applied sciences

MDPI

Article

Hierarchical Feature Aggregation from Body Parts for Misalignment Robust Person Re-Identification †

Yuting Liu [1,2], Hongyu Yang [1,2] and Qijun Zhao [1,2,*]

1 College of Computer Science, Sichuan University, Chengdu 610065, China; yuting.liu@stu.scu.edu.cn (Y.L.); yanghongyu@scu.edu.cn (H.Y.)
2 National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China
* Correspondence: qjzhao@scu.edu.cn; Tel.: +86-028-85417865

† This paper is an extended version of our paper published in The 2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA).

Received: 8 March 2019; Accepted: 27 May 2019; Published: 31 May 2019

Abstract: In this work, we focus on the misalignment problem in person re-identification. Human body parts commonly contain discriminative local representations relevant with identity recognition. However, the representations are easily affected by misalignment that is due to varying poses or poorly detected bounding boxes. We thus present a two-branch Deep Joint Learning (DJL) network, where the local branch generates misalignment robust representations by pooling the features around the body parts, while the global branch generates representations from a holistic view. A Hierarchical Feature Aggregation mechanism is proposed to aggregate different levels of visual patterns within body part regions. Instead of aggregating each pooled body part features from multi-layers with equal weight, we assign each with the learned optimal weight. This strategy also mitigates the scale differences among multi-layers. By optimizing the global and local features jointly, the DJL network further enhances the discriminative capability of the learned hybrid feature. Experimental results on Market-1501 and CUHK03 datasets show that our method could effectively handle the misalignment induced intra-class variations and yield competitive accuracy particularly on poorly aligned pedestrian images.

Keywords: person re-identification; misalignment; hierarchical feature aggregation

1. Introduction

Typical person re-identification (re-ID) systems [1–3] can be broken down into three modules, i.e., person detection, person tracking, and person retrieval. It is generally believed that the first two modules are independent computer vision tasks, thus most re-ID methods focus on the last module, i.e., person retrieval. In this paper, if not specified, person re-ID refers to the person retrieval module. Defined as a classical image retrieval problem, person re-ID is considered as a process of matching identity classes between person-of-interest (query) and detected objects (large galleries) across cameras, which is a fundamental task in several fields such as surveillance, robotics, multimedia and forensics. It has been an area of intense research in the past few years.

Despite years of great efforts, person re-ID remains a challenging task due to the dramatic appearance variations in illumination, human pose, occlusion, and background. The varying poses or poorly detected bounding boxes often lead to misalignment of detected pedestrians (e.g., excessive background and missing or mis-aligned body parts), which is a critical challenge to robust person re-ID systems. The useless background noise and information loss due to misalignment can significantly compromise the feature learning and matching process. Figure 1 shows examples of mis-aligned pedestrian images.

Figure 1. Examples of mis-aligned pedestrian images in Market-1501 dataset caused by pose variations and detection errors. The corresponding image patches of same identity are semantically unmatched (e.g., human head to background).

To handle this problem, early works [4–8] extract features from predefined image patches such as grid cell and horizontal stripes to construct the globally aligned representations for person re-ID. These methods subjectively suppose that every person appears in a similar pose within a tightly surrounded bounding box, ignoring the complex realistic conditions. Thus, they fail to perform well on more difficult databases [5,9]. More reasonable body part partition fashion [10–13] has then been exploited to generate finely aligned representations. With the development of pose estimation techniques [14–18], the above mentioned works have been re-studied. The adapted methods either intuitively perform affine transformation in order to get standard pose-aligned images (PoseBox) [19] or implicitly learn the proper transformation parameters and generate modified pose images with the help of impactful spatial transformer network [20]. However, highly-accurate pose estimation was required to prevent abnormal pose-normalized pedestrian images. To mitigate the problems, we proposed in [21] to apply alignment on feature level by pooling the features around the body parts. Alignment on feature level can not only avoid unnecessary geometric deformation in image but also make full use of the context-aware information encoded in middle convolution layers that can compensate detection errors. Meanwhile, the pooling operation also favors translation and rotation. All these factors make our method more robust to pose estimation errors compared to previous image-level-alignment-based methods. Recent methods [22,23] share similar insights with us in implementing feature level alignment.

Hierarchical-based learning methods are widely used in many tasks. The methods in [24,25] use the hierarchical Hidden Markov Model (HMM) to estimate and synthesize the motion of fingers or full-body while the method in [26] proposes a Bayesian hierarchical model to learn and recognize natural scene categories. These works adopt hierarchies of models to describe the intermediate states or themes of complex motions and scenes. The method in [27] takes advantage of Convolutional Neural Networks to learn hierarchies of features for Scene Labeling. Such hierarchies of features assemble pixel inputs into elements from low-level details to high-level semantic concepts and form

good internal representations that are helpful for various visual perception tasks. Similar to these hierarchical-based learning methods, we propose to aggregate features from body parts with different levels of semantics.

Specifically, we construct a deep joint learning (DJL) network to learn misalignment robust feature representations from body parts for person re-ID. We propose to locally align the human bodies based on their landmarks, and pool the features around the body parts on feature maps rather than on original images. This way, our method can effectively handle the misalignment induced intra-class variations even though semantically corresponding body parts are not well aligned on the original images or the detected landmarks deviate from their true positions. As features from multiple layers abstract different level visual patterns of the same pedestrian image, we adopt a Hierarchical Feature Aggregation mechanism to enrich the feature representations for a pedestrian image by aggregating body part features with different levels of semantics. Besides, a Region Re-weighting strategy is applied to learn the importance weight of each body part as well as to mitigate the scale differences [28] among multiple convolution layers. Evaluation experiments on two public benchmark databases prove the effectiveness of our proposed method compared with existing state-of-the-art methods.

This paper is an extended version of our previous conference paper [21] with the following incremental contributions: (i) We further explore the identification performance of multiple layers for re-ID tasks from low-level to semantic-level and propose a Hierarchical Feature Aggregation (HFA) mechanism to take full advantage of different levels of features. (ii) We adopt a Region Re-Weighting (RRW) strategy to learn optimal weight of each body part as well as to mitigate the scale difference of multiple layers. (iii) We get further performance boost, obtaining 88.39% and 85.90% on Market-1501 and CUHK03 datasets. The rest of this paper is organized as follows. Section 2 reviews related work on deep learning based person re-ID methods, global and local features for re-ID and the pedestrian misalignment problem. Section 3 introduces in detail our proposed method, and Section 4 then reports our evaluation experiments. Finally, Section 5 concludes the paper.

2. Related Works

2.1. Deep Learning for Person Re-ID

Early methods solve the person re-ID problem mainly from two aspects, feature extraction and metric learning. Typical features used for person re-ID include color histograms [29–31], color names [9,32], local binary patterns (LBP) [30,33], gabor features [34] and scale invariant local ternary patterns (SILTP) [29,35]. Some researchers apply metric learning methods to seek for effective distance metrics for computing similarity between detected persons [6,29,30,36,37]. The emerging deep learning (DL) technology provides effective approaches for learning both feature representations and distance metrics. These DL-based person re-ID methods are dominating the re-ID community. Recently, attributes [38], transfer learning [39,40], re-ranking [41], mutual learning [42] and different levels of supervision [40,43,44] have also been studied.

