symmetry

MDPI

Article

An Improved Deep Mutual-Attention Learning Model for Person Re-Identification

Miftah Bedru Jamal [1], Jiang Zhengang [1] and Fang Ming [1,2,*]

[1] School of Computer Science, Changchun University of Science and Technology, Changchun 130022, China; miftahbedru@gmail.com or ieec2016174@mails.cust.edu.cn (M.B.J.); jiangzhengang@cust.edu.cn(J.Z.)
[2] School of Artificial Intelligence, Changchun University of Science and Technology, Changchun 130022, China
* Correspondence: fangming@cust.edu.cn

Received: 26 January 2020; Accepted: 17 February 2020; Published: 2 March 2020

Abstract: Person re-identification is the task of matching pedestrian images across a network of non-overlapping camera views. It poses aggregated challenges resulted from random human pose, clutter from the background, illumination variations, and other factors. There has been a vast number of studies in recent years with promising success. However, key challenges have not been adequately addressed and continue to result in sub-optimal performance. Attention-based person re-identification gains more popularity in identifying discriminatory features from person images. Its potential in terms of extracting features common to a pair of person images across the feature extraction pipeline has not been be fully exploited. In this paper, we propose a novel attention-based Siamese network driven by a mutual-attention module decomposed into spatial and channel components. The proposed mutual-attention module not only leads feature extraction to the discriminative part of individual images, but also fuses mutual features symmetrically across pairs of person images to get informative regions common to both input images. Our model simultaneously learns feature embedding for discriminative cues and the similarity measure. The proposed model is optimized with multi-task loss, namely classification and verification loss. It is further optimized by a learnable mutual-attention module to facilitate an efficient and adaptive learning. The proposed model is thoroughly evaluated on extensively used large-scale datasets, Market-1501 and Duke-MTMC-ReID. Our experimental results show competitive results with the state-of-the-art works and the effectiveness of the mutual-attention module.

Keywords: person re-identification; mutual-attention; classification; verification

1. Introduction

Person re-identification task aims at making a correspondence between pedestrian images across non-overlapping camera views captured at different times. It is a key task for surveillance systems and applications involving human-computer interaction. It draws increasing attention in video surveillance due to its greater role in applications like suspicious threat detection, person retrieval, and multi-camera tracking. It saves a great deal of human labor on exhaustively searching for a target person in a large crowd of people. A full-fledged person re-identification can be used in tracing crime perpetrators and missing persons in a crowd with a target person fed to the system. Given an input image, called the query image from one camera, the goal is to retrieve an image or sets of images from a different camera, called the gallery set, based on the similarity to the query image. It became a problem of great interest bearing great challenges due to the fact that the appearance of a person keeps on changing across different camera views due to the aggregated effect of variations resulting from the change in light, pose, occlusion, view point, and even in some cases a pedestrian undergoing an instantaneous change such as a change of clothing, carrying bag, or putting on a cap. Some variations/situations not

only pose a challenge in re-identifying, but can lead to misleading conclusions by the re-identification system when two images of the same person look different (inter-class difference) or of different persons looking the same (intra-class difference). Many classic methods approached these challenges by using discriminative feature representation and the similarity measure as a method to judge the degree of similarity among a pair of images. The expectation for a well performing re-identification system is to return the best match for a given query image from the gallery set in the form of the degree of similarity as quantified by the highest probability value for a gallery that has the same identity as the query after the gallery set is sorted based on the similarity to the current query image. Deep learning approaches achieve tremendous success in computer vision tasks including person re-identification. Deep learning methods, particularly that of that of the deep Convolutional Neural Network (CNN), have been the most competitive approaches applied to robust feature extraction with great success in areas of computer vision such as object recognition [1,2], image and video classification [3–5], image forgery identification [6], speech recognition [7], and semantic segmentation tasks [8], to mention a few. Features learned by deep networks are automatically learned and are not easily understood even by a human observer, nor can they be crafted and interpreted the way traditional features are designed in a fixed and recurring deterministic algorithm. However, as deep learning models have learnable features, they can be guided to attend to and learn features from more visually distinct and informative regions within the feature maps of a given image. Such visual attention can be exploited to detect regions of an image that are potentially important to better train the deep network. From this point of view, learning the attention map is as essential as learning the features themselves to focus on regions of interest with relevant features as demanded by the target. The attention map, being produced from learned features, tends to give silent regions a stronger response. Numerous deep neural network based on the attention scheme have been studied in an effort to solve a wide range of visual representation in tasks like scene generatio [9], fine-grain recognition [10], and image captioning [11]. Their usage ranges from learning discriminative regions of a given image to classifying different objects and tracking objects by merging multiple attention types to leverage their joint potential. Likewise, attention-based methods have recently been used for person re-identification problems [12–14] with the intention of searching for the most responsive regions that play a role in identifying pedestrians. However, many of these methods naively feed-forward pedestrian input images in a sequence of layers to extract features without leveraging the feature dependency across input pairs to emphasize correlated regions with high responses. For person re-identification, it was studied in some literature [12–15] that feature representation that considers the interaction among a pair of images has been shown to be effective. Given two images, for an attention map to learn the relevant features, it needs to consider features that are common to both images. In this regard, an independent attention map without considering cross-image attention may lead to sub-optimal performance in similarity measurement. The attention method that can encode mutually discriminative regions and rescale the features accordingly for efficient similarity computation is a natural choice. Some methods [16] proposed the attention scheme that inferred the spatial attention from a single image to enhance the learned feature. However, the feature learned and scene detected in the deep convolutional neural network possess spatial and channel components [17]. In order to learn robust attention, decomposing attention channel and spatial-wise and exploiting their relative merits are crucial to learn attention distributed along spatial and channel dimensions and make comparable matching between pairs of input. This is very crucial for learning a mutually robust feature. Motivated by these intuitions, we propose a mutual-attention-based deep learning model that not only computes the attention map for individual input images, but also uses an additional joint attention, called mutual-attention, to capture cross-input attention. Our proposed novel mutual-attention map is learned from the self-attention map computed from individual pairs of input images. Further, we follow a scheme that decomposes attention from the feature map into spatial and channel components in order to lead the model to learn features that are relevant and discriminative with respect to each input image across their channel and spatial dimension. The mutual-attention map re-scales feature maps inferred from a pair of

inputs, and the original feature maps are added back to consolidate the feature learned across layers. As person re-identification is positioned as a person retrieval and ranking task, we design the model as a classification and verification task to learn subtle intra-class variation and misleading inter-class variation and exploit the complimentary advantage of the two tasks. The final person re-ID model is based on an end-to-end trainable deep neural network with verification and classification loss. Training results show that the model is capable of boosting co-dependent features across input pairs and achieves competitive performance over widely used benchmark datasets. The rest of this paper is organized as follows. Section 2 presents the related works. The proposed method is discussed in Section 3. The loss function and similarity learning are detailed in Section 4. The experimental setting and the discussion on the result are elaborated in Sections 5 and 6, respectively. Section 7 puts forth the conclusions and future directions.

