used a new proposed in-video ReID dataset called PoseTrack-ReID.

Here is the introduction for the three public cross-camera ReID datasets:

Market-1501 was collected at Tsinghua University using 6 cameras. There were 1501 identities and 32,668 bounding box images in total. Those bounding box images had a fixed size of 128×64 and were generated by a Deformable Part Model (DPM) [58] pedestrian detector.

DukeMTMC-reID is an image-based person re-identification based on the DukeMTMC dataset [8], in the format of the Market-1501 dataset. It crops pedestrian images from the videos every 120 frames, yielding in total 36,411 bounding boxes with IDs.

CUHK03 was proposed in [6], collecting images using 10 cameras. Each identity was observed by a pair of 2 cameras, and there were 1467 identities in all. Both DPM [58] detected and manually labeled bounding boxes were provided. We used the labeled version in this paper. A new recent protocol proposed in [57] was adopted in this paper with a fixed split for which 7368 images of 767 identities were used for training and 6728 of 700 identities were used for the testing.

Despite the fact that the above three public cross-camera datasets are widely used, they are not applicable to the in-video ReID task. To solve the in-video ReID problem, we needed a dataset to train and evaluate an in-video ReID method, which should (1) be video-based, (2) have annotations for persons appearing in the same frame, (3) be large enough, and (4) provide full images so the pedestrian detection errors can be examined. However, no existing ReID dataset satisfied those demands. Therefore, we proposed a new large-scale video-based in-video ReID dataset, called PoseTrack-ReID.

Our PoseTrack-ReID raw images came from a pose tracking dataset, PoseTrack [59]. The original PoseTrack [59] dataset contained human keypoint annotations, but did not have the bounding box of a human body. Based on the keypoint annotations, we first removed persons with less than 6 keypoints and obtained the bounding box of all keypoints. Then, we adjusted those bounding boxes to surround the whole body. We also removed frames with less than two persons present.

We split PoseTrack-ReID into a fixed train/val/test division with 250/37/50 videos, respectively. Statistical information and a comparison with the above three cross-camera datasets and some video-based datasets are shown in Table 1. Samples are shown in Figure 5.

Table 1. Comparison of the statistical information between PoseTrack-ReID and some widely-used cross-camera ReID datasets. # is the number of the corresponding item. MTMC, Multi-Target, Multi-Camera.

Name	#identities	#bboxes	Video?	Full Image?
CUHK03 [6]	1467	13,164		
Market-1501 [1]	1501	32,217		
DukeMTMC-reID [7]	1812	36,441		✓
PRID2011 [60]	934	24,541	✓	
iLIDS-VID [61]	300	42,495	✓	
MARS [62]	1261	1,191,003	✓	
PoseTrack-ReID	3088	84,443	✓	✓

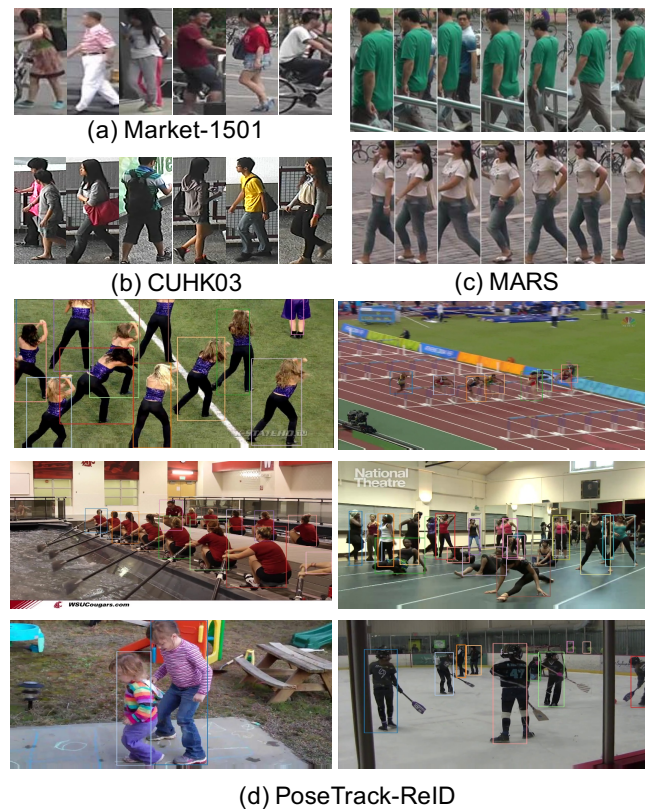


Figure 5. Samples from the proposed in-video (d) PoseTrack-ReID dataset and some other widely-used cross-camera ReID datasets, including (a) Market-1501 [1], (b) CUHK03 [6], and (c) MARS [62].