2.2. Global and Local Features

Human visual system leverages both global (contextual) and local (saliency) information concurrently [45,46]. This observation supports that global and local features have correlated complementary information in different contexts. Most deep learning methods for person re-ID [47–49] follow the classical image classification mode [50], which favors intrinsically in learning global feature representations. However, these methods ignore the importance of local information. Some methods [5,6,51] utilize local information by decomposing images into horizontal stripes and learning effective local features in each patch. These local stripes in essence globally align the images of detected persons, and are thus still sensitive to misalignment of human bodies in different images.

2.3. Pedestrian Misalignment

Pedestrian misalignment caused by detectors or pose variations is a main challenge for feature matching across images. Most previous works partition pedestrian bounding box into grids or horizontal stripes to handle misaligned pedestrian images [5,9,29,51]. Nevertheless, these methods only work under the assumption of slight vertical misalignment but not for severe misalignment. Some methods [11,12] use the pictorial structure to construct well aligned pedestrian images. However, they only use local body parts while ignoring the global context, which results in suboptimal feature learning.

The recent PIE method [19] proposes a PoseBox fusion (PBF) CNN architecture that takes the original image, the PoseBox, and the pose estimation confidence as input to achieve a globally optimized tradeoff between the global and local feature representations. The PoseBox structure is similar to the pictorial structure [11,12] in enabling well-aligned pedestrian matching. The PDC method [52] first crops part regions and then transforms each part by a Pose Transformation Network (PTN) to automatically learn transformations such as translation, rotation and scale. The PTN outputs the final transformed part images and hence learns partly aligned representations. These methods all attempt to solve the misalignment problem at image level, with few exceptions that directly handle learned features. For example, Zhao et al. [22] followed human body structure to iteratively decompose and fuse features from different semantic region; Li et al. [53] exploited attention models to implicitly learn effective part representations without guidance of body part locations; and Wang et al. [23] encoded human poses in feature maps through bilinear pooling which aggregates appearance and part maps to compute part-aligned representations. Our method differs from them in the following three aspects.

- Our work constructs the "PoseBox" at feature level instead of the image level. We find that the image level PoseBox would lose their discriminative property due to pose estimation errors. In addition, the affine transformation employed by the PIE method may result in unwanted geometric distortion and deteriorating the intrinsic structure of human body. Figure 2 shows some examples of good and bad PoseBox constructed by PIE. Instead of image level affine transformation, we directly pool local body part features on feature maps, and organize them in a fixed order for feature level alignment (concatenate each body part features along channel dimensions). Meanwhile, we propose to model the spatial dependencies between those local body parts through cross-channel convolution computation. Thanks to the capability of CNN feature maps in context-aware semantic information, we suppose that the feature level alignment would be more robust to pose estimation errors.
- We apply max pooling inside local body part regions so as to find the most salient local details. HFA mechanism and RRW strategy are proposed to make the best of multi-level body part features. Our joint optimization of both global and local features further enhances the discriminative capability of learned feature representations for person re-ID.
- By avoiding complicated affine transformation, we can obtain pose aligned features in a simple and efficient way. Moreover, our method can be easily integrated with different person re-ID networks, and effectively enhance their identification accuracy.

Figure 2. Examples of good and bad PoseBox constructed by PIE: (Top row) original bounding boxes with detection errors/occlusions; and (Bottom Row) corresponding PoseBoxes.

3. Proposed Method

As shown in Figure 3, our proposed DJL network consists of three main components: the global branch base network, the local branch sub-network, and the multi-loss module. First, the input human body image is segmented into a number of body part regions (Section 3.1). The global branch base network extracts global representations from the original image (Section 3.2). The local branch sub-network then constructs misalignment robust local features according to the segmented body part regions and middle layer feature maps generated by global branch. With three Softmax losses, the multi-loss module optimizes global and local features jointly (Section 3.3). In this section, we introduce first the process of body part segmentation, then the global branch base network, and finally the proposed DJL network.

3.1. Body Part Segmentation

We first segment human body parts through deep pose estimation method CPM [16]. CPM outputs the coordinates of a set of 14 body parts and the corresponding confidence scores, i.e., head, neck, left and right shoulders, left and right elbows, left and right wrists, left and right hips, left and right knees, and left and right ankles. Several previous works [4,6,19] show that the torso and legs make the largest contributions and that integration of the head may introduce noise due to the unstable head detection. In this paper, we thus choose ten of the body parts as region boxes for local feature extraction, including left and right shoulders, left and right elbows, left and right hips, left and right knees, and left and right ankles. Figure 4 shows an illustration of the chosen body parts.

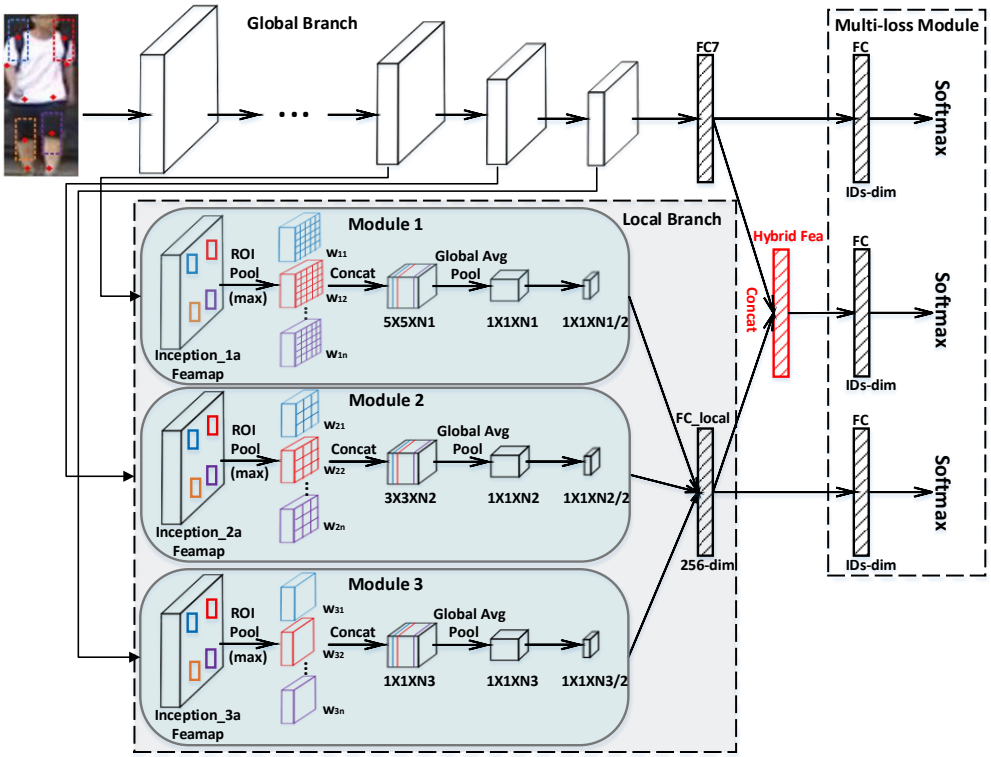


Figure 3. The proposed DJL network with InceptionNet as the base network. The input to DJL includes a pedestrian image and the human body landmarks. We segment ten body part regions according to the landmarks (Section 3.1). A local branch sub-net (Section 3.3) is specially designed in this paper to pool and aggregate multi-level body part representations from the feature maps generated by the global branch base network (Section 3.2). The multi-loss module then optimizes the global and local features jointly.

