Local-Aware Global Attention Network for Person Re-Identification Based on Body and Hand Images

Nathanael L. Baisa

Abstract—Learning representative, robust and discriminative information from images is essential for effective person re-identification (Re-Id). In this paper, we propose a compound approach for end-to-end discriminative deep feature learning for person Re-Id based on both body and hand images. We carefully design the Local-Aware Global Attention Network (LAGA-Net), a multi-branch deep network architecture consisting of one branch for spatial attention, one branch for channel attention, one branch for global feature representations and another branch for local feature representations. The attention branches focus on the relevant features of the image while suppressing the irrelevant backgrounds. The global and local branches intends to capture global context and fine-grained information, respectively. A set of ablation study shows that each component contributes to the increased performance of the LAGA-Net. Extensive evaluations on four popular body-based person Re-Id benchmarks and two publicly available hand datasets demonstrate that our proposed method consistently outperforms existing state-of-the-art methods.

Index Terms—Person re-identification, Deep representation learning, Attention mechanisms, Global features, Part-level features.

I. INTRODUCTION

Person re-identification (Re-Id), matching a particular person across different times, places, or cameras, has recently received a lot of attention from both industry and academia for different applications such as intelligent video surveillance. It is currently the main component of visual tracking in both single camera [1] and multiple camera [2]. It is also similar to image retrieval in many aspects. Given a query image, person Re-Id ranks gallery images in terms of similarity to the query image. The image can be of a person’s body, hand, face, etc. In this process, each image is represented with a feature embedding. Learning robust and discriminative feature representations is very crucial to overcome the many challenges the person Re-Id is facing. These challenges [3] include pose variations, occlusion, view point changes, lighting changes, background clutter, noisy labels, etc. Various efforts have been made to address these challenges [4], for instance, considering whole body [5–7], body parts [8–10] and attention mechanisms [11–13] for learning robust and discriminative feature representations from person body images in uncontrolled environments for better matching.

Person Re-Id for biometric application has recently also received a lot of attention [14]. Hand images, one of the primary biometric traits [15], [16], deliver discriminative features for biometric person recognition. Hand images not only have less variability when compared to other biometric modalities but also have strong and diverse features which remain relatively stable after adulthood [16–18]. Because of this, there is a strong potential to investigate hand images captured by digital cameras for person recognition, especially for criminal investigation in uncontrolled environments since they are often the only available information in cases of serious crime such as sexual abuse.

In this work, we propose a compound approach for end-to-end discriminative deep feature representations learning for person Re-Id based on both body and hand images. We carefully design the Local-Aware Global Attention Network (LAGA-Net), a multi-branch deep network consisting of 4 branches to learn deep global, attentive and local (part-level) feature representations which are robust and discriminative enough for dealing with person Re-Id challenges. Each branch has its own merit within the goal of the proposed network architecture for learning robust and discriminative deep feature representations. Specifically, the attention branches, channel attention branch and spatial attention branch, focus on the relevant features of the image while suppressing the irrelevant backgrounds. The spatial attention branch incorporates relative positional encodings into spatial attention module to maintain translation equivariance. The global and local branches intends to capture global context and fine-grained information, respectively. By carefully designing this end-to-end compound method, we have shown that it is possible to effectively learn robust and discriminative feature representations for person Re-Id based on both body and hand images, unlike the previous methods [5–13]. Our contributions can be summarized as follows.

1) We propose a multi-branch deep network by incorporating both channel and spatial attention modules in branches in addition to global (without attention) and local branches for person Re-Id based on both body and hand images which is efficient computationally and flexible in terms of the backbone architecture.

2) We include relative positional encodings into the spatial attention module, considering height and width independently, to capture the spatial positions of pixels in order to overcome the weakness of the attention mechanisms - equivariant to pixel shuffling, for efficiently re-identifying a person based on body and hand images.

3) We make extensive evaluations on four popular body datasets: Market-1501 [19], DukeMTMC-Re-ID [20], CUHK03 [5], MSMT17 [3] and two publicly available hand datasets: 11k [21], HD [22], and LAGA-Net significantly outperforms existing state-of-the-art methods.
on these datasets.