2. Related Works

A number of classical and deep learning models have been proposed in the literature to solve the person re-identification problem. In this section, we briefly present the literature of some of the key works. Many of the existing methods for person re-identification focus on two key tasks: learning robust feature extraction and learning the distance metric. Feature extraction from raw input falls into one of two broad categories, namely hand-crafted and deep neural network learned features. There is a good number of person re-identification works that are based on hand-crafted features and some statistical method as the feature descriptor for the region of images. The most commonly used feature descriptor for pedestrians includes color, texture, and histogram of color. Hand-crafted-based methods are one of the classic approaches that handle feature extraction and the similarity measure as independent tasks that do not complement one another. Some work attempted to exploit the merits of handcrafted and deep neural network features to gain complementary advantages. The work in [18] is a typical example of such a trend, where the entire convolutional network is jointly trained with the feature extracted by CNN and hand-crafted methods. Such a fusion serves as a way to constrain and regularize the feature extracted by CNN and compliment it with the hand-crafted feature. Although handcrafted features showed notable success in the past, they remained the same in recent years due to the huge success of deep learning methods. Deep learning led to a series of breakthroughs in many computer vision and other tasks. They were capable of naturally and automatically integrating low-, mid-, and high-level features from raw images in multiple layers with adjustable weights that could be learned through a process known as back propagation. A comprehensive survey and its detail can be found in [19–21]. They have increasingly obtained popularity in many areas such as classification tasks for image [3,4] and video classification [5], translation [22], captioning [23,24] and description generation for images [25], image recognition [1,2] and detection [26,27], and semantic image segmentation [8]. Deep neural network architectures, particularly the deep Convolutional Neural Network (CNN), have shown impressive success in many image classification tasks. Due to the close relationship between classification and person re-ID, such success was easily extended to the re-ID task in the works that followed in the subsequent years, and CNN has become a major building unit of virtually all state-of-the-art re-ID models in the literature. The following sections present related works for different methods.

2.1. Siamese-Based CNN Model

The Siamese-based model consists of two or more branches of CNN layers intended to learn features and the similarity measure in parallel. The Siamese neural network was originally proposed as a signature verification method in [28] where two sub-networks compared the distance between previously stored signature feature vectors and a signature feature vector currently to be verified. Some of the person re-ID work worthwhile to mention in this regard includes [29,30]. They proposed a Siamese model to judge if a pair of images belongs to the same identity Siamese mode with CNN as its core component, which has gained more popularity, and it was the choice of many researchers following

the works in [29,30]. The architecture of Siamese model makes it a natural choice for person re-ID due to the fact that person re-ID mostly involves pairwise comparison of images and the number of training samples for each identity is mostly limited (usually two). Unlike the classification/identification-based model [31], the Siamese model does not make full use of labels for the input. It rather uses a weak label assigned to a pair of images indicating if the pair of images belongs to the same or a different identity. Although the Siamese model commonly contains two branches taking pairs of input images, there is research work that considered multiple branches like a triplet and a quadruplet network. The work in [32] proposed a unified multi-scale triplet convolutional neural network consisting of triplet training input, one deep, and two shallow tied layers that were optimized with the comparative similarity loss function called the L2-norm. Similarity loss on the image triplet is aimed to give a higher score to the pair of images from the same person than those from different persons. However, the triplet network still does not adequately address the generalization ability and may not fully address the problem of inter- and intra-class variation when the aim is to have smaller inter-class variation and larger intra-class variation. To this end, the authors in [33] proposed a quadruplet network. However, forming a quadruplet training sample for such a network can be quite overwhelming; the authors proposed a threshold scheme to select positive and negative samples and minimize the effort in generating training samples. The Siamese the model can benefit from hand-crafted and deeply learned features that can complement each other. A Siamese-based deep neural network that co-learns color, texture, and the metric for image similarity was proposed in [34]. The author used two branches of a symmetric three layer sub-network that shared similar structures in the corresponding layers and joined them by a bounded cosine function.

2.2. Patch-Based Method

Some works split an input image into several strips and extract local features from each strip. In [30], the input image was split into three adjacent non-overlapping strips, and each part was fed into two independent convolutional layers and a fully connected layer that merged them to produce vectors of the images. The co-sine function was used to determine the degree of similarity. Such an approach, however, cannot precisely address spatial misalignment. The work in [35] aimed to address this issue by exploiting the shortest path algorithm between sets of local features between two images. The architecture proposed in [36] followed a similar model, but was different in that the horizontal patch-wise matching was included wherein the feature response across each patch was multiplied by every other patch sample from the horizontal strip in other image feature map. This work was subsequently improved in [37], where the author used the neighborhood difference between pixel values of the feature of one image and the neighborhood location of another image feature. The author asserted that such an approach added robustness to the positional difference in corresponding features of the two input images. The work in [38] proposed very similar work and extended the work in [37] by computing the pixel value difference around the neighborhood followed by the computation of the correlation between feature vectors. Further, they used a wider search region along with an inexact matching technique to overcome the challenges such as pose change. The authors argued that a wider search and inexact match were crucial to overcoming the challenges such as pose caused by viewpoint variation, illumination change, and partial occlusions.

2.3. Local/Global Feature and Scale Learning Methods

In [39], the importance of the scale and spatial importance of features was studied to determine which scale (global and local regions) was proper for matching. A similar work in [40] proposed a model with a global and local feature selection model constrained by the same class label and used the cross-entropy classification loss function, which avoided the need to form image pairs for training and increase the scalability of the model. The Deep pyramid feature Learning (DPFL) model proposed in [41] asserted that the scale specific feature and multi-scale feature overcame appearance variation across images taken at different scale. The model overcame problems such as the cross-scale feature

learning discrepancy challenge by being guided by inter-level feature interaction and multi-scale complementary feature selection. In [42], a similar work was proposed wherein two streams of the convolutional neural network with a weighted objective were able to learn temporal traits like gait and spatial information to form discriminative features. Human body parts play key role in forming correspondence among the pair of image. The re-ID model that focuses on extracting features from human body parts is one of the potential approaches expected to make the model resistant to pose and changes in the spatial distribution of body parts across images views. In [43], a deep CNN model used the part-aligned weighted feature extraction of regions in the image. Body part extraction and representation were modeled as a joint operation. The features extracted were concatenated, aggregated, and normalized before the overall score, and triplet based loss was computed. The work in [11] proposed a CNN-based model having a sub-network that was capable of detecting, rotating, normalizing, and relocating human body parts with affine transformation to a reasonable region. The work in [39] proposed a triplet lose branch model based on utilizing both ranking and classification and used cross-domain knowledge transfer by training the model with a bigger dataset and testing it with a smaller dataset.