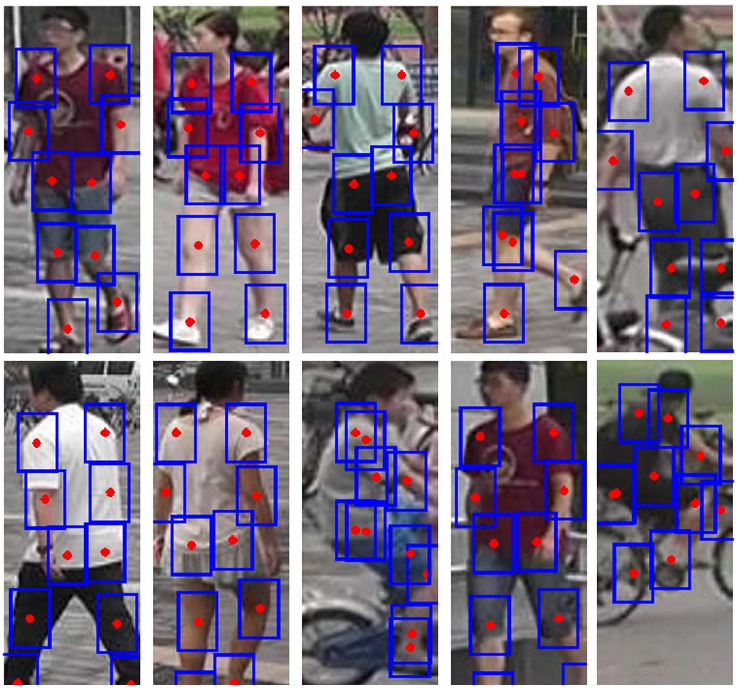


Figure 4. Examples of the segmented ten body parts used in our DJL network.

3.2. Base Networks

We utilize the widely used AlexNet [50], Residual-50 [54] and InceptionNet [48] as the base networks in our proposed method. We refer readers to respective papers for detail network descriptions. We adopt Identification model in this paper and edit the last FC layer to have the same number of neurons as the number of distinct IDs in the training set. As described in [49], the identification model yields superior performance to verification model for the reason that the former makes full use of the re-ID labels while the latter takes limited relationships into consideration, i.e., whether two input images belong to the same person.

3.3. The Deep Joint Learning Network

Two pairs of feature maps extracted by the base network are provided in Figure 5 to give insights into the model design. We observe that high responses are mostly concentrated on the local body parts and they often present attribute-relevant information (e.g., clothing type, color, accessories, etc.), and, when reasonably exploited, those body part features may be helpful to distinguish individuals. Motivated by this, we integrate body part features from low level to semantic level, resulting in misalignment-robust representations for matching.

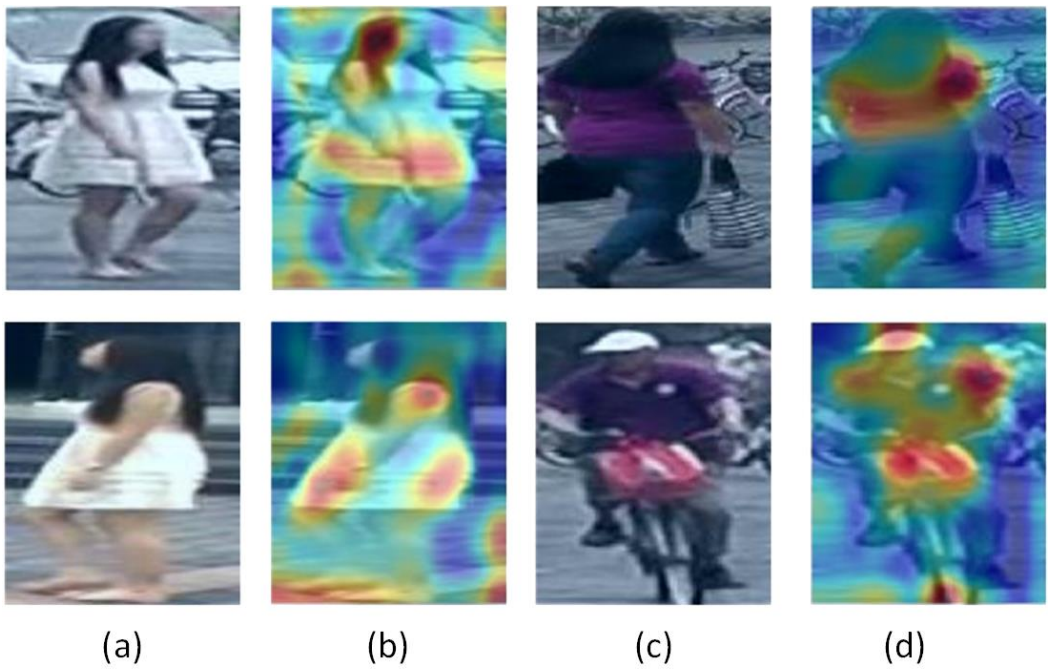


Figure 5. Two examples to show the effectiveness of the local body part features: (**a**) two images of the same person; (**b**) corresponding feature maps of (**a**); (**c**) two different persons; and (**d**) corresponding feature maps of (**c**).

3.3.1. Network Structure

The input to the DJL network contains a pedestrian image and its ten body parts. Each body part is represented by its position. The global branch of DJL is composed by the base networks, as previously described in Section 3.2. Its objective is to extract global features of pedestrians.

The local branch aims to learn misalignment-robust feature representations from low level to semantic level. It consists of several similar modules, each of which takes as input the output feature maps of a specific middle convolution layer from base network and generates local descriptors of that level. As shown in Figure 3, for a single module, RoI pooling layer [55] is adopted to learn sparse representations of each local body part. The RoI pooling layer uses max pooling to convert the features inside any region of interest window of size $h \times w$ into a small feature map with a fixed spatial extent

of $H \times W$, where H and W are layer hyper-parameters. It works by dividing the $h \times w$ RoI window into an $H \times W$ grid of sub-windows of approximate size $h/H \times w/W$ and then max-pooling the values in each sub-window into the corresponding output grid cell. Pooling is applied independently to each sub-window as in standard max pooling. Figure 6 shows an illustration of the RoI pooling operation. Given the middle-layer feature maps and coordinates of body part regions, we perform RoI pooling inside each region to select the most discriminative features. Then, those local body part features are concatenated along channel dimensions in a fixed order, and a global average pooling layer and a convolution layer follow to get the dimension-reduced local descriptors.