The rest of the paper is organized as follows. After the discussion of related work in Section II the proposed method is described in Section III including the attention modules, the overall architecture of the LAGA-Net and the loss functions. The experimental results are analyzed and compared in Section IV followed by the discussion in Section V and the main conclusion along with suggestion for future work in Section VI.

II. RELATED WORK

A. Body-based Person Re-Id

Many person Re-Id methods have been proposed over the last few years, with more performance gain obtained by methods based on deep learning, using both supervised [8–13] and unsupervised [23], [24] learning approaches. Some of these works are based on learning global deep feature representations [5]–[7]. To overcome the performance limitations of the person Re-Id methods based on global features, researchers shifted their attention to learn deep local (part-level) feature representations by considering person poses [8]. In this case, external pose (skeleton) estimation methods have been used to leverage human part cues. However, this also comes with another disadvantage since the errors in the pose estimation can propagate to the re-identification stage. Mask is also used as external cues to remove the background clutters in pixel level to retrieve body shape information [25]. Using these external methods bring additional computational burden. To overcome this, uniform partitioning of the images without relying on external methods was introduced for body-based person Re-Id in [9], [10] and for hand-based person identification in [16]. Though this approach helped in gaining performance boost over the previous methods, the performance of the methods based on this approach is not sufficient to handle the challenges the person Re-Id methods are facing, probably due to pose misalignment problem.

Recently, self-attention mechanism, an integral component of Transformers [26], has received great attention in deep learning. The self-attention mechanism captures long-term information and interactions amongst all entities (e.g. pixels, channels, sequence elements, etc.) of the input data. It updates each pixel, for instance, by aggregating global information from all pixels in the input image. The attention mechanism has been used in [11]–[13] for person Re-Id by considering both channel and spatial attentions which compute correlations between all the channels and all the pixels of the input feature map, respectively. However, these methods used the attention modules across the entire network layers which make it computationally inefficient since the self-attention computation is very expensive if the dimension of the input data is very large. Furthermore, these methods have limited performance when applied to person body and hand images. By default, the attention mechanism does not model relative or absolute position information. To overcome this, a relative position representation of a sequence element with respect to its neighbours has been proposed in [27] for natural language processing, and later on incorporated into a standalone global

attention-based deep network for images without using convolutions for modeling pixel interactions [28]. Unlike these methods, we use the attention mechanism along with the convolution operations by applying the attention modules only to low-resolution feature maps in later stages of a deep network in branches along with global (without attention) and local branches for better computational efficiency as well as accuracy for person Re-Id based on not only body images but also hand images.

B. Hand-based Person Re-Id

Both traditional and deep learning approaches have been combined in [17], [21] to develop person identification using hand images. After training a convolutional neural network (CNN) on digital hand images (RGB), the method in [21] used the network as a feature extractor to obtain CNN-features which have been fed into a set of support vector machine (SVM) classifiers. The work in [17] used rather similar approach with additional data type for fusion, near-infrared (NIR) images. These methods are not an end-to-end. An end-to-end approach considering both horizontal and vertical uniform partitioning has been proposed in [16]. However, all these methods have a limited performance. Unlike these methods, our proposed method is a compound approach considering multi-branch network architecture which is efficient not only on body images but also on hand images for re-identifying individuals.

III. PROPOSED METHOD

In this section, we introduce the two attention modules followed by the overall architecture of the LAGA-Net and the used loss functions. The goal of the attention modules is to suppress irrelevant backgrounds while focusing on discriminative information of person appearances.