Note that the MARS [62] dataset contained more bounding boxes from a long continuous sequence, but it had fewer identities and missed full-frame images. MARS [62] contained only campus images of Tsinghua University. DukeMTMC-reID [7] also only contained campus images of Duke University. As a comparison, our PoseTrack-ReID dataset contained a large-scale of identities and bounding boxes, with full-image videos available in diverse scenes. As shown in Figure 5d, each video contained multiple persons, and the persons could be very similar to each other.

Each video had about 31 labeled frames, and unlabeled images before and after those labeled frames were also provided for future unsupervised methods.

4.2. Evaluation Protocol

Cross-camera ReID: We used the standard single query setting for cross-camera ReID as in [1]. For each query image, a model was used to retrieve images belonging to the same person from different cameras.

In-video ReID: For each video in PoseTrack-ReID, we left the last 15 labeled frames in each sequence of the test set as gallery images only, and the rest of frames were used as query images and potential gallery images as well. We performed an in-video person ReID in the recent subsequent frames, more specifically, for persons in a frame, and we identified the same person in the next several frames. In our experiments, we chose the frame interval $G \in \{1, 5, 10, 15\}$ to identify query persons in the image that was G frames after the query image. For different G , the total query bounding boxes were slightly different. We assigned a detected bounding box with the Intersection over Union (IoU) larger than 0.5 with a ground truth box the corresponding ID label and an ID of -1 otherwise.

Evaluation metric: We used rank-1 accuracy as our evaluation metric for both the cross-camera ReID task and the in-video ReID task. Given a query image, a ranking list was obtained based on the similarities between the query image and all candidates. The first image (rank-1) in the ranking list

was the most similar image, and the percentage that it belonged to the same person with the query image was computed.

4.3. Implementation Details

We implemented our network using the PyTorch framework. In all experiments, the margin constant α was fixed to 0.3. We made a clean network design so that the results could be easy to be reproduced, and our code will also be publicly released. Two networks were designed:

ReID-Head: The input image size was 300×300 . We used an SSD network pre-trained on the PASCAL VOC dataset [63] with an mAP of 77.4% on the VOC2007 Test. To compare different ReID methods with the same detection performance fairly, we fixed the detection part and only trained the proposed ReID-Head. Besides, the relationship between detection and ReID performance had already been fully investigated in [19]. Because the ReID-Head was light-weight, we only trained it for 10 epochs using an Adam [64] optimizer with the learning rate linearly increasing from 5^{-5} to 5^{-4} .

Baseline for cross-camera ReID: To evaluate the efficiency of our proposed instance hard triplet loss on cross-camera ReID task, we compared it with the widely-used state-of-the-art metric batch triplet [14] loss on three public cross-camera ReID datasets. As for the network structure, we chose the IDE network [30], which was commonly used as a baseline, like in [7,30,65–68]. The IDE network [30] uses a standard ResNet-50 structure [49] and the pre-trained weights on ImageNet for initialization. After the last convolution layer, a global average pooling follows to get the final ReID feature, which is a 2048-dimension vector for every sample. The input image size we chose was 256×128 , and only random horizontal flipping data augmentation was used. We used an Adam [64] optimizer with an initial learning rate 5^{-5} and linearly increased the learning rate to 1^{-3} within 20 epochs. We kept this learning rate for 60 epochs and then lowered it to 1^{-4} for 20 epochs and 1^{-5} for another 80 epochs. We used a PK sampling strategy [14] with $P = 32$ and $K = 4$, which meant there were 32 persons and 4 images for each person in a mini-batch. The same network structure was used for instance hard triplet loss. Except for the loss function, all the remaining settings were identical, and the results on three public cross-camera datasets are reported.