2.4. Attention-Based Methods

Attention-based re-identification has gained growing popularity in recent years. In [44], the paper proposed a network called the harmonious attention convolutional neural network with feature representation and an attention mechanism focusing on soft pixel and hard regions. The author in [17] proposed a multi-branch attention framework to identify discriminative whole body and local parts of person images by leveraging an encoder-decoder style network. They used an inter- and intra-attention map inferred from local body parts and inter-attention inferred from global image. However, their framework required extensive training for each local and global intra-attention branch. The model may inherit error on detection as it employed joint detection-based body part estimation. The work in [15] proposed a CNN-based mode that used a differentiable gate function that served as a smooth switch and attention to select the extracted pattern in the feature map based on the Euclidean distance computed along the dimension of summarized features. In [45], the authors proposed a model based on a human semantic parsing scheme to use local cues belonging to different body parts. They used these cues for human body parts and performed the element-wise product with the feature pulled from the global image to pay attention to body regions. In [46], the authors proposed a co-attention model to learn the relative representation of input pairs and used an iterative recurrence comparator to learn similarity. This work was similar to ours in that the model learned features from input pairs and concurrently detected the most distinct pattern from the pair of images, then fused them in similarity learning. However, the significance of attention for features in the spatial and channel dimension was not explicitly considered. Our proposed mutual-attention-based model leverages feature correspondence across input pairs by fusing their respective self-attention map to boost feature points that have higher activation across the pairs and facilitate end-to-end learning with codependent features, as well as aides the subsequent similarity computation.

3. Proposed Method

This section presents the proposed architecture for person re-identification, which is comprised of the identification and verification branch for person re-ID aided by our novel cross-input (mutual-attention).

3.1. Model Architecture

Person re-ID models need to be robust to identity a probe person in the gallery set. It also needs to be robust in capturing pairwise differences among the pair of images. We follow this line and formulate person re-identification as a joint classification and verification problem. In the classification mode, the model leverages the label for the input and learns the feature in a supervised manner to

classify each input person to one of the unique classes. The verification mode aims to learn the feature embedding space by comparing the feature from the pair of input person images, pull images of the same identity close, and push the images of different identities further away in the embedding space. The verification task considers the limited relationship among the dataset, while the classification task does not explicitly consider the similarity measure between samples in the dataset. Hence, formulating the model from the two tasks enables the model to leverage the complimentary advantage of the two tasks and address the person re-ID objective.

We adopted the popular CNN ResNet-50 architecture up to Stage 4 as the backbone to extract the feature as it has shown good performance in a wide range of computer vision tasks. The base ResNet-50 was built with five blocks, each block having batch normalization, ReLU, and a downsampling layer that progressively reduces the feature map size by half form the previous block. The design of the ResNet-50 is shown in Figure 1. We used a Siamese model with two shared branches as the feature extraction pipeline. At Layer 4 and Layer 3, we computed the self-attention map for each branch based on the spatial and the channel attention from the feature map produced by the two layers. Following this, the cross-input mutual-attention was computed from the self-attention map of each branch. The mutual-attention map then re-scaled the feature map of each branch in such a way that feature activation at different spatial and channel locations across the two branches were re-emphasized and highlighted by the mutual-attention map. This served as a mechanism to boost the features that were mutually relevant across the two feature maps and improve the similarity computation for the verification part of the model. To consolidate the feature from the original feature map, we added the original feature to the newly computed feature map by our mutual-attention map. The design of the overall architecture is illustrated in Figure 2. All inputs were re-sized to 256×188. Given the pair of inputs, the proposed model simultaneously predicted the class of each input and computed the similarity score between the input pairs. The model had two ImageNet pre-trained CNN branches. The parameters were shared among the two branches to minimize the model complexity and training effort. The entire model was jointly optimized with the weighted loss from the classification and verification part of the model. In the following sections, we discuss the self-attention and mutual-attention map.

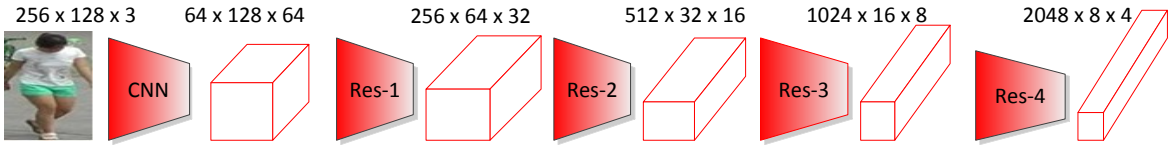

Figure 1. Basic ResNet-50 model with the corresponding dimensions across layers for a given input dimension. We use high-level feature maps from Layer 3 (Res-3) and Layer 4 (Res-4) to compute our mutual-attention layer.

3.2. Self- and Mutual-Attention

The spatial attention layer for feature maps across channels aimed to emphasize the spatial importance of each feature stacked across the depth of the feature map. The global average pooling appended at the penultimate layer of the model neglected the relative importance of spatial features irrespective of their location. Certain features, which might add robustness to the model, were compromised by the global average pooling operation. Moreover, features from non-corresponding spatial locations in the feature maps played some role in emphasizing discriminative patterns. In this regards, considering channel attention, which was computed from the spatial location of the same feature map, could complement the spatial attention in leveraging the importance of the feature across both the spatial and channel dimension.

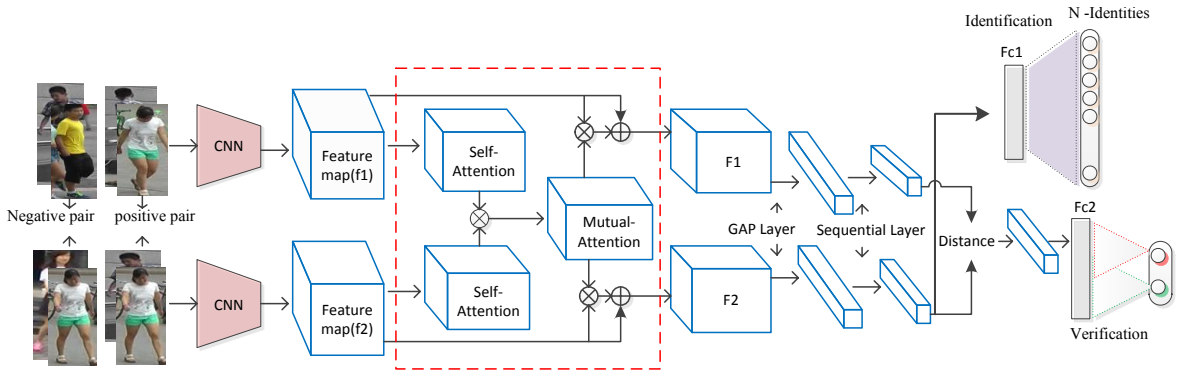

Figure 2. Architecture of the proposed deep mutual attention learning model. One mutual attention layer (shown with a red dotted line) is shown here for brevity, but the mutual-attention layer is inserted for both Layer 3 and Layer 4 of the model. CNN refers to ResNet-50 up to Layer 2. For each pair, the verification parts compute the similarity, and the identity (class) of each image is predicted by the identification part.(Best viewed in colour).