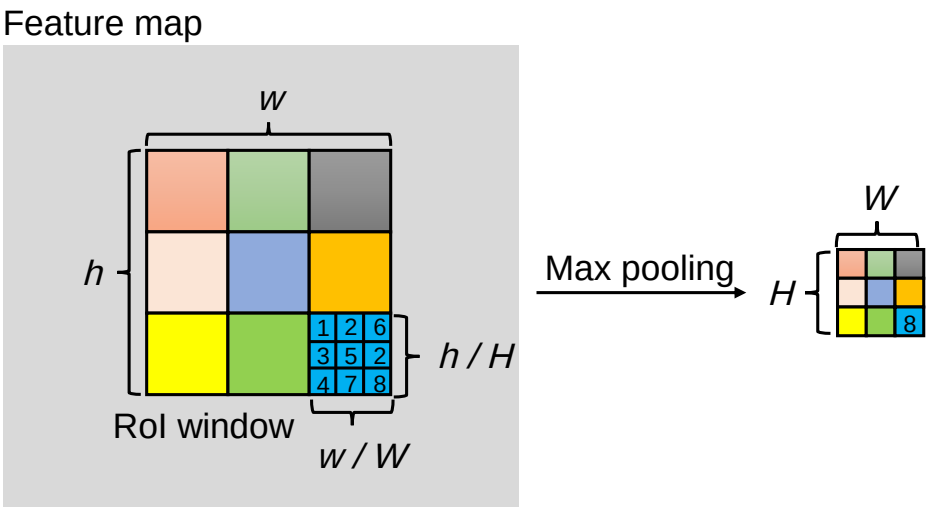


Figure 6. Illustration of the RoI pooling operation.

The multi-loss module consists of three full connection (FC) layers before Softmax loss computation. The sum of the three Softmax losses is used for loss computation. Dimensions of these FC layers are the number of distinct IDs in the training set. In Figure 3, as denoted by the red FC layer, the learned hybrid feature representation for final matching is defined as the concatenated FC7 activations (FC_ local + FC7). The motivation of our multiple loss module is to integrate the discriminative power of global and local features.

3.3.2. Hierarchical Feature Aggregation

Inspired by neuroscience, reasoning across multiple levels of hierarchies has been proven beneficial in some computer vision problems [24,26,27,56,57]. On the one hand, it has been demonstrated that details can be well captured by low-level features from shallow convolution layer rather than by high-level features. On the other hand, high-level features from deeper convolution layer get complementary semantic information as neurons in these layers have lager receptive fields. We thus adopt a Hierarchical Feature Aggregation mechanism to pool features from shallow to deep convolution layers of base network and aggregate the learned local descriptors from detail to semantic. For example, as shown in Figure 3, we perform RoI pooling at Inception_ 3a, Inception_ 2a, Inception_ 1a for InceptionNet with different pooling scales ($H \times W$). The output spatial extents are, respectively, 1×1, 3×3, and 5×5. Here, we adopt coarse spatial division 1×1 in deep layers and fine spatial division 5×5 in shallow layers to capture fine-grained features corresponding to local salient details. Finally, the pose aligned body part features from each module are concatenated to form the final multi-level local descriptors (denoted by FC_ local). We also adopt a Region Re-Weighting strategy (see Section 3.3.3) to make the Hierarchical Feature Aggregation mechanism more effective.

3.3.3. Region Re-Weighting

For the reason that pose estimation method (CPM) may induce ill-positioned body parts and different body part regions may have different importance for person re-identification, we intend to learn the importance weight of each body part region during training procedure. We call this strategy Region Re-Weighting (RRW). RRW performs an element-wise product between body part region features and the corresponding region weights. Formally, for each pooled body part feature of d-dimension $X_i = (x_{i1}, \cdots, x_{id})$, we introduce a weight parameter w_i, which scales per region features as $Y_i = (w_i \cdot x_{i1}, \cdots, w_i \cdot x_{id})$. During training, letting L be the loss we want to minimize, we use back propagation and chain rule to compute derivatives with respect to the weight factor w_i and body part features X_i.

$$\frac{\partial L}{\partial X_i} = \frac{\partial L}{\partial Y_i} \cdot w_i \qquad \frac{\partial L}{\partial w_i} = \sum_{j=1}^{d} \frac{\partial L}{\partial y_{ij}} \cdot x_{ij} \qquad (1 \leq i \leq 10) \tag{1}$$

As mentioned in [28], scales and norms of feature vectors from multiple layers may be quite different, and directly concatenating multi-level features may leads to poor performance as the "larger" features dominate the "smaller" ones. We find that combining RRW with HFA makes the training more stable and enables further performance improvements.

4. Experiments

4.1. Datasets and Protocol

4.1.1. Datasets

This study used CUHK03 [5] and Market-1501 [9] datasets for evaluation. The Market-1501 dataset is featured by 1501 IDs (750 for training and 751 for testing) with 32,668 cropped pedestrian bounding boxes. It contains 3368 query images and 19,732 gallery images (including 2793 distractors). For each query, we aimed to retrieve the ground-truth images from the 19,732 candidate images. This dataset is one of the largest benchmark datasets for person re-identification. Pictures were captured by six cameras: five high-resolution cameras and one low-resolution camera. The CUHK03 dataset contains 13,164 cropped pedestrian bounding boxes of 1360 identities (1160 for training, 100 for validation and 100 for testing) captured by six cameras. Each identity appears in two disjoint camera views (i.e., 4.8 images in each view on average). The bounding boxes of the pedestrians used in this study were generated by the DPM detector [58] instead of human annotated. This was to make the evaluation results more practical as in real-world automatic person re-ID systems.

4.1.2. Protocol

Cumulative Matching Characteristic (CMC) curve and mean average precision (mAP) are commonly used metrics for evaluating person re-ID methods. The CMC curve reflects retrieval precision, while the mAP reflects the recall. On CUHK03, we followed Li et al. [5] to repeat 20 times of random 1160/100 training/test splits and report the results under the single-shot evaluation setting. On Market-1501, the standard training/test split (750/751) was used.

4.2. Implementation Details

This work was implemented using Caffe [59], an open source deep learning framework. Original images were resized to 256×256 (then randomly cropped to 227×227 for AlexNet and 224×224 for Residual-50). As for InceptionNet, original images were resized to 160×64 (then randomly cropped to 144×56). All input images were mirrored randomly for data augmentation. Both AlextNet and Residual-50 were pre-trained on ImageNet dataset [60], while InceptionNet was directly trained from scratch (refer to [48]).

4.2.1. Training Base Networks

We adopted the mini-batch stochastic gradient descent (SGD) algorithm to update the network parameters. The batch size was set to 64 for AlexNet, 16 for Residual-50 and 100 for InceptionNet. The maximum number of training epochs was set to 50, 62, and 232 for AlexNet, Residual-50, and InceptionNet, respectively. AlexNet was trained with an initial learning rate of 0.001 and then reduced by 10 every 20 epochs. Residual-50 was trained with learning rate initialized at 0.001 and reduced by 10 every 25 epochs. For InceptionNet, the initial learning rate was set to 0.1 and was decreased by 4% for every four epochs until it reached 0.0005. The learning rate was then fixed at this value for a few more epochs until convergence.