Baseline for in-video ReID: To evaluate the efficiency of the proposed ReID-Head network and instance hard triplet loss on in-video ReID task, we compared it with a state-of-the-art cross-camera ReID method, Part-based Convolutional Baseline (PCB) [2]. For the PCB network, following [69–71], we used an input image size of 256×128 . The number of horizontal stripes was fixed to 4 accordingly. The PK -sampling strategy with $P = 12$ and $K = 4$ was used to help the network converge. We trained it on the widely-used DukeMTMC-reID [7] cross-camera ReID dataset and then evaluated its performance on the in-video ReID task.

4.4. Cross-camera ReID Results

We first compared our proposed instance hard triplet loss with the widely-used batch hard triplet loss on the IDE [30] network for the cross-camera ReID task. The result is shown in Table 2. As we can see, on all three datasets, our instance hard triplet loss outperformed the batch hard triplet loss. In the meantime, our new loss could train the network faster with an iteration time of 0.53 s, saving about 35% training time, which meant the proposed loss was both better and faster.

The faster speed could be attributed to the smaller computation complexity, and the better performance could be attributed to the new hard triplets mining mechanism with more balanced positive and negative pairs.

Table 2. Comparison of two loss functions on the IDE [30] network for the cross-camera ReID task. Rank-1 accuracy is reported. The time for an iteration (forward + backward time) was tested on an NVIDIA Titan Xp GPU. Our proposed instance hared triplet loss achieved a better performance with less training time.

Method	Market-1501	DukeMTMC-reID	CUHK03	Time
BHT [14]	85.9	78.1	58.6	0.81s
IHT (ours)	87.0	80.3	59.2	0.53s

With a large input image size (like 384×128), more data augmentation (like random crop and random erasing [68]), a stronger backbone (like ResNet-101 [49]), multiple part-based features (like six horizontal features in PCB+RPP [2]), and other tricks, we may achieve a better performance, but that was not our main purpose. We kept our network design simple to demonstrate the flexibility, speed, and efficiency of our new loss; hence, the results were easy to reproduce, and we will also release our implementation.

4.5. In-video ReID Results

We also investigated the generality of our novel loss function and the efficiency of the ReID-Head network for the in-video ReID task.

We trained our ReID-Head network with different K in the PK -sampling strategy for the ablation study, where K is the number of images for each person in a mini-batch. The results are shown in Table 3. As we can see, when $K = 2$, the ability of hard example mining could not be fully utilized, thus resulting in an inferior performance. When $K = 8$, the time span was large, so a more diverse appearance could be seen in the training time. However, in this way, as we discussed in the Introduction part, short-term features were discarded. When K was very large, the task was more like a cross-camera ReID, but without a camera difference, viewpoint changes, and scene transition. In our in-video ReID problem, we cared more about ReID in the next few frames, so too big K was unnecessary, and note that we only had a length of about 31 labeled frames in each video. Therefore, in the following experiments, we chose $K = 6$.

Table 3. Experimental results on PoseTrack-ReID with different K values by our ReID-Head network. K is the number of images for each person in a mini-batch, and G is the frame interval. Rank-1 accuracy is reported.

K	G = 1	G = 5	G = 10	G = 15
2	39.1	37.9	36.3	34.9
4	39.9	39.2	37.8	36.4
6	40.0	39.2	37.8	35.9
8	40.0	39.0	37.9	36.3

We also evaluated the results using two sets of bounding boxes in the gallery: one was detected using the SSD trunk, and the other was the labeled ground truth. As shown in Table 4, using the labeled bounding boxes yielded a much better result, which revealed the huge influence of the detection accuracy on ReID performance. With better detection, we could achieve a better ReID result, because a good bounding box contained more body areas and less background distracting information. Some examples of both detected and labeled results are shown in Figure 6.