3.3. Self-Attention Map

A convolutional network that used *C* number of filters convolved though the input or feature map and yielded *H XW X C* feature maps where each feature map was presumed to detect a certain pattern across the spatial dimension. Hence, feature detection by the convolutional operation could be perceived as involving spatial and channel components. Following this intuition, we designed a novel mutual-attention map to infer mutually significant visual patterns from two input images or feature maps. The spatial significance of the spatial feature in the feature map was encoded to form the spatial attention map. Such a map took the form of a single-channel spatial map *H x W x1* where each location in the spatial map was the summary (mean) of all corresponding locations across all feature maps along the channel. Given feature map f of C channels, the value of the spatial attention $SPA_{(i,j)}$ was computed as the mean of all feature points at the corresponding spatial location across the channel C:

$$SPA_{(i,j)} = \frac{1}{C}\sum_{c=1}^{c} f_c(i,j) \tag{1}$$

where $f_{(i,j)}$ is the activation value of spatial point (i,j) on the channel c and SPA(i,j) represents the value of the spatial attention score at (i,j). These summarized spatial positions represented their aggregated significance through the depth of the feature maps and encoded the spatial relation among activation maps. During training, the network emphasized the spatial points with a higher activation value, and the proportional gradient flows through them during back propagation. While the spatial attention maps encoded the importance of the corresponding spatial point across the depth, the relevance of some non-corresponding locations within the same feature map or different feature maps across the channel direction might be overlooked in the process. Channel-wise features serve as the encoder of the feature for different semantic attributed generated by filters and are stacked as different feature maps. Hence, incorporating the channel-wise attention map along with the spatial attention map enabled the model to further infer co-related feature importance along the channel dimension of the pair of inputs. To incorporate this merit, the channel attention map, which was computed from a given feature map, was employed. Concretely, given feature map f of width W, height H, and channels C, average pooling was applied to each feature map f_i to get channel feature $Q_i \in \Re^{1x1xC}$:

$$Q_i = AP(f_i, W_I) \tag{2}$$

where AP is the average pooling and W is the parameter for average pooling. The channel-wise attention map $CA_i \in \Re^{1x1xC}$ was obtained by applying one convolutional layer.

$$CA_i = f(Q_i, W_i) \tag{3}$$

where f is the convolutional operation, C_i is the feature map at the i^{th} channel, and W is the parameter of the convolutional operator. The semantic attributes aggregated at different features were exclusively exploited by the channel-wise attention map besides the spatial attention map. The final Self-Attention map (SA) was obtained by combining spatial and channel attention. The spatial and channel-wise attentions were combined by multiplying them and passing the result through a 1×1 convolutional operation:

$$SA = f((SPA * CA_k), W)) \tag{4}$$

where f is the convolutional operation and W the corresponding parameters. The self-attention map computed from the spatial and channel attention fairly encoded the importance of different activations at the same spatial locations and also benefited from activation from different activations along the channel. With the learned weight being proportional to the mean of activations along the channel and across the spatial dimension, the spatial and channel relationships were captured, and they were used to compute the mutual-attention between two input images in a versatile way. The following section presents mutual attention map.

3.4. Mutual Attention Layer

When comparing visual similarity or variation for the learned feature map of a pair of input images, discriminative and salient patterns needed to have more attention paid to them. While the visually discriminative part in each input was relevant, patterns that were common across the two feature maps played an important role in giving a better comparison of the learned features. Inspired by this intuition, we modeled a mutual-attention map with the goal of making an effective local feature similarity comparison from visually richer higher level features. The model weighed common and co-dependent local patterns based on the similarity score and loss computed in the subsequent soft-max layer. This way, the inferred mutual pattern enabled the lower layers to learn filters that could distinguish the local pattern of positive pairs from negative pairs through the gradient that flowed during back propagation. The mutual-attention was computed from the self-attention map of the pair of feature map as follows:

$$MA_{(i,j)} = SA_1(i,j) * SA_2(i,j) \tag{5}$$

where MA is the mutual attention and SA1 and SA2 are the self-attention map for the pair of feature maps, respectively. * is the element-wise multiplication. The feature point with higher activation in a similar location in both feature maps inferred by the Self-Attention map (SA) also obtained a higher value in the the mutual attention map. The mutual attention map hence re-emphasized mutually discriminative visual cues and robust features across the two inputs. To make the mutual-attention map learnable and adjust its weights based on the proportion of correlated pattern, we passed the mutual attention map through a 1×1 convolutional layer as follows:

$$MA_i = f(MA_i, W_i) \tag{6}$$

where f is the convolutional layer with parameter W. The mutual-attention map was used to rescale the original feature map of the two inputs and enhance their respective activations. Given a feature map of two input images, *f1* and *f2*, and their mutual-attention map MA, the new feature maps *F1* and *F2* were enhanced by the mutual-attention map proportional to their relative magnitude. To consolidate the features learned across a layer, the original features were added back. Hence, the newly mutually boosted features maps *F1* and *F2* are given as follows:

$$F_1(i,j) = f_1(i,j) * MU(i,j) + f_1(i,j) \tag{7}$$

$$F_2(i,j) = f2(i,j) * MU(i,j) + f2(i,j) \tag{8}$$

where * refers to the element-wise multiplication. We inserted the mutual attention map at Layer 3 and Layer 4 of the model, where compact and high level features were learned. The classification task of the model learned to identify each person's identity. This assisted the verification part to learn person related features from the pair of inputs for which our mutual-attention map co-related visually salient concurrent regions previously boosted by the self-attention map of each branch. A detailed discussion and analysis of the effective regions learned by the mutual-attention layer are given in Section 6.

4. Loss Function and Similarity Learning

The model was based on two branch Siamese architectures, and parameters were shared between the two branches. The batch of input pairs was fed to the model, and two identity labels were predicated for each input pair. The final fully connected layer in ResNet-50 was replaced with a 512-dimensional sequential layer. The resulting feature vector was connected to an N-dimensional fully connected layer where N is the number of unique identities in the dataset in consideration. With the soft-max unit, the final image descriptor of size $1 \times 1 \times N$ was normalized to the N-dimensional vector representing predicted multi-class probabilities for N identities. Given f as the feature descriptor, the class probability prediction and cross-entropy losses are given as:

$$\hat{y} = Softmax(Sq(\Phi_c, f)) \tag{9}$$

$$Cls(f, y, \Phi_c) = \sum_{i=1}^{N} -y \log(y_i) \tag{10}$$

where $\hat{y}$ is the predicated probability, is S_q the sequential layer, and Φ_c is a parameter of the sequential layer. Cls –identification loss y is the ground truth for the target class.