4.2.2. Training DJL Network

Once the base network was pre-trained, we fine-tuned our Deep Joint Learning network. During training, the coordinates of body parts were transformed along with random image cropping and mirror operation. We set the position of invisible parts as zero. We empirically set the w/h of each body part region as 24/16 for InceptionNet (32/32 for AlexNet and Residual-50). When a body part was invisible, the features corresponding to its region were set to zero. The learning rate policy was changed to decay polynomially from 0.01 with the power parameter set to 0.5 and the whole network was trained for only around 20 epochs.

4.2.3. Testing

Given a pedestrian image of fixed size (227×227 for AlexNet, 224×224 for Residual-50, and 144×56 for InceptionNet), we extracted as features the FC7 activations for AlexNet, Pool5 activations for Residual-50, and FC7 activations for InceptionNet. We measured the similarity between two pedestrian images by the Euclidean distance between the L2-normalized features of them.

4.3. Performance Evaluation

We defined a simple version DJL network (DJL-S) which only contained one module in its local branch and compared it with the complete DJL network (DJL-HFS) with Hierarchical Feature Aggregation mechanism and Region Re-Weighting strategy. We adopted DJL-S structure with different base networks to validate the generalization ability of the proposed method and compared with the PIE method for the sake of fairness. We choose Conv4, Res4a and Inception_3a feature maps to generate the local features for AlexNet, Residual-50 and InceptionNet, respectively. Here, the output spatial extent of the RoI pooling layer was 1×1. To show the effectiveness of the Hierarchical Feature Aggregation as well as Region Re-Weighting strategy, further experiments were designed for InceptionNet based implementation with DJL-HFA structure.

4.3.1. Improvement over Base Networks

We first evaluated the proposed DJL-S network using various base networks on Market-1501 and CUHK03 benchmarks. The overall results are shown in Tables 1 and 2. The improvements over both AlexNet and Residual-50 base networks were significant. When using AlexNet, Rank-1 accuracy on Market-1501 rose from 57.75% to 67.64% and mAP rose from 33.80% to 43.60%. On CUHK03 dataset, Rank-1 accuracy rose by +18.92% for AlexNet. When using Residual-50, Rank-1 accuracy on CUHK03 arrived at 80.83%. On Market1501, consistent improvement could also be observed. Best performance appeared using InceptionNet [48], which obtained Rank-1 accuracy of 85.12% on Market-1501 and 84.25% on CUHK03. These results prove the effectiveness of our DJL-S network.

4.3.2. Comparison with The PIE Method

Our method shares a similar nature with the recent PIE [19] method, which learns pose invariant embedding from both well aligned PoseBox and original image. We compared our method with it

under the same experimental settings. Rank-1 accuracy improvement over base networks was used as the measurement criteria here. According to the results in Table 3, our observation was two-fold.

Table 1. Comparison with the three base networks, AlexNet, Residual-50 and InceptionNet on Market-1501 (by adopting the proposed DJL-S structure) in terms of identification accuracy (%) and mAP (%).

Method	Market-1501				
	Rank-1	**Rank-5**	**Rank-10**	**Rank-20**	**mAP**
AlexNet	57.75	77.52	84.47	89.46	33.80
Residual-50	72.42	86.49	91.03	94.42	48.01
InceptionNet	79.66	91.51	94.54	96.50	56.59
Proposed (AlexNet)	67.64	84.80	89.88	93.53	43.60
Proposed (Residual-50)	78.86	90.38	93.91	96.35	55.49
Proposed (InceptionNet)	**85.12**	**93.91**	**95.69**	**97.51**	**64.82**

Table 2. Comparison with the three base networks, AlexNet, Residual-50 and InceptionNet on CUHK03 (by adopting the proposed DJL-S structure) in terms of identification accuracy (%).

Method	CUHK03			
	Rank-1	**Rank-5**	**Rank-10**	**Rank-20**
AlexNet	53.03	79.53	87.82	94.21
Residual-50	61.79	85.46	92.31	97.86
InceptionNet	80.85	95.90	98.17	99.48
Proposed (AlexNet)	71.95	90.30	94.91	98.16
Proposed (Residual-50)	80.83	95.92	98.66	99.54
Proposed (InceptionNet)	**84.25**	**97.40**	**98.86**	**99.67**

Table 3. Rank-1 accuracy improvement (%) over base networks compared with the PIE method.

Base Network	Market-1501		CUHK03	
	DJL-S	**PIE**	**DJL-S**	**PIE**
AlexNet	**+9.89**	+9.12	**+18.92**	+2.65
Residual-50	**+6.44**	+5.66	**+19.04**	+5.50

First, for both base networks, DJL-S achieved better accuracy than PIE on both databases. This validated the superiority of our proposed local body part features as we did alignment at feature level instead of image level. As for PIE, image level alignment by affine transformation performed worse due to pose estimation errors. The higher accuracy achieved by our proposed method might be owing to two factors. For one thing, we pool body part features on the feature maps that are generated by the middle convolution layers in the base network. These layers have larger receptive fields and thus capture more context-aware information that can compensate misalignment errors of detected persons. For another, discriminative detail information can be learned through max-pooling operation inside local body part regions, which should be helpful to identify individuals with slight difference.

Second, we found that our method obtained significant improvement on CUHK03. We speculate that the higher image resolution in CUHK03 benefited the learned features. We discuss this in detail in Section 4.3.4.

4.3.3. Comparison with More State-of-The-Arts

We compared our DJL with the current state-of-the-art DL-based methods. For ease of comparison, those methods are summarised into two categories: Pose-irrelevant DL-based methods and Pose-relevant DL-based methods. Their results on Market-1501 and CUHK03 are shown in Tables 4 and 5. The proposed DJL-S structure achieved comparable Rank-1 accuracy among the

methods, i.e., 85.12% and 84.25% on Market-1501 and CUHK03, respectively. When adopting DJL-HFS structure and combining other re-ranking method (RK) [41], the performance was further boosted, reaching 88.39% on Market-1501. Furthermore, our Deep Joint Learning pipeline can be easily integrated with other state-of-the-art person re-ID networks.

Table 4. Comparison with state-of-the-arts on Market-1501. Rank-1 accuracy (%) and mAP (%) are shown. The best result is marked in bold while the second best in gray.