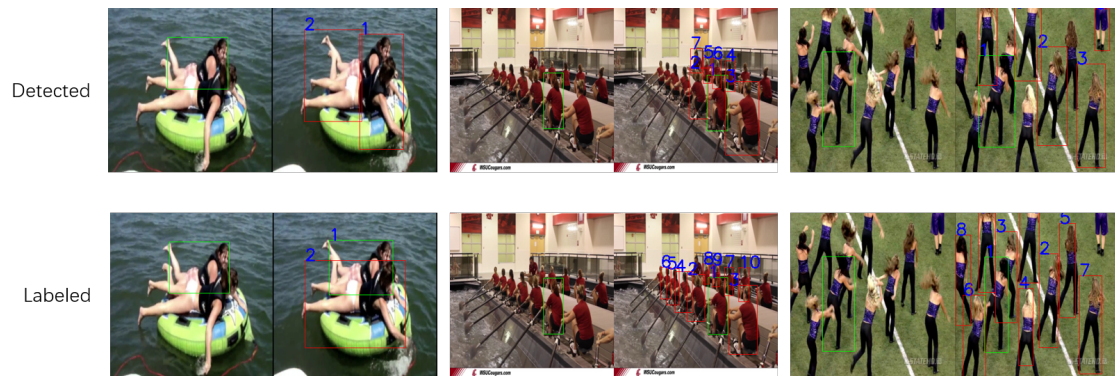


Figure 6. Some sample results obtained by our in-video ReID method. For each column, the first image is the query frame, and the second is the gallery frame. Query and candidates' bounding boxes are drawn. The number in the gallery is the ordinal number in the ranking list, so the number 1 is the most similar match. Wrong matches are drawn in red color, and correct matches are drawn in green.

Many failure cases were caused by the wrong detection, as shown in the first column of Figure 6; all candidate bounding boxes were wrong, so the ReID-Head could not find a correct match. Detection affected ReID so much that we designed our ReID-Head as a separate independent module. Unlike some other rigid network designs, in this way, as the detection area progressed, we could constantly shift to the new better detection frameworks and achieve a better ReID performance. Moreover, our ReID-Head could be plugged into both one-stage and two-stage detection networks with good flexibility.

Another thing that needs to be mentioned is that rank-1 accuracy when $G = 1$ using labeled ground truth bounding boxes was as high as 99.5%. After investigating the image results, we found it normal because when $G = 1$, the query frame and that right after the next frame (gallery frame) were very similar. For a 24 fps video, $G = 1$ meant only a $1/24$ second time interval, and the images did not change much. Therefore, the query and gallery images were almost the same.

Table 4. Results on the PoseTrack-ReID using detected and labeled bounding boxes, respectively. Rank-1 accuracy is reported.

Bounding Boxes	G = 1	G = 5	G = 10	G = 15
detected	40.0	39.2	37.8	35.9
labeled	99.5	93.6	87.1	81.1

We also evaluated the performance of a state-of-the-art cross-camera ReID method, the PCB [2] model, on the in-video ReID task. We first trained the PCB model on a popular large-scale cross-camera ReID dataset, the DukeMTMC-reID [7] dataset, and our implementation achieved a rank-1 accuracy of 85.0%, which was slightly better than the 83.3% rank-1 accuracy reported in the original paper. We then evaluated this model on the PoseTrack-ReID dataset to obtain in-video ReID performance.

Results are shown in Table 5, and there were three interesting observations that can be obtained from the results:

- (1) Even a state-of-the-art model trained on a popular large-scale cross-camera ReID dataset still performed badly on the PoseTrack-ReID dataset, because in-video ReID was a different problem with cross-camera ReID, and we needed to train a new model to fit for the new job, that is why we proposed a new dataset, a new network structure, and a new loss function to answer how to train for the new in-video problem;
- (2) Unlike ReID-Head, the performance of PCB did not descend with a larger G , because cross-camera ReID is a long-term problem, which discards short-term clues, while in-video ReID is a short-term problem;

- (3) Our light-weight ReID-Head was much faster than the PCB model. Unlike the traditional two-stage way with independent detection and the ReID model, our ReID-Head could achieve real-time multi-person ReID feature extracting with almost no increasing time by fully reusing features when the number of persons in a video increased. When bounding box number was larger than the maximum batch size of the GPU, traditional ReID models would need multiple forwards, costing even more time.

Table 5. Comparison of PoseTrack-ReID for the in-video ReID task with the state-of-the-art PCB [2] model. Feature extracting time t on an NVIDIA Titan Xp GPU for a bounding box and for an image with 10 bounding boxes are also provided. Rank-1 accuracy is reported. PCB, Part-based Convolutional Baseline.