For the verification task, we used the feature from the higher layer, which was enhanced and rescaled by the mutual-attention layer. This layer encoded aggregated, richer, and condensed activations. Hence, similarity was computed from features at this layer learned from input pairs. Given a 512-dimensional feature embedding f_1 and f_2 for the pair of input images, the Euclidean distance between these features was computed to give a 512-dimensional f_d distance feature vector encoding similarity between the inputs. The 512-dimensional sequential layer followed by the soft-max layer to encode f_d as a two-dimensional probability vector for the input pair indicating if the pair were similar or different. The verification task was formulated as binary classification. Binary Cross-Entropy (BCE) loss for predicated class probabilities is given as:

$$\hat{y} = Softmax(Sq(\Phi_c, f_d) \tag{11}$$

$$V(f_1, f_2, d, \Phi_c) = \sum_{i=1}^{N} -y_i \log(y_i) \tag{12}$$

where f_1 and f_2 are the feature descriptors for the pair of 512-dimensional vectors, d is the target label for input pair ([1, 0], the same; [0, 1], different), and $\hat{y}$ predicts the label of the inputs. V refers to the verification loss.

5. Experiment

5.1. Experimental Settings

- Input preparation: We used pre-trained ResNet-50 trained on ImageNet and used the feature extracted from the third and fourth residual block to compute the self- and mutual-attention map. We re-sized all input images to a resolution of 256 × 188, horizontally flipped, and the mean image computed from all training was subtracted from all the images. We also used random-erasing to

regularize and make the model robust. Positive and negative pairs were randomly chosen and shuffled in each mini-batch to avoid the model benefiting from a fixed input sequence and to avoid overfitting.

- Training: We used the Pytorch deep learning framework to implement the proposed model. The mini-batch size was set to 32 and with initial learning initialized as 1×10^{-2} The learning rate gradually faded by a factor of 1×10^{-1} between the 40^{th} and 60^{th} epoch. We used stochastic gradient descent to optimize the model and trained for 60 epochs. We maintained a drop-out rate of 0.75 for the fully connected layer to reduce the risk of overfitting. To maintain the stability of training and avoid vanishing and/or exploding gradients, the weights for the model were initialized using Kaiming initialization. As the model was trained with the classification and verification tasks, two cross-entropy losses, namely multi-class cross-entropy and binary cross-entropy loss, jointly optimized the training. We experimentally set the regulating weight coefficient of $\alpha = 0.5$ for verification loss (V_{loss}). Total training loss from the two tasks is computed as:

$$L_{total} = \alpha * V_{loss} + ID_{loss} \tag{13}$$

- Testing: During testing, the feature was first extracted from the gallery and queried using the trained model by feed-forwarding test dataset images of 256×188 and obtained the person descriptors of 512 dimensions. The final ranking was performed by calculating the Euclidean distance between each query image and all galleries. We used the commonly used Cumulative Match Curve (CMC) and mean Average Precision (mAP) for the performance evaluation of the model.

5.2. Datasets and Protocols

We conducted experiments on two large-scale datasets, namely Market-1501 [47] and Duke MTMC-ReID [48] to train and test the performance of the proposed model. Market-1501 was the largest re-ID dataset containing 32,668 manually annotated images boxes with 1501 pedestrians' identities captured with at most six different cameras with different views. The total identities were split into 751 training IDs and 750 for testing.

There was a total of 3368 images for query, which were randomly selected from each camera, making it possible to perform cross-camera search. The search for query images was performed from a gallery set containing 19,732 images and another 6796 distractor junk images. Some sample images from this dataset are shown in Figure 3.

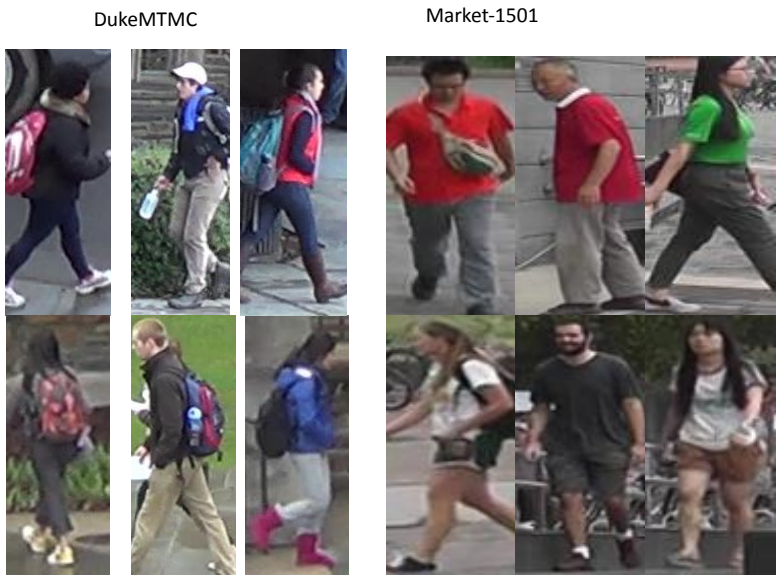

Figure 3. Sample pedestrian image from the Market-1501 and DukeMTMC-reID datasets.

The DukeMTMC-reID dataset was a subset of the DukeMTMC prepared for image-based person re-ID. The dataset was collected with eight different cameras. DukeMTMC-reID followed the same format and protocol as Market-1501 with 16,522 training images of 702 identities and 2228 query images of another 702 identities. The gallery set contained 17,661 images. It was the largest image dataset to have images that were cropped by hand-drawn bounding boxes. Some sample images from DukeMTMC-reID are shown in Figure 2.

6. Result and Discussion

We trained and evaluated our model on the Market-1501 and DukeMTMC-reID datasets. The results for Rank 1, Rank 5, Rank 10, and Rank 20 were used as performance measurements. The experimental results for the Market-1501 and DukeMTMC-reID datasets are shown in Table 1. Furthermore, a comparison with other works is given in Tables 2 and 3, respectively. Our proposed mutual-attention-based model outperformed the baseline methods by a considerable margin. As our model was designed with the standard ResNet-50 model as the backbone, we examined the improvement observed over ResNet-50 without the mutual-attention layer. The model also showed very competitive performance on similar supervised methods indicating the effectiveness of our approach and its prospect for being a competitive baseline technique. All results compared were with the single shoot scenario without re-ranking. We present the result analysis for Market-1501 and DukeMTMC-reID in the following subsections, and further analysis on the effect of the mutual-attention layer is discussed in detail in Section 6.3.