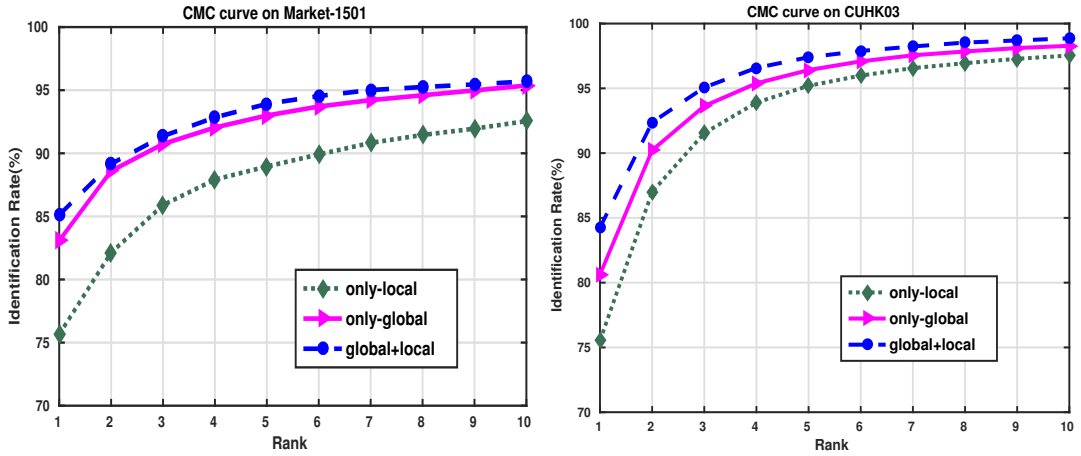
Methods	**Rank-1**	**mAP**
Pose-irrelevant DL-based Methods		
APR [38]	84.29	64.67
DLCE [49]	79.51	59.87
DML [42]	87.73	68.83
Gate-SCNN [7]	65.88	39.55
JLML [51]	85.10	65.50
X-Corr [61]	-	-
Ours		
DJL-S	85.12	64.82
DJL-HFA	85.99	65.65
DJL-HFA(RK)	88.39	**79.97**
Pose-relevant DL-based Methods		
DLPA [53]	81.0	63.4
MSCAN [62]	80.31	57.53
PABP [23]	**88.8**	74.5
PDC [52]	84.14	63.41
PIE [19]	78.65	53.87
PIE + KISSME [19]	79.33	55.95
Spindle [22]	76.90	-

Table 5. Comparison with state-of-the-arts on CUHK03. Rank-1 accuracy (%) is shown. The best result is marked in bold while the second best in gray.

Methods	**Rank-1**
Pose-irrelevant DL-based Methods	
APR [38]	-
DLCE [49]	83.4
DML [42]	-
Gate-SCNN [7]	68.10
JLML [51]	80.60
X-Corr [61]	72.04
Ours	
DJL-S	84.25
DJL-HFA	85.90
DJL-HFA(RK)	85.12
Pose-relevant DL-based Methods	
DLPA [53]	81.6
MSCAN [62]	67.99
PABP [23]	**88.0**
PDC [52]	78.29
PIE [19]	62.40
PIE + KISSME [19]	67.10
Spindle [22]	-

4.3.4. Further Analysis and Discussion

- **Body part segmentation**
  To evaluate the impact of body part segmentation errors on our method, we randomly disturbed the position of each body part during training. Here, we adopted two settings: small disturbance (Disturb-small) and violent disturbance (Disturb-violent). We translated the coordinates of each body part up to 6% of input image size for small disturbance and 30% for violent disturbance. Tables 6 and 7 show the results of DJL-S on Market-1501 and CUHK03, respectively. Generally, accuracy changed a little under slight disturbances (from 67.64% to 68.82% for AlexNet on Market-1501) while varied dramatically under large disturbances (still better than base networks). This demonstrates that our proposed method can effectively cope with human body misalignment.
- **Low resolution**
  We evaluated the impact of image resolution on our method. Experiments were conducted on CUHK03. We down-sampled all images in CUHK03 to half of their original size and used those low resolution images for training and testing. The results in Table 7 show that low image resolution degrades the performance of DJL-S.
- **RoI pooling effects at different layers**
  An important part of our method is to apply the RoI pooling operation to different middle layers. In Tables 8 and 9, we systematically explore the identification performance of different middle convolution by performing RoI pooling on each of them. We experimented with various network structures (AlexNet, Residual-50 and InceptionNet) and found that pooling at relative deeper layer obtains better performance improvements over the base networks. This observation shows that deeper, semantic CNN features contribute more to person re-ID task.
- **Hierarchical Feature Aggregation and Region Re-Weighting**
  We evaluated the effects of Hierarchical feature aggregation and Region Re-Weighting using the base Inception network with different variants of DJL: DJL-S, DJL-S + RRW, DJL-HFA(w/o RRW), and DJL-HFA. DJL-S denotes pooling body part features from a single convolution layer (Inception_3a). DJL-S + RRW further combines RRW strategy with DJL-S. DJL-HFA (w/o RRW) means DJL-HFA without applying RRW strategy, and DJL-HFA is the full version of our proposed method. As depicted in Tables 10 and 11, the DJL-S + RRW achieves performance gain in Rank-1 accuracy compared with the DJL-S network on both Market-1501 and CUHK03 datasets. When adopting DJL-HFA(w/o RRW), the Rank-1 accuracy improved on CUHK03 dataset while dropped slightly on Market-1501 dataset. We believe the performance drop is due to the inconsistent scale and norm of multiple layers (the "larger" features would dominate the "smaller" ones) [28]. As Region Re-Weighting would automatically learn the scale of features during training procedure, we speculate that integrating RRW with HFA would achieve more performance gain in Rank-1 accuracy. The results in Tables 10 and 11 also demonstrate this: the Rank-1 accuracy arrived at 85.99/85.90 on Market-1501/CUHK03 when using DJL-HFA. Furthermore, we give some illustrations about the learned weight parameters in Table 12, which show the scale and importance differences across multiple layers regions.
- **Complementary effects**
  We evaluated the effects of individual local feature (FC_local), global feature (FC7) as well as their combination on Market-1501 and CUHK03. The results on the two databases are shown in Figure 7. These results demonstrate that, although global and local feature representations alone are competitive for re-ID, further performance gain can be obtained by combining them using our proposed method. This proves that our proposed method can effectively explore the complementary discriminative information in global and local features for more accurate person re-ID. Two example results are shown in Figure 8. As can be seen, even when the probe and gallery pedestrian images have obviously different poses (i.e., they are not well aligned), our proposed method can still correctly retrieve the corresponding gallery images among the first ten ranks.



(a) Market-1501 (b) CUHK03

Figure 7. CMC curves on Market-1501 and CUHK03 when using local, global and hybrid features (global+local) extracted by DJL-S (based on InceptionNet).

Figure 8. Example person re-ID results by using the base Residual-50 network (Res-50) and the proposed DJL network on Market-1501 database. Correct retrievals are surrounded with green bounding boxes while wrong retrievals are surrounded with red bounding boxes.

Table 6. Identification accuracy (%) and mAP (%) of the proposed method with different base networks on Market-1501 when different disturbances are applied to the segmented body parts. The best results under different settings are marked in bold.

Base Network	Setting	Market-1501				
		Rank-1	**Rank-5**	**Rank-10**	**Rank-20**	**mAP**
AlexNet	Base	57.75	77.52	84.47	89.46	33.80
	Proposed	67.64	84.80	**89.88**	**93.53**	43.60
	Disturb-small	**68.82**	**84.95**	89.31	93.50	**44.89**
	Disturb-violent	64.79	82.21	88.15	92.22	40.84
Residual-50	Base	72.42	86.49	91.03	94.42	48.01
	Proposed	**78.86**	**90.38**	**93.91**	**96.35**	**55.49**
	Disturb-small	77.76	89.88	92.96	96.02	54.62
	Disturb-violent	75.95	88.60	92.37	95.19	52.71
InceptionNet	Base	79.66	91.51	94.54	96.50	56.59
	Proposed	**85.12**	**93.91**	95.69	97.51	64.82
	Disturb-small	84.53	93.79	95.93	97.54	**64.89**
	Disturb-violent	83.61	93.65	**95.99**	**97.60**	63.44

Table 7. Identification accuracy (%) of the proposed method with different base networks on CUHK03 when different disturbances were applied to the segmented body parts and when low resolution images were used. The best results under different settings are marked in bold.