Method	G = 1	G = 5	G = 10	G = 15	t/bbox	t/img
PCB [2]	20.2	20.6	20.9	20.8	10.5 ms	14.7 ms
ReID-Head	40.0	39.2	37.8	35.9	1.7 ms	1.7 ms

The key point was not the network structure, but the loss function. A state-of-the-art cross-camera ReID model with classification or batch hard triplet loss is not applicable to the in-video ReID task due to the varying number of persons in each frame. With the help of the proposed instance hard triplet loss, even a simple network structure like ReID-Head in this paper could achieve a better in-video ReID performance, which demonstrated the efficiency of our proposed loss function.

4.6. Unsupervised In-video ReID Results

We leveraged the proposed reciprocal identity association method to match the same person across frames. We used those generated data as a training set to train a ReID-Head network. The results are shown in Table 6. As we can see, when tested on detected bounding boxes, the unsupervised version was even slightly better than the supervised version. The explanation was that when training with ground truth bounding boxes, detection errors were missing. Using RIA-generated data was actually acting as data augmentation, so the model was more robust to detection errors. When it came to labeled test data, the RIA-generated model was inferior to the model trained on human-labeled data.

Table 6. Comparison of models trained on human-labeled data and Reciprocal Identity Association (RIA)-generated data, respectively, on the PoseTrack-ReID dataset. Rank-1 accuracy is reported. Results on both detected bounding boxes and labeled bounding boxes are reported.

Training Data	Test Bbox	G = 1	G = 5	G = 10	G = 15
Human-labeled	detected	40.0	39.2	37.8	35.9
RIA-generated	detected	40.2	39.2	37.9	36.4
Human-labeled	labeled	99.5	93.6	87.1	81.1
RIA-generated	labeled	99.4	92.6	84.8	76.9

Those results demonstrated the efficiency of our RIA method and implied that our unsupervised model could achieve competitive accuracy in scenarios with no labeled data available.

5. Conclusions

In this paper, we investigated the in-video person ReID task. To address this problem, we proposed a new large-scale video-based in-video ReID dataset, PoseTrack-ReID, with full images available. We also proposed a ReID-Head network, which could incorporate both one-stage and two-stage human detectors, realizing a real-time multi-person ReID feature extracting with reused features. We designed a novel instance hard triplet loss, which could be applied in both cross-camera and in-video ReID problems even with an indefinite quantity of persons and bounding boxes. An unsupervised

reciprocal instance association method was also proposed so that we could obtain an in-video model in an unsupervised way, further extending the potential applications of in-video ReID. Next, we will integrate the in-video ReID method into a human tracking system to improve its performance.

Author Contributions: Conceptualization, X.F.; methodology, X.F. and W.J.; software, X.F.; validation, X.F., H.L. and W.J.; formal analysis, X.F., W.J., and W.M.; investigation, H.L.; resources, W.J., W.M., and H.Y.; data curation, X.F.; writing, original draft preparation, X.F.; writing, review and editing, X.F., H.L., and W.J.; visualization, X.F.; supervision, W.J., W.M. and H.Y.; project administration, W.J., W.M., and H.Y.; funding acquisition, W.J., W.M., and H.Y. All authors read and agreed to the published version of the manuscript.

Funding: Supported by the National Natural Science Foundation of China (No. 61633019), the Public Projects of Zhejiang Province (No. LGF18F030002), and the Science Foundation of Chinese Aerospace Industry (JCKY2018204B053).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable Person Re-Identification: A Benchmark. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
2. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
3. Ristani, E.; Tomasi, C. Features for Multi-Target Multi-Camera Tracking and Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6036–6046.
4. Zhang, Z.; Wu, J.; Zhang, X.; Zhang, C. Multi-Target, Multi-Camera Tracking by Hierarchical Clustering: Recent Progress on DukeMTMC Project. *arXiv* **2017**, arXiv:1712.09531.
5. Ning, G.; Huang, H. LightTrack: A Generic Framework for Online Top-down Human Pose Tracking. *arXiv* **2019**, arXiv:1905.02822.
6. Li, W.; Zhao, R.; Xiao, T.; Wang, X. DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014.
7. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled Samples Generated by Gan Improve the Person Re-Identification Baseline in Vitro. *arXiv* **2017**, arXiv:1701.07717.
8. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016.
9. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Mask-Guided Contrastive Attention Model for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.
10. Ge, Y.; Li, Z.; Zhao, H.; Yin, G.; Yi, S.; Wang, X.; Li, H. FD-GAN: Pose-Guided Feature Distilling GAN for Robust Person Re-Identification. In Proceedings of the Advances in Neural Information Processing Systems 31 (NIPS 2018), Montréal, QC, Canada, 3–8 December 2018.
11. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Advances in Neural Information Processing Systems 27 (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
12. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
13. Raaj, Y.; Idrees, H.; Hidalgo, G.; Sheikh, Y. Efficient Online Multi-Person 2D Pose Tracking with Recurrent Spatio-Temporal Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
14. Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. *arXiv* **2017**, arXiv:1703.07737.