A. Channel Attention Module

Channel attention module (CAM) aims to aggregate channel-wise feature-level information since some channels in higher convolutional layers are semantically related i.e. CAM computes the correlations between all the channels. The structure of CAM is given in Fig. 13. Given the input feature map $E_i \in \mathbb{R}^{C \times H \times W}$ where $C$, $H$, $W$ are the number of channels, height and width of the feature map, respectively, we first reshape it to produce the matrices of keys, queries and values, respectively, as $K \in \mathbb{R}^{C \times HW}$, $Q \in \mathbb{R}^{C \times HW}$ and $V \in \mathbb{R}^{C \times HW}$. Then, the global channel attention map $A_c \in \mathbb{R}^{C \times C}$ is computed using the dot-product of the query with all keys as

$$A_c = \rho(KQ^T)$$

where $Q^T$ denotes the matrix transpose of $Q$, and $\rho$ represents the softmax normalization along each row separately. The self-attended output feature map $E_o \in \mathbb{R}^{C \times H \times W}$ for the channel attention is given by

$$E_o = \gamma(A_cV) + E_i$$

where $\gamma$ is a parameter to adjust the output feature map. The attention mechanism aims to aggregate the most relevant feature maps for each channel by computing their correlations with all other feature maps in the input feature map. The softmax normalization is used to ensure that the output feature maps are normalized and represent the relative importance of each channel.

B. Spatial Attention Module

Spatial attention module (SAM) extends the channel attention module to aggregate spatial information. It takes the channel-attended feature map and computes spatial correlations between all the spatial locations in the feature map.

$$A_s = \rho(KQ^T)$$

where $K \in \mathbb{R}^{H \times HW}$ and $Q \in \mathbb{R}^{H \times HW}$ are the matrices of keys and queries, respectively. The output feature map is then computed by applying the softmax normalization along each column separately:

$$E_o = \gamma(A_sV) + E_i$$

The spatial attention module is used to aggregate spatial information by considering the most relevant spatial locations in the feature map. It is important to note that the channel- and spatial-attention modules are applied sequentially in the proposed method to obtain a more accurate representation of the input feature map.

The LAGA-Net architecture consists of multiple branches with both channel and spatial attention modules. The overall architecture is shown in Fig. 14. The network is trained using the proposed attention modules along with the modified cross-entropy loss function. The experiments are conducted on the CPHD and CUB datasets, and the results show significant improvements in terms of accuracy compared to existing methods.
where $\gamma$ is initialized as 0 and gradually learns to assign more weight to adjust the impact of the CAM. $A_s V$ means that the matrix of values $V$ is weighted by the attention score $A_s$.

**B. Spatial Attention Module with Relative Positional Encodings**

The goal of spatial attention module is to aggregate the semantically similar pixels in the spatial domain of the input feature map. Though the spatial attention mechanism attends to the entire input feature map based on content (pixel values), it does not take into account the spatial positions of pixels which makes it equivariant to pixel shuffling. To overcome this, we incorporate the relative positional encodings along the rows (height) and columns (width), which is computationally efficient, so that it maintains translation equivariance i.e. translating (shifting) the input pixel also translates the output pixel by the same amount. The structure of the Spatial Attention Module with Relative Positional Encodings (SAM-RPE) is given in Fig. 1b. Given the input feature map $E_i \in \mathbb{R}^{d_h \times d_w \times d_k}$, the re-indexing tensor $I_H \in \mathbb{R}^{d_H \times (2H-1)}$ and its corresponding self-attended feature map $E_H \in \mathbb{R}^{C \times H \times W}$ can be obtained by transforming it through defined learnable weight matrices $W_K$, $W_Q$ and $W_V$, respectively, where $d_k = \frac{C}{s}$ is the channels dimension of the keys and queries. The learnable weight matrices $W_K$, $W_Q$ and $W_V$ are implemented using independent pointwise $(1 \times 1)$ convolution layers with batch normalization and ReLU activation. Thus, the global spatial attention map $A_s \in \mathbb{R}^{H \times W}$ is computed as

$$A_s = \rho(K^T Q)$$

where $K^T$ denotes the matrix transpose of $K$, and $\rho$ represents the softmax normalization along each row separately.