Base Network	Setting	CUHK03			
		Rank-1	**Rank-5**	**Rank-10**	**Rank-20**
AlexNet	Base	53.03	79.53	87.82	94.21
	proposed	**71.95**	**90.30**	**94.91**	**98.16**
	Disturb-small	68.31	89.19	93.86	97.07
	Disturb-violent	62.07	84.84	91.49	96.03
	Low-resolution	60.35	83.71	90.59	95.49
Residual-50	Base	61.79	85.46	92.31	97.86
	proposed	**80.83**	95.92	98.66	99.54
	Disturb-small	80.53	**96.45**	**99.01**	**99.71**
	Disturb-violent	73.40	93.23	96.75	99.25
	Low-resolution	75.58	93.68	97.28	9 9.15
InceptionNet	Base	80.85	95.90	98.17	99.48
	proposed	**84.25**	97.40	**98.86**	99.67
	Disturb-small	83.38	**97.49**	98.81	99.52
	Disturb-violent	82.41	97.42	98.84	**99.69**
	Low-resolution	82.80	97.12	98.64	99.63

Table 8. Identification accuracy (%) and mAP (%) of the proposed method with performing RoI pooling at different middle layers on Market-1501. The best result over various pooling layers is marked in bold.

Base Network	Pooling Layer	Market-1501				
		Rank-1	Rank-5	Rank-10	Rank-20	mAP
AlexNet	Conv3	66.30	84.29	89.58	93.23	42.75
	Conv4	67.64	84.80	89.88	93.53	43.60
	Conv5	**69.83**	**85.66**	**90.65**	**94.00**	**45.17**
Residual-50	Res3a	77.02	89.88	93.29	95.87	54.17
	Res4a	78.86	90.38	93.91	**96.35**	55.49
	Res5a	**79.48**	**91.30**	**94.51**	96.20	**57.53**
InceptionNet	Inception_1a	82.90	92.99	95.16	97.00	61.24
	Inception_2a	84.59	**94.83**	**96.53**	**98.19**	**65.89**
	Inception_3a	**85.12**	93.91	95.69	97.51	64.82

Table 9. Identification accuracy (%) of the proposed method with performing RoI pooling at different middle layers on CUHK03. The best result over various pooling layers is marked in bold.

Base Network	Pooling Layer	CUHK03			
		Rank-1	Rank-5	Rank-10	Rank-20
AlexNet	Conv3	68.13	89.12	94.84	97.94
	Conv4	71.95	90.30	94.91	98.16
	Conv5	**74.22**	**91.25**	**95.37**	**98.64**
Residual-50	Res3a	77.40	94.63	98.44	99.57
	Res4a	80.83	95.92	98.66	99.54
	Res5a	**83.97**	**96.97**	**98.67**	**99.61**
InceptionNet	Inception_1a	82.66	96.66	98.42	99.33
	Inception_2a	83.01	96.98	98.75	99.53
	nception_3a	**84.25**	**97.40**	**98.86**	**99.67**

Table 10. Effects of Region Re-Weighting and Hierarchical Feature Aggregation using the base Inception network on Market-1501. Identification accuracy (%) and mAP (%) are reported. The best Rank-1 result is marked in bold.

Methods	Market-1501				
	Rank-1	Rank-5	Rank-10	Rank-20	mAP
DJL-S	85.12	93.91	95.69	97.51	64.82
DJL-S + RRW	85.21	93.74	95.78	97.60	65.38
DJL-HFA (**w/o RRW)**	84.95	94.15	96.38	97.74	65.29
DJL-HFA	**85.99**	94.15	96.29	97.77	65.65

Table 11. Effects of Region Re-Weighting and Hierarchical Feature Aggregation using the base Inception network on CUHK03. Identification accuracy (%) is reported. The best Rank-1 result is marked in bold.

Methods	CUHK03			
	Rank-1	Rank-5	Rank-10	Rank-20
DJL-S	84.25	97.40	98.86	99.67
DJL-S + RRW	84.29	97.15	98.80	99.67
DJL-HFA (**w/o RRW)**	84.72	97.17	98.40	99.34
DJL-HFA	**85.90**	97.79	98.90	99.40

Table 12. The learned weights of ten body parts at different pooling layers. The initial weight parameter of each body part region was set to 10.

Body Parts	Pooling Layer		
	Inception_1a	Inception_2a	Inception_3a
Rshoulder(w0)	8.30	7.21	5.56
Lshoulder(w1)	8.35	7.20	5.39
RElbow(w2)	8.96	8.97	6.60
LElbow(w3)	9.97	8.49	6.56
RHip(w4)	8.68	7.24	5.21
LHip(w5)	8.66	7.54	5.76
Rknee(w6)	9.99	7.67	6.07
Lknee(w7)	8.70	7.81	5.46
RAnkle(w8)	10.09	9.24	8.82
LAnkle(w9)	9.92	9.86	8.56

5. Conclusions

This paper proposes a Deep Joint Learning (DJL) network to learn better feature representation from both entire image and local body parts. The local features are pooled from the feature maps generated by the convolution layers, which capture the salient details and are robust to handle pedestrian misalignment. Hierarchical Feature Aggregation mechanism and Region Re-Weighting strategy effectively improve our feature representation by optimally aggregating body parts features from low-level to semantic-level. Multiple Softmax losses are used to integrate the discriminative power of global and local features. Extensive evaluations on Market1501 and CUHK03 benchmarks validated the advantages of the proposed DJL network.

Author Contributions: Conceptualization, Y.L.; Methodology, Y.L.; Supervision, H.Y. and Q.Z.; Validation, Y.L.; Writing—original draft, Y.L.; and Writing—review and editing, H.Y. and Q.Z.

Acknowledgments: This work is supported by the National Key Research and Development Program of China (No. 2017YFB0802300), and the Miaozi Key Project in Science and Technology Innovation Program of Sichuan Province, China (No. 2017RZ0016) and the Shenzhen Fundamental Research fund (JCYJ20180305125822769).

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

References

1. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person re-identification: Past, present and future. *arXiv* **2016**, arXiv:1610.02984.
2. Gong, S.; Cristani, M.; Yan, S.; Loy, C.C. *Person Re-Identification*; Springer: Berlin, Germany, 2014.
3. Bedagkar-Gala, A.; Shah, S.K. A survey of approaches and trends in person re-identification. *Image Vis. Comput.* **2014**, *32*, 270–286. [CrossRef]
4. 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.
5. 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, Columbus, OH, USA, 23–28 June 2014; pp. 152–159.
6. 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, 27–30 June 2016; pp. 1335–1344.