15. Geng, M.; Wang, Y.; Xiang, T.; Tian, Y. Deep transfer learning for person re-identification. *arXiv* **2016**, arXiv:1611.05244.
16. Matsukawa, T.; Suzuki, E. Person re-identification using CNN features learned from combination of attributes. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 2428–2433.
17. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Yang, Y. Improving person re-identification by attribute and identity learning. *arXiv* **2017**, arXiv:1703.07220.
18. Fan, X.; Luo, H.; Zhang, X.; He, L.; Zhang, C.; Jiang, W. SCPNet: Spatial-Channel Parallelism Network for Joint Holistic and Partial Person Re-Identification. In Proceedings of the Asian Conference on Computer Vision, ACCV, Singapore, 1–5 November 2018.
19. Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; Tian, Q. Person Re-Identification in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
20. Chen, W.; Chen, X.; Zhang, J.; Huang, K. Beyond triplet loss: a deep quadruplet network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; Volume 2.
21. Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; Zheng, N. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1335–1344.
22. Liu, H.; Feng, J.; Qi, M.; Jiang, J.; Yan, S. End-to-end comparative attention networks for person re-identification. *IEEE Trans. Image Process.* **2017**, *26*, 3492–3506. [[CrossRef](#)]
23. Varior, R.R.; Haloi, M.; Wang, G. Gated siamese convolutional neural network architecture for human re-identification. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 791–808.
24. Bai, X.; Yang, M.; Huang, T.; Dou, Z.; Yu, R.; Xu, Y. Deep-Person: Learning Discriminative Deep Features for Person Re-Identification. *arXiv* **2017**, arXiv:1711.10658.
25. Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, Q.; Jiang, W.; Zhang, C.; Sun, J. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv* **2017**, arXiv:1711.08184.
26. Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Wang, X.; Tang, X. Spindle Net: Person Re-Identification With Human Body Region Guided Feature Decomposition and Fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
27. Wei, L.; Zhang, S.; Yao, H.; Gao, W.; Tian, Q. Glad: global-local-alignment descriptor for pedestrian retrieval. In Proceedings of the 2017 ACM on Multimedia Conference, Mountain View, CA, USA, 23–27 October 2017; ACM: New York, NY, USA, 2017; pp. 420–428.
28. Li, J.; Wang, J.; Tian, Q.; Gao, W.; Zhang, S. Global-Local Temporal Representations for Video Person Re-Identification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019.
29. Zhao, Y.; Shen, X.; Jin, Z.; Lu, H.; Hua, X.S. Attribute-Driven Feature Disentangling and Temporal Aggregation for Video Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
30. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person Re-Identification: Past, Present and Future. *arXiv* **2016**, arXiv:1610.02984.
31. Ding, S.; Lin, L.; Wang, G.; Chao, H. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognit.* **2015**, *48*, 2993–3003. [[CrossRef](#)]
32. Ahmed, E.; Jones, M.; Marks, T.K. An improved deep learning architecture for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3908–3916.
33. Xiao, Q.; Luo, H.; Zhang, C. Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-identification. *arXiv* **2017**, arXiv:1710.00478.
34. Yu, R.; Dou, Z.; Bai, S.; Zhang, Z.; Xu, Y.; Bai, X. Hard-Aware Point-to-Set Deep Metric for Person Re-Identification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.