We consider height (row) and width (column) attentions due to the relative spatial positions for computational efficiency. To compute these relative positional attentions, we first need to represent relative shifts along the height or width of the input feature map. Let a relative position embedding for the height that needs to be learned be $R_H \in \mathbb{R}^{d_h \times (2H-1)}$ where $H$ is the height of the input feature map and $d_h = \frac{C}{s}$ is the number of channels. A possible vertical shift, from $-(H-1)$ to $H-1$, corresponds to each row of $R_H$. The relative shifts in the matrix $R_H$ need to be represented using absolute shifts. To do this, the re-indexing tensor $I_H \in \mathbb{R}^{H \times W \times (2H-1)}$, which is used as a mask, can be defined as

$$I_{h,i,r}^H = \begin{cases} 1, & \text{if } i - h = r \ \& \ |i - h| \leq H \\ 0, & \text{otherwise} \end{cases}$$

where $h \in \{0, ..., H-1\}$, $i \in \{0, ..., W-1\}$ and $r \in \{-(H-1), ..., 0, ..., H-1\}$.

By reshaping $I_H$ to $I_H \in \mathbb{R}^{HW \times (2H-1)}$, a position embedding tensor with indices of absolute shifts for the height $P_H \in \mathbb{R}^{HW \times d_k}$ is given by

$$P_H = I_H R_H$$

Then, the self-attended feature map $E_H \in \mathbb{R}^{C \times H \times W}$ corresponding to the height relative position embedding $R_H$, which is used as keys implicitly ($P_H$ explicitly), is computed as

$$E_H = V(P_H Q)$$

where $P_H Q$ corresponds to the height relative positional attention.

The relative position embedding for the width $R_W \in \mathbb{R}^{(2W-1) \times d_w}$, the re-indexing tensor $I_W \in \mathbb{R}^{HW \times (2W-1)}$ and its corresponding self-attended feature map $E_W \in \mathbb{R}^{C \times H \times W}$ can be obtained with similar approach to the above height formulation since they are symmetric.

Thus, the final self-attended output feature map $E_o \in \mathbb{R}^{C \times H \times W}$ for the SAM-RPE is given by

$$E_o = \gamma(VA_s + BN(E_H) + BN(E_W)) + E_i$$

where $\gamma$ is initialized as 0 and gradually learns to assign more weight to adjust the impact of the SAM-RPE, and $BN$ is a batch normalization.

**C. Network Architecture Overview**

The overall architecture of the proposed LAGA-Net is given in Fig. 2. It incorporates two complementary attention modules, channel and spatial. These attention modules are used at higher level of the network, for computational efficiency, in branches along with the global (without attention) branch and the local branch which is obtained by performing uniform horizontal partitioning. As a backbone network, we use ResNet50 [29] pre-trained on ImageNet due to its precise architecture with competitive performances in some person Re-Id works [9], [10], [12], [13]. Obviously, any network designed for image classification can be adapted, for example Inception network [30] and DenseNet [31]. We keep the structure of the original ResNet50 before layer 3 (inclusive) remain the same when we modify the backbone network to produce the LAGA-Net. We create 4 independent branches just after the layer 3 of the ResNet50 to incorporate the channel and the spatial (with relative positional encodings) attention modules in branches by keeping one (without attention) global branch and one additional local branch for which we generate 3 horizontal stripes uniformly from the output feature map.

**Spatial attention branch:** This branch aggregates the semantically similar pixels in the spatial domain of the input feature map and it uses a Global Average Pooling (GAP) layer to summarize the 3D tensor of activations to form a 2048-dimensional column feature vector $s$.

**Channel attention branch:** This branch aggregates the correlations between all the channels of the input feature map and it uses the GAP layer to summarize the 3D tensor of activations to form a 2048-dimensional column feature vector $c$.

**Global branch:** This branch aims to maintain global context information for discriminative feature learning, and the GAP layer is used to summarize the 3D tensor of activations to form a 2048-dimensional column feature vector $g$.

**Local branch:** For this branch, we change from the GAP layer to conventional average pooling (AP) layer to create uniform
horizontal partitions (stripes) on the 3D tensor of activations to learn 2048-dimensional part-level features $p_i$ where integer $i \in [1, 3]$; the total number of partitions used is 3.