Table 1. Performance of the deep mutual attention model on the Market-1501 and DukeMTMC-reID datasets for Sing-shoot, multi-shoot, and re-ranking.

Setting	Rank	Market-1501	DukeMTMC-reID
Single-shoot	R = 1	90.74	80.83
	R = 5	90.36	90.08
	R = 10	97.86	93.49
	R = 20	98.57	94.74
	mAP	76.92	64.52
Multi-shoot	R = 1	93.82	-
	R = 5	97.86	-
	R = 10	98.81	-
	R = 20	99.34	-
	mAP	83.55	-
Re-ranking	R = 1	91.77	85.18
	R = 5	95.39	91.29
	R = 10	96.70	93.49
	R = 20	98.07	95.51
	mAP	87.30	80.65

6.1. Result on DukeMTMC-reID

For the DukeMTMC-reID dataset, we conducted an experiment on the single query setting. We also tested the model with re-ranking. Our approach with the mutual-attention layer outperformed the baseline ResNet-50 without the mutual-attention layer by a margin of 5.56% for Rank 1. Compared with the Deep Co-attention-based Comparator (DCC) [14], which employed a closer scheme to our method, the proposed model outperformed it by a margin of 0.53% and even by greater margin compared to many supervised methods, as shown in Table 2.

Table 2. Performance comparison with the state-of-the-art work for the deep mutual attention model on DukeMTMC-reID. DCC, Deep Co-attention-based Comparator.

Methods	Rank 1	mAP
LOMO+XQDA [49]	30.7	17.04
BoW+Kissme [47]	25.13	12.17
GAN(R) [50]	67.68	47.13
SPGAN [51]	46.4	26.2
IDE [31]	66.7	46.3
GOG [52]	65.8	-
GAN(R) [22]	67.68	47.13
LSRO [53]	67.7	47.1
SVDNet[54]	76.70	56.80
DCC [46]	80.3	59.2
PAN [55]	71.6	51.5
DPFL [41]	79.2	60.6
HA-CNN [44]	80.50	63.80
ResNet-50 Baseline	75.27	57.13
Ours	**80.83**	**64.52**

We fine-tuned and trained ResNet-50 baseline on the DukeMTMC-reID dataset for comparison purposes with our mutual-attention-based model.

6.2. Result on Market-1501

We conducted an experiment on Market-1501 on the single and multi-shoot scenarios. We also evaluated the model with re-ranking to further analyze the performance. Results for Rank 1, Rank 5, Rank 10, and Rank 20 and mAP are shown in Table 1. Likewise, the proposed model was compared with similar attention-based methods and with other supervised methods, as shown in Table 3. Our proposed model with the mutual-attention map outperformed the baseline ResNet-50 method without mutual-attention by a margin of 2.77% and 4.46% on Rank 1 and Rank 5, respectively. Compared with similar work on DCC in [14], our model outperformed by a margin of 4.04%, 0.66%, and 0.76% on Rank 1, Rank 5, and Rank 10, respectively.

Table 3. Performance comparison with state-of-the-art work for the deep mutual attention model on Market-1501. MA, Mutual-Attention.

Method	Single-Query		Multi-Shoot	
	Rank 1	mAP	Rank 1	mAP
PUL [56]	45.5	-	-	-
BoW [47]	34.4	14.1	-	-
OSML [53]	42.6	-	-	-
PIE [52]	65.7	41.1	-	-
S -CNN [15]	76.04	48.45	-	-
MSCAN [57]	80.3	57.5	-	-
SpindleNet [58]	76.9	-	-	-
LSRO [53]	83.9	66.1	-	
Part-aligned [43]	81.0	-	-	-
VGG16-Basel [59]	65.02	38.27	74.14	52.25
CaffeNet-Basel [3]	50.89	26.79	59.80	36.50
ResNet-50-Basel [27]	73.69	51.48	81.47	63.95
DCC [14]	86.7	69.4	-	-
ResNet-50 Base	87.97	72.46	-	-
Ours with MA	**90.74**	**76.92**	**93.82**	**83.55**

We fine-tuned and trained ResNet-50 baseline on the Market-1501 dataset for comparison purposes with our mutual-attention-based model.

6.3. Ablation Study

To see the effectiveness of the proposed mutual-attention map, we studied the activation map produced at the layer where we used the mutual-attention map. In this section, we analyze how the mutual-attention filtered out and emphasized key mutually relevant regions in input pairs. The mutual attention was inserted at Layer 3 and Layer 4 of the model. We extracted the feature map at Layer 4 for some samples to demonstrate how the learned features were improved with the help of the mutual-attention layers. The first row in Figure 4 shows the activation of regions across the image from

the baseline model without mutual-attention being used. The second row shows activation across the image region learned and guided by our mutual-attention map. While person related features were learned with supervision from the ground truth in the classification part, the mutual-attention assisted the learning by inferring visual cues at corresponding regions of input pairs and re-emphasized them to facilitate effective comparison in the verification part. It is clear from the figure that features learned with the help of the mutual-attention layer focused more on person body parts, and they were fairly distributed compared to the feature from the baseline model with no mutual-attention. For instance, closely watching the activation of the fourth image in Figure 4d, it can be noted that the salient regions learned from the baseline focused on the object carried by the person, whereas our model with mutual-attention expanded the focus from the object to more regions of the pedestrian like arms, legs, and some part of torso, which are important in identifying a person. For the image in the second column (b), we can also note that our model effectively paid attention to part of the heads, arms, and lower legs, whereas the baseline model attended only to part of the arm. The model gained this capability by virtue of mutually learned common features from input pairs conditioned on each other's spatial pattern, and features getting higher activation during training were implicitly selected and further boosted. In the meantime, the gradient produced from verification loss ensured the proportional magnitude to flow across these regions so that in the long run, the model relied on the cross-input feature across intermediate layers for final similarity judgment as opposed to independently learned feature maps solely from the last layer. Note that mutual-attention was used for pairs of input that were to be processed by the verification part. The classification task took the original input to make the predication for individual person classes.