35. Ye, M.; Li, J.; Ma, A.J.; Zheng, L.; Yuen, P.C. Dynamic Graph Co-Matching for Unsupervised Video-Based Person Re-Identification. *IEEE Trans. Image Process.* **2019**, *28*, 2976–2990. [[CrossRef](#)] [[PubMed](#)]
36. Zajdel, W.; Zivkovic, Z.; Krose, B. Keeping track of humans: Have I seen this person before? In Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, 18–22 April 2005; pp. 2081–2086.
37. Zhou, S.; Ke, M.; Qiu, J.; Wang, J. A Survey of Multi-object Video Tracking Algorithms. In Proceedings of the International Conference on Applications and Techniques in Cyber Security and Intelligence, Shanghai, China, 11–13 July 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 351–369.
38. Cho, Y.J.; Kim, S.A.; Park, J.H.; Lee, K.; Yoon, K.J. Joint Person Re-identification and Camera Network Topology Inference in Multiple Cameras. *arXiv* **2017**, arXiv:1710.00983.
39. Jiang, N.; Bai, S.; Xu, Y.; Xing, C.; Zhou, Z.; Wu, W. Online inter-camera trajectory association exploiting person re-identification and camera topology. In Proceedings of the 2018 ACM Multimedia Conference on Multimedia Conference, Seoul, Korea, 22–26 October 2018; ACM: New York, NY, USA, 2018; pp. 1457–1465.
40. Xiao, T.; Li, S.; Wang, B.; Lin, L.; Wang, X. Joint Detection and Identification Feature Learning for Person Search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
41. Xiao, T.; Li, S.; Wang, B.; Lin, L.; Wang, X. End-to-End Deep Learning for Person Search. *arXiv* **2016**, arXiv:1604.01850.
42. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015.
43. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
44. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
45. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
46. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
47. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the European Conference on Computer Vision ECCV, Amsterdam, The Netherlands, 8–16 October 2016.
48. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
49. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
50. Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
51. Lin, T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
52. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
53. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional Pose Machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
54. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
55. Weinberger, K.Q.; Saul, L.K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J. Mach. Learn. Res. (JMLR)* **2009**, *10*, 207–244.

56. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A Unified Embedding for Face Recognition and Clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
57. Zhong, Z.; Zheng, L.; Cao, D.; Li, S. Re-Ranking Person Re-Identification with k-Reciprocal Encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
58. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. [[CrossRef](#)]
59. Andriluka, M.; Iqbal, U.; Insafutdinov, E.; Pishchulin, L.; Milan, A.; Gall, J.; Schiele, B. PoseTrack: A Benchmark for Human Pose Estimation and Tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
60. Hirzer, M.; Beleznaï, C.; Roth, P.M.; Bischof, H. Person re-identification by descriptive and discriminative classification. In Proceedings of the Scandinavian Conference on Image Analysis, Ystad, Sweden, 23–25 May 2011; pp. 91–102.
61. Wang, T.; Gong, S.; Zhu, X.; Wang, S. Person Re-Identification by Video Ranking. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014.
62. Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; Tian, Q. Mars: A Video Benchmark for Large-Scale Person Re-Identification. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016.
63. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (Voc) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [[CrossRef](#)]
64. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
65. Zheng, L.; Huang, Y.; Lu, H.; Yang, Y. Pose Invariant Embedding for Deep Person Re-Identification. *IEEE Trans. Image Process.* **2019**, *28*, 4500–4509. [[CrossRef](#)] [[PubMed](#)]
66. Sun, Y.; Zheng, L.; Deng, W.; Wang, S. SVDNet for Pedestrian Retrieval. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
67. Zheng, Z.; Zheng, L.; Yang, Y. Pedestrian Alignment Network for Large-Scale Person Re-Identification. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 3037–3045. [[CrossRef](#)]
68. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), San Francisco, CA, USA, 4–9 February 2017.
69. Zhong, Z.; Zheng, L.; Luo, Z.; Li, S.; Yang, Y. Invariance Matters: Exemplar Memory for Domain Adaptive Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
70. Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; Kautz, J. Joint Discriminative and Generative Learning for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
71. Yang, W.; Huang, H.; Zhang, Z.; Chen, X.; Huang, K.; Zhang, S. Towards Rich Feature Discovery With Class Activation Maps Augmentation for Person Re-Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.