Each reduction layer for each branch (for each stripe in the case of the local branch) is implemented using a new fully-connected layer (FC), batch normalization (BN), leaky rectified linear unit (LReLU) and dropout with probability of 0.5 to reduce possible over-fitting. The reduction layers are employed to convert 2048-dimensional column feature vectors obtained after the GAP and conventional AP layers to 1024-dimensional feature vectors which in turn are fed into the classification layers. Each classification layer, which is implemented using a FC layer followed by a softmax function, predicts the identity (ID) of each input. In addition, we change the last stride from 2 to 1 in the backbone network i.e. remove the last spatial down-sampling operation, which increases the size of the tensor of each branch for improved performance as observed in [7].

D. Loss Functions

The LAGA-Net is optimized during training by minimizing the loss function $\mathcal{L}$ consisting of the sum of cross-entropy losses over the 6 ID predictions for identification (classification) and the sum of hard mining triplet losses over the 6 ID predictions for metric learning i.e. each classifier predicts the identity of the input image as shown in Fig. 2. The total loss function $\mathcal{L}$ is given by

$$\mathcal{L} = \sum_{l=1}^{6} \mathcal{L}_{l,xent} + \beta \sum_{l=1}^{6} \mathcal{L}_{l,triplet}$$

where $\beta$ is a hyperparameter balancing the two types of losses.

For the learned features $f_i$, the cross-entropy loss (softmax loss) with label smoothing [32] is given as:

$$\mathcal{L}_{l,xent} = -\sum_{i=1}^{N} q_{y_i} \log \frac{e^{W_c^T f_i + b_c}}{\sum_{c=1}^{C} e^{W_c^T f_i + b_c}}$$

where $N$ is the batch-size, $C$ is the number of classes (identities) in the training dataset, and $W_c$ and $b_c$ are weight vector and bias for class $c$, respectively. Note that $z_c = W_c^T f_i + b_c$ are the logits or unnormalized probabilities. The ground-truth distribution over labels $q_{y_i}$ by including label smoothing can be given as

$$q_{y_i} = \begin{cases} 1 - \frac{C-1}{C} \epsilon, & \text{if } y_i = y \\ \frac{1}{C} \epsilon, & \text{otherwise} \end{cases}$$

where $y$ is ground-truth label and $\epsilon$ is a smoothing value.
Fig. 2: Structure of LAGA-Net. Four separate 3D tensors (one for spatial attention branch, one for channel attention branch, one for global branch and the other for local branch) are obtained by passing the input image through the stacked convolutional layers from the backbone network. S3 and S4 are the SAM-RPE after layer 3 and layer 4 (L4) of the ResNet50, respectively. Similarly, C3 and C4 are the CAM after layer 3 and layer 4 of the ResNet50, respectively. Three horizontal partitions (stripes) are also performed on L4 to produce the local branch. Given an input image, six separate 2048-D column feature vectors are obtained by passing it through the backbone network with the 4 branches (the local branch has 3 horizontal stripes). Each classifier predicts the identity (ID) of the input image during training. In case of hand-based person Re-Id, hand input images are used.

Similarly, the batch hard mining triplet loss [33] is given as follows:

$$\mathcal{L}_{ triplet} = \sum_{i=1}^{P} \sum_{a=1}^{K} \left[ \alpha + \max_{p=1, \ldots, K} \|f_a(i) - f_p(i)\|_2 \right]^{\text{hardest positive}} - \min_{n=1, \ldots, K, j=1, \ldots, P, j \neq i} \|f_a(i) - f_n(j)\|_2^{\text{hardest negative}}$$

where $[A]^+ = \max(A, 0)$, $\alpha$ is the margin hyperparameter that controls the distance differences of intra and inter classes, and $f_a(i), f_p(i), f_n(j)$ are the features extracted from anchor, positive and negative samples, respectively. The positive and negative samples refer to the persons with same or different identity with the anchor. The candidate triplets are constructed by the furthest positive and closest negative sampled pairs. These are basically the hardest positive and hardest negative pairs in a mini-batch $N = PK$ with $P$ selected identities and $K$ instances (images) per identity. In our experimental settings, we use $\beta = 0.1$, $\alpha = 1.2$, $P = 5$, $K = 4$ and $N = PK = 20$.