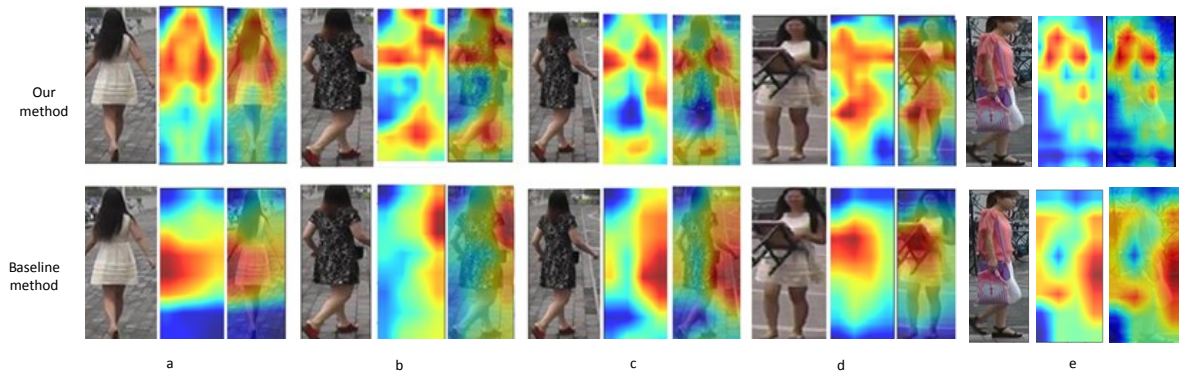

Figure 4. Visualization of the feature learned by the baseline and our model. The top list shows the feature learned by our proposed mutual-attention model, and the second row shows the feature learned by the baseline model. Each column (**a**–**e**), refers to raw input, activation map and activation map superimposed on raw input respectively. For each sample input image, salient regions learned and the salient region superimposed over the input are indicated. The features learned by our model are fairly focused on the salient person body regions and are distributed, while the features learned by the baseline method fail to emphasize the corresponding salient regions. (Best viewed in colour).

The Market-1501 dataset had multiple matched for each query image. The multi-shoot rank result was given in the earlier section. We further studied the model's retrieval performance. It is evident from Figure 5, given a query image on the left side, that our trained model was effective at retrieving multiple matching gallery persons (shown on the right side).

The first pedestrians image shown on the left side is a query image, and the rest along the same row from left to right are retrieved matches from the gallery set based on the similarity with the query in consideration. The correct matches retrieved are shown in the green bounding box with a few wrong matches shown in the red bounding box. The wrong matches were mostly hard negatives bearing a strong resemblance to the query image, like the query image shown in the last row. We also note that the model convergence was a little faster compared with the baseline model, as shown in the plot in

Figure 6 for the training loss verification and identification tasks. It could be noted that the verification loss obtained greatly reduced, and the model converged faster. This was in line with the assertion that features learned with the aid of the mutual-attention map facilitated similarity computation and back propagation to the lower layers, enabling learning effective features, which led to the loss incurred to be minimized faster, as demonstrated in Figure 6.

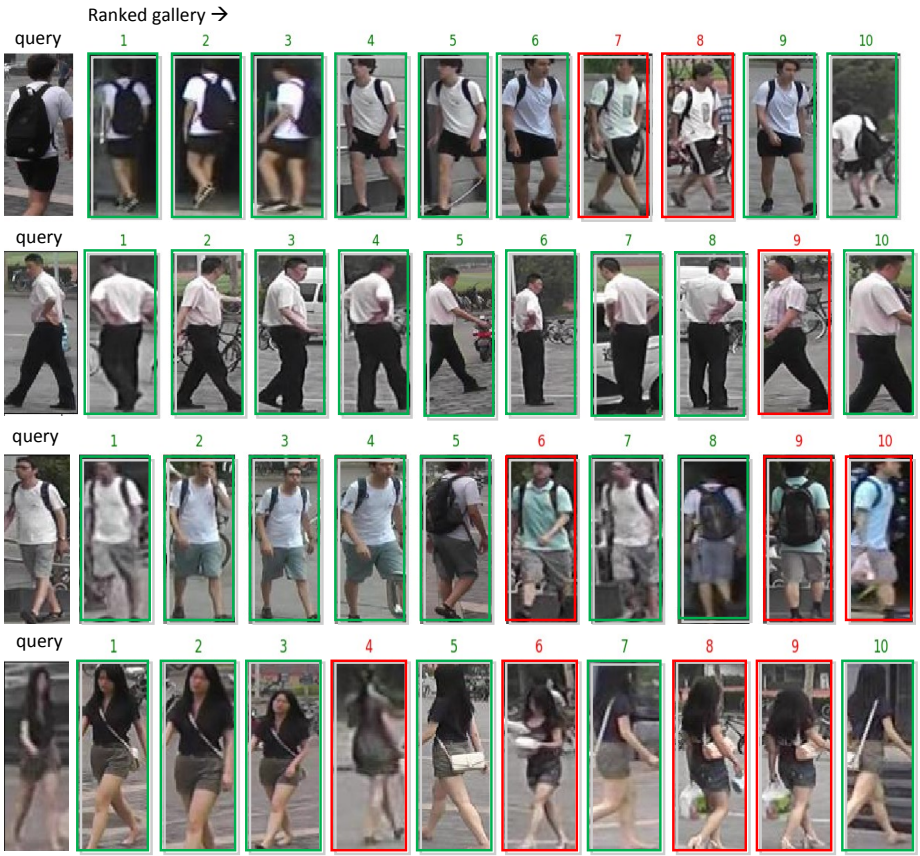

Figure 5. Retrieval performance of the model for a given query from the gallery set where each query has multiple matched in the gallery set.(Best viewed in colour).

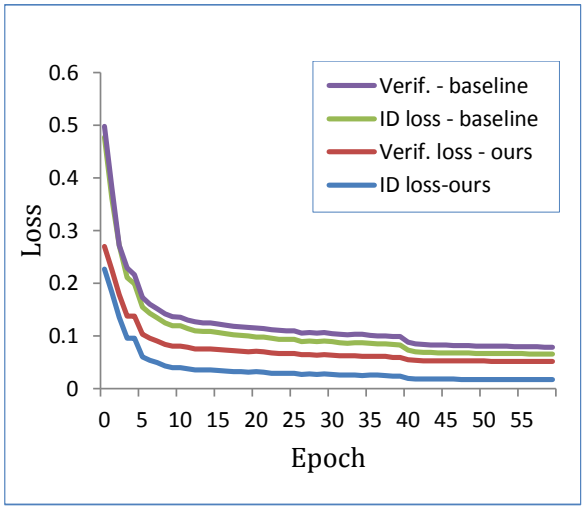

Figure 6. Convergence plot for training loss for the model with the DukeMTMC-reID dataset. ID, Identification; Verif., Verification.

7. Conclusions and Future Work

In this paper, we presented a deep mutual-attention layer based person re-identification model framed as identification and verification tasks. Our proposed model was comprised of a mutual-attention layer that bridged between two branches of the feature extraction layer in relating spatially active regions across the inputs and boosting them to favor an effective similarity judgment in the subsequent layers. Our mutual-attention was computed from the self-attention layer of the high-level branch, which summarized the feature map across the channel and spatial dimensions. This enabled the model to harness the interaction between both the channel and spatial layers of the input pairs. Key to the superiority of the proposed model was that the deep mutual-attention computed from the intermediate layer helped the model infer the common features across the input pairs and propagate this context to similarity measurement. Our experimental results and further analysis showed the effectiveness of the proposed mutual-attention layer, and the model outperformed many supervised model sand similar attention-based models. As future extension to this work, body part-based mutual-attention can be considered to alleviate the interference of background clutter.