7. 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; pp. 791–808.
8. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Deep metric learning for person re-identification. In Proceedings of the International Conference on Pattern Recognition, Stockholm, Swede, 24–28 August 2014; pp. 34–39.
9. 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, Las Condes, Chile, 11–18 December 2015; pp. 1116–1124.
10. Corvee, E.; Bremond, F.; Thonnat, M. Person re-identification using spatial covariance regions of human body parts. In Proceedings of the 2010 IEEE 7th International Conference on Advanced Video and Signal Based Surveillance (AVSS), Boston, MA, USA, 29 August–1 September 2010; pp. 435–440.
11. Cheng, D.S.; Cristani, M. Person re-identification by articulated appearance matching. In *Person Re-Identification*; Springer: Berlin, Germany, 2014; pp. 139–160.
12. Cheng, D.S.; Cristani, M.; Stoppa, M.; Bazzani, L.; Murino, V. Custom pictorial structures for re-identification. In Proceedings of the 22nd British Machine Vision Conference, Dundee, UK, 29 August–2 September 2011; Volume 1, p. 6.
13. Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; Cristani, M. Person re-identification by symmetry-driven accumulation of local features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2360–2367.
14. Plagemann, C.; Ganapathi, V.; Koller, D.; Thrun, S. Real-time identification and localization of body parts from depth images. In Proceedings of the 2010 IEEE International Conference on Robotics and Automation, Anchorage, AK, USA, 3–8 May 2010; pp. 3108–3113.
15. Mousas, C.; Anagnostopoulos, C.N. Performance-driven hybrid full-body character control for navigation and interaction in virtual environments. *3D Res.* **2017**, *8*, 18. [CrossRef]
16. Wei, S.E.; Ramakrishna, V.; Kanade, T.; Sheikh, Y. Convolutional pose machines. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4724–4732.
17. 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, Honolulu, HI, USA, 21–26 June 2017; pp. 7291–7299.
18. Fang, H.; Xie, S.; Tai, Y.W.; Lu, C. Rmpe: Regional multi-person pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2334–2343.
19. Zheng, L.; Huang, Y.; Lu, H.; Yang, Y. Pose invariant embedding for deep person re-identification. *arXiv* **2017**, arXiv:1701.07732.
20. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, USA, 7–12 December 2015; pp. 2017–2025.
21. Liu, Y.; Wu, Z.; Zhao, Q. Pooling body parts on feature maps for misalignment robust person re-identification. In Proceedings of the 2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA), Singapore, 11–12 January 2018; pp. 1–8. [CrossRef]
22. 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, Honolulu, HI, USA, 21–26 June 2017; pp. 1077–1085.
23. Suh, Y.; Wang, J.; Tang, S.; Mei, T.; Lee, K.M. Part-Aligned Bilinear Representations for Person Re-identification. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018, pp. 402–419.
24. Mousas, C.; Anagnostopoulos, C.N. Real-time performance-driven finger motion synthesis. *Comput. Graph.* **2017**, *65*, 1–11. [CrossRef]
25. Mousas, C. Full-body locomotion reconstruction of virtual characters using a single inertial measurement unit. *Sensors* **2017**, *17*, 2589. [CrossRef] [PubMed]
26. Fei-Fei, L.; Perona, P. A bayesian hierarchical model for learning natural scene categories. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 524–531.

27. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1915–1929. [CrossRef] [PubMed]
28. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
29. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person re-identification by Local Maximal Occurrence representation and metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
30. Xiong, F.; Gou, M.; Camps, O.; Sznaier, M. Person re-identification using kernel-based metric learning methods. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 1–16.
31. Zhao, R.; Ouyang, W.; Wang, X. Person re-identification by salience matching. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2528–2535.
32. Yang, Y.; Yang, J.; Yan, J.; Liao, S.; Yi, D.; Li, S.Z. Salient color names for person re-identification. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 536–551.
33. Koestinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Large scale metric learning from equivalence constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2288–2295.
34. Li, W.; Wang, X. Locally aligned feature transforms across views. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3594–3601.
35. Liao, S.; Zhao, G.; Kellokumpu, V.; Pietikäinen, M.; Li, S.Z. Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1301–1306.
36. Liao, S.; Li, S.Z. Efficient psd constrained asymmetric metric learning for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015; pp. 3685–3693.
37. Jose, C.; Fleuret, F. Scalable metric learning via weighted approximate rank component analysis. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 875–890.
38. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Yang, Y. Improving person re-identification by attribute and identity learning. *arXiv* **2017**, arXiv:1703.07220.
39. Shi, Z.; Hospedales, T.M.; Xiang, T. Transferring a semantic representation for person re-identification and search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4184–4193.
40. Peng, P.; Xiang, T.; Wang, Y.; Pontil, M.; Gong, S.; Huang, T.; Tian, Y. Unsupervised cross-dataset transfer learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1306–1315.
41. 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, Honolulu, HI, USA, 21–26 June 2017; pp. 3652–3661.
42. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep Mutual Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4320–4328.
43. Fan, H.; Zheng, L.; Yan, C.; Yang, Y. Unsupervised Person Re-identification: Clustering and Fine-tuning. *ACM Trans. Multimed. Comput. Commun. Appl.* **2018**, *14*, 83:1–83:18. [CrossRef]
44. Wu, Y.; Lin, Y.; Dong, X.; Yan, Y.; Ouyang, W.; Yang, Y. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5177–5186.
45. Navon, D. Forest before trees: The precedence of global features in visual perception. *Cogn. Psychol.* **1977**, *9*, 353–383. [CrossRef]
46. Torralba, A.; Oliva, A.; Castelhano, M.S.; Henderson, J.M. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychol. Rev.* **2006**, *113*, 766–786. [CrossRef] [PubMed]

47. 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, Honolulu, HI, USA, 21–26 June 2017; pp. 403–412.
48. Xiao, T.; Li, H.; Ouyang, W.; Wang, X. Learning deep feature representations with domain guided dropout for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1249–1258.
49. Zheng, Z.; Zheng, L.; Yang, Y. A Discriminatively Learned CNN Embedding for Person Reidentification. *ACM Trans. Multimed. Comput. Commun. Appl.* **2017**, *14*, 13:1–13:20. [CrossRef]
50. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 1097–1105.
51. Li, W.; Zhu, X.; Gong, S. Person re-identification by deep joint learning of multi-loss classification. *arXiv* **2017**, arXiv:1705.04724.
52. Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-driven deep convolutional model for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3960–3969.
53. Zhao, L.; Li, X.; Zhuang, Y.; Wang, J. Deeply-Learned Part-Aligned Representations for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 3239–3248.
54. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
55. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Las Condes, Chile, 11–18 December 2015; pp. 1440–1448.
56. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 346–361.
57. Hariharan, B.; Arbeláez, P.; Girshick, R.; Malik, J. Hypercolumns for object segmentation and fine-grained localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 447–456.
58. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
59. Jia, Y.; Shelhamer, E.; Donahue, J.; Karayev, S.; Long, J.; Girshick, R.; Guadarrama, S.; Darrell, T. Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 675–678.
60. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
61. Subramaniam, A.; Chatterjee, M.; Mittal, A. Deep neural networks with inexact matching for person re-identification. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 2667–2675.
62. Li, D.; Chen, X.; Zhang, Z.; Huang, K. Learning deep context-aware features over body and latent parts for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 384–393.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).