For both losses, we use the learned embeddings $s, c, g, p_1, p_2, p_3$ (in place of $f$ in the Eqs. (9) and (11)) during training (see in Fig. 2). However, during testing, we concatenate all the 2048-D feature vectors of the 6 branches as the final feature embedding, just after the GAP and conventional AP i.e. $\mathcal{F} = [s, c, g, p_1, p_2, p_3]$ which becomes 12288-D feature vector, and then compare feature vector of each query image with gallery feature vectors using cosine distance. In fact, we put image $I$ and its horizontally flipped image $I'$ into the model, and get their embeddings $\mathcal{F}$ and $\mathcal{F}'$. Then their mean $\frac{\mathcal{F} + \mathcal{F}'}{2}$ is used as the embedding of image $I$ during testing which improves the Re-Id performance.

**IV. EXPERIMENTS**

**A. Datasets**

In this section, we describe both body person Re-Id datasets such as Market-1501 [19], DukeMTMC-Re-ID [20], CUHK03 [5] and MSMT17 [3], and hand datasets such as 11k hands dataset [21] and Hong Kong Polytechnic University Hand Dorsal (HD) dataset [22].

**Market-1501** [19] is captured by 6 cameras comprising 32,668 labeled images of 1,501 identities. 751 identities (12,936 images) are used for training whereas the rest are used for testing. The testing data is sub-divided into test probe (query) set and test gallery set. The test probe set has 3,368 images of 750 identities, while 2,793 additional distractors are also included into the test gallery set. The overall statistics of body-based person Re-Id datasets used in this paper is given in Table I.

**DukeMTMC-Re-ID** [20] is captured by 8 cameras containing 36,411 images of 1,812 identities. Among these identities, 1,404 identities appear in more than 2 cameras while 408 identities (distractors) appear in only one camera. 702 identities are randomly chosen for training from the 1,404 identities and the rest are used for testing. One query image for each identity per camera is chosen from the testing set for the probe set
while all remaining images including the distractors are used for gallery set. 

CUHK03 [5] contains 13,164 images of 1,467 persons, and each identity only appears in two disjoint camera views. The new training and testing protocol proposed in [34] is adopted, in which 767 identities are used for training and 700 for testing. Both labeled and detected bounding boxes are given in CUHK03, we perform experiments on the labeled (L) bounding boxes.

MSMT17 [3] is captured by a 15-camera network (12 outdoor, 3 indoor) and is currently the largest publicly available person Re-Id dataset with 126,441 images of 4,101 identities. The training set contains 32,621 images with 1041 persons (identities), while the test set contains 93,820 images with 3060 persons. For the test set, 11,659 images are randomly selected as query, and the other 82161 images are used as gallery i.e. we use the training-testing split of [3]. The video is collected with different weather conditions at three-time slots (morning, noon, afternoon). The annotations include camera IDs, weathers and time slots. Thus, the MSMT17 is significantly more challenging than the other three due to its massive scale, more complex and dynamic scenes.

11k hand dataset [21] has 190 subjects (identities). We use the same partitioning strategy of the dataset as in [16]. As in [16], this dataset is divided into right dorsal, left dorsal, right palmar and left palmar sub-datasets to train a hand-based person Re-Id (recognition) model. After excluding accessories and dividing the dataset as in [16], right dorsal has 143 identities, left dorsal has 146, right palmar has 143 and left palmar has 151 identities. The first half and the second half of each sub-dataset based on identity are used for training and testing, respectively. For instance, for right dorsal, the first 72 identities are used for training and the last 71 identities are used for testing. Similarly, the first 73, 72 and 76 identities are used for training phase for left dorsal, right palmar and left palmar sub-datasets, respectively. The remaining identities of each sub-dataset (73 for left dorsal, 71 for right palmar, 75 for left palmar) are used for testing. From each identity of the test set of each sub-dataset, we randomly choose one image and put in a common gallery for all the 11k sub-datasets. The remaining images of each identity of the test set of each sub-dataset are used as a query set for that sub-dataset. Accordingly, the 11k gallery has 290 images and the query has 971 images for right dorsal, 988 images for left dorsal, 917 images for right palmar and 948 images for left palmar. A randomly chosen image of each identity of the training set of each sub-dataset is used as a validation for monitoring the training process. This procedure is repeated for 10 times and the average performance is reported.