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

Funding: This work was supported by The National Key Research and Development Program of China (2017YFC0108303), the Scientific and Technology Development Plan of Jilin Province, China (20170307002GX, 20190302112GX).

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

Abbreviations

The following abbreviations are used in this manuscript:

SA	Self-Attention
MU	Mutual-Attention
Cls	Classification Loss
V	Verification Loss

References

1. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
2. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
3. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Twenty-Sixth Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 1097–1105.
4. Zhang, G.P. Neural networks for classification: A survey. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2000**, *30*, 451–462. [CrossRef]
5. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
6. Dang, L.M.; Min, K.; Lee, S.; Han, D.; Moon, H. Tampered and Computer-Generated Face Images Identification Based on Deep Learning. *Appl. Sci.* **2020**, *10*, 505. [CrossRef]
7. Kubanek, M.; Bobulski, J.; Kulawik, J. A Method of Speech Coding for Speech Recognition Using a Convolutional Neural Network. *Symmetry* **2019**, *11*, 1185. [CrossRef]

8. Gadde, R.; Jampani, V.; Kiefel, M.; Kappler, D.; Gehler, P.V. Superpixel convolutional networks using bilateral inceptions. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 597–613.
9. Gregor, K.; Danihelka, I.; Graves, A.; Rezende, D.J.; Wierstra, D. Draw: A recurrent neural network for image generation. *arXiv* **2015**, arXiv:1502.04623.
10. Zhao, B.; Wu, X.; Feng, J.; Peng, Q.; Yan, S. Diversified visual attention networks for fine-grained object classification. *IEEE Trans. Multimed.* **2017**, *19*, 1245–1256. [CrossRef]
11. 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.
12. Bai, X.; Yang, M.; Huang, T.; Dou, Z.; Yu, R.; Xu, Y. Deep-person: Learning discriminative deep features for person re-identification. *Pattern Recognit.* **2020**, *98*, 107036. [CrossRef]
13. 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]
14. Wu, L.; Wang, Y.; Gao, J.; Li, X. Deep adaptive feature embedding with local sample distributions for person re-identification. *Pattern Recognit.* **2018**, *73*, 275–288. [CrossRef]
15. Varior, R.R.; Haloi, M.; Wang, G. Gated siamese convolutional neural network architecture for human re-identification. In *European cOnference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 791–808.
16. Wang, H.; Fan, Y.; Wang, Z.; Jiao, L.; Schiele, B. Parameter-Free Spatial Attention Network for Person Re-Identification. *arXiv* **2018**, arXiv:1811.12150.
17. Yang, F.; Yan, K.; Lu, S.; Jia, H.; Xie, X.; Gao, W. Attention driven person re-identification. *Pattern Recognit.* **2019**, *86*, 143–155. [CrossRef]
18. Wu, S.; Chen, Y.C.; Li, X.; Wu, A.C.; You, J.J.; Zheng, W.S. An enhanced deep feature representation for person re-identification. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 March 2016; pp. 1–8.
19. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
20. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef] [PubMed]
21. Abdel-Hamid, O.; Deng, L.; Yu, D. Exploring convolutional neural network structures and optimization techniques for speech recognition. *Interspeech* **2013**, *2013*, 1173–1175.
22. Venugopalan, S.; Xu, H.; Donahue, J.; Rohrbach, M.; Mooney, R.; Saenko, K. Translating videos to natural language using deep recurrent neural networks. *arXiv* **2014**, arXiv:1412.4729.
23. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE cOnference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
24. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
25. Kiros, R.; Salakhutdinov, R.; Zemel, R.S. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv* **2014**, arXiv:1411.2539.
26. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
27. 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, Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; Shah, R. Signature verification using a "siamese" time delay neural network. In Proceedings of the Advances in Neural Information Processing Systems Conference, San Francisco, CA, USA, November 1993; pp. 737–744.
29. 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.

30. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Deep metric learning for person re-identification. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 34–39.
31. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person re-identification: Past, present and future. *arXiv* **2016**, arXiv:1610.02984.
32. Liu, J.; Zha, Z.J.; Tian, Q.; Liu, D.; Yao, T.; Ling, Q.; Mei, T. Multi-scale triplet cnn for person re-identification. In Proceedings of the 24th ACM international conference on Multimedia, Amsterdam, The Netherlands, October 2016; pp. 192–196.
33. 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 July 2017; pp. 403–412.
34. Rahimpour, A.; Liu, L.; Taalimi, A.; Song, Y.; Qi, H. Person re-identification using visual attention. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 4242–4246.
35. 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.
36. Farenzena, M.; Bazzani, L.; Perina, A.; Murino, V.; Cristani, M. Person re-identification by symmetry-driven accumulation of local features. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2360–2367.
37. 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.
38. 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 Conference, Barcelona, Spain, 5–10 December 2016; pp. 2667–2675.
39. Li, W.; Zhu, X.; Gong, S. Person re-identification by deep joint learning of multi-loss classification. *arXiv* **2017**, arXiv:1705.04724.
40. Chen, W.; Chen, X.; Zhang, J.; Huang, K. A multi-task deep network for person re-identification. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
41. Chen, Y.; Zhu, X.; Gong, S. Person re-identification by deep learning multi-scale representations. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2590–2600.
42. Chung, D.; Tahboub, K.; Delp, E.J. A two stream siamese convolutional neural network for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1983–1991.
43. Zhao, L.; Li, X.; Zhuang, Y.; Wang, J. Deeply-learned part-aligned representations for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3219–3228.
44. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2285–2294.
45. Kalayeh, M.M.; Basaran, E.; Gökmen, M.; Kamasak, M.E.; Shah, M. Human semantic parsing for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1062–1071.
46. Wu, L.; Wang, Y.; Gao, J.; Tao, D. Deep Co-attention based Comparators For Relative Representation Learning in Person Re-identification. *arXiv* **2018**, arXiv:1804.11027.
47. 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, Santiago, Chile, 7–13 December 2015; pp. 1116–1124.
48. Gou, M.; Karanam, S.; Liu, W.; Camps, O.; Radke, R.J. DukeMTMC4ReID: A large-scale multi-camera person re-identification dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017.

49. 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.
50. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 3754–3762.
51. Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; Jiao, J. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 994–1003.
52. Matsukawa, T.; Okabe, T.; Suzuki, E.; Sato, Y. Hierarchical gaussian descriptor 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. 1363–1372.
53. Bak, S.; Carr, P. One-shot metric learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 2990–2999.
54. Wang, Y.; Wu, L.; Lin, X.; Gao, J. Multiview spectral clustering via structured low-rank matrix factorization. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 4833–4843. [CrossRef]
55. 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]
56. 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. [CrossRef]
57. 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 Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 384–393.
58. 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 Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 1077–1085.
59. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.