B. Implementation Details

We implemented the LAGA-Net using PyTorch deep learning framework and trained it on NVIDIA GeForce RTX 2080 Ti GPU.

Body-based person Re-Id: The input images of size $384 \times 128$ are resized to $\frac{3}{4}$ times of the size of the input images and then randomly cropped to $384 \times 128$, augmented by random horizontal flip, normalization, color jittering and random erasing [35] during training. The test images are resized to $384 \times 128$ and augmented only by normalization.

Hand-based person Re-Id: The input images are resized to $356 \times 356$ and then randomly cropped to $324 \times 324$, augmented by random horizontal flip, normalization and color jittering during training. However, only normalization is utilized during testing with the test images resized to $324 \times 324$, without a random crop.

In both cases, a random order of images are used by reshuffling the dataset. We use a combination of cross-entropy and hard mining triplet losses over the 6 ID predictions as in Eq. [6] to train the LAGA-Net. To prevent over-fitting and over-confidence, label smoothing [32] with smoothing value ($\epsilon$) of 0.1 is also used with the cross-entropy loss. We train the model for 70 epochs with mini-batch size of 20 and Adam optimizer with the weight decay factor for L2 regularization of $5 \times 10^{-4}$. For the first 10 epochs, we use a warmup strategy [7], increasing a learning rate linearly from $8 \times 10^{-6}$ to $8 \times 10^{-4}$, and then it is decayed to $4 \times 10^{-4}$, $2 \times 10^{-4}$ after 40 and 60 epochs, respectively. The learning rate is divided by 10 for the existing layers of the backbone network i.e. ten times bigger learning rate is given to the newly added layers (FC layers and batch normalizations) and the attention modules (embedding functions and batch normalizations), with appropriate weight and bias initializations.

C. Evaluation Metrics

We use the standard person Re-ID evaluation metrics, particularly, Cumulative Matching Characteristics (CMC) [19] (rank-1 or top-1 matching accuracy) and mean Average Precision (mAP) [19] to evaluate our proposed person Re-Id method.

D. Ablation Analysis

Our proposed method, as described in Section III incorporates two complementary attention modules, channel and
TABLE I
Statistics of body-based person Re-Id datasets used in this paper: Market-1501 [19], DukeMTMC-Re-ID [20], CUHK03 [5] and MSMT17 [3]. Number of identities (ids), number of images and number of cameras are shown for train set, query set and gallery set of each dataset.

| Subset | Market-1501 | DukeMTMC | CUHK03 (L) | MSMT17 |
|--------|-------------|-----------|------------|--------|
|        | # ids # images # cameras | # ids # images # cameras | # ids # images # cameras | # ids # images # cameras |
| Train  | 751 12,936 6 | 702 16,522 8 | 767 7,365 2 | 1,041 30,248 15 |
| Query  | 750 3,368 6 | 702 2,228 6 | 700 1,400 2 | 3,060 11,659 15 |
| Gallery| 751 15,913 6 | 1,110 17,661 8 | 700 5,332 2 | 3,060 82,161 15 |

TABLE II
Statistics of hand-based person Re-Id datasets used in this paper: 11k [21] and HD [22]. Number of identities (ids) and number of images are shown for train set, query set and gallery set of each dataset. Only one camera is used to capture each of these datasets.

| Subset | D-r of 11k | D-l of 11k | P-r of 11k | P-l of 11k | HD |
|--------|------------|------------|------------|------------|----|
|        | # ids # images | # ids # images | # ids # images | # ids # images | # ids # images |
| Train  | 72 962 | 73 808 | 72 977 | 76 1,004 | 251 2,407 |
| Query  | 71 971 | 73 988 | 71 917 | 75 948 | 251 1,992 |
| Gallery| 290 290 | 290 290 | 290 290 | 290 290 | 464 1,593 |

spatial with relative position, at a higher level of the network in branches along with the global (without attention) and local branches. The ablation analysis of these components is given in Table III with evaluations on the Market-1501 [19] body dataset and palmar right (P-r) of the 11k hands dataset [21]. As shown in this table, each component contributes to a performance gain. For body-based person Re-Id, for instance, the global branch (ResNet50 with some modifications) gives rank-1 and mAP of 93.92% and 83.21%, respectively. Incorporating the local branch with three horizontal partitions (stripes) boosts the performance to 94.41% rank-1 and 85.30% mAP. Rank-1 accuracy and mAP of 95.12% and 86.91%, respectively, are obtained by integrating the channel attention module (CAM). Incorporating the spatial attention module with relative positional encodings (SAM-RPE) contributes to the performance gain as well, giving the overall LAGA-Net performance of 96.18% rank-1 and 88.76% mAP on the Market-1501 dataset. Similarly, each component contributes to a performance gain on palmar right (P-r) of the 11k hands dataset as shown in this table.

TABLE III
Ablation analysis on components of LAGA-Net on Market-1501 [19] and palmar right (P-r) of 11k [21]. Global + Local + CAM + SAM-RPE gives LAGA-Net. The results are shown in rank-1 accuracy (%) and mAP (%).

| Method          | Market1501 rank-1 mAP | Market1501 P-r (11k) mAP | Palmar Right rank-1 mAP | Palmar Right P-r (11k) mAP |
|-----------------|------------------------|--------------------------|-------------------------|---------------------------|
| Global          | 93.92 83.21            | 95.43 95.95              | 96.52 96.97             | 97.14 97.45               |
| + Local         | 94.41 85.30            | 96.43 96.95              | 97.52 97.59             | 98.16 98.54               |
| + CAM           | 95.12 86.91            | 96.52 96.97              | 97.14 97.45             | 98.16 98.54               |
| + SAM-RPE       | 96.18 88.76            | 98.16 98.54              | 98.16 98.54             | 98.16 98.54               |

E. Comparison with the State-of-the-art Methods

Body-based person Re-Id: We evaluate our model and report the results using rank-1 matching accuracy and mAP on Market-1501 [19], DukeMTMC-Re-ID [20], CUHK03 [5] and MSMT17 [3] datasets. For fair comparison, we did not use post-processing such as re-ranking [34] or multi-query [19]. We compare our proposed method, LAGA-Net, to many existing state-of-the-art methods and report the quantitative performance comparison in Table IV. As shown in this table, our method outperforms all other methods across all datasets in both rank-1 accuracy and mAP evaluation metrics except on CUHK03 where our method is ranked 2nd in rank-1 accuracy. This indicates that our method is more generalizable than the other methods across all datasets. Specifically, our proposed method outperforms the existing part-based methods such as PCB+RPP [9] and MGN [10] by large margin. For instance, the LAGA-Net outperforms the MGN by 13.73% in rank-1 accuracy and 10.54% mAP on CUHK03 dataset. Similarly, the LAGA-Net outperforms the existing attention-based methods such as MHN [11], ABD-Net [12] and RGA-Net [13]. For instance, the LAGA-Net outperforms the RGA-Net by 2.26% in rank-1 accuracy and 3.47% in mAP on MSMT17 dataset. Our proposed method outperforms not only these supervised body-based person Re-Id methods, but also the recent unsupervised person Re-Id methods such as RLCC [23] and ICS [24] as shown in Table IV.

Hand-based person Re-Id: We compare our proposed method to many existing state-of-the-art hand-based Re-Id (recognition) methods such as GPA-Net [16], MBA-Net [39], RGA-Net [13] and ABD-Net [12]. The GPA-Net was designed for hand-based person identification, however, both RGA-Net and ABD-Net were designed for body-based person re-identification. Therefore, we trained both RGA-Net and ABD-Net on hand datasets using the same experimental settings (loss function, optimizer, hyperparameters, etc.) as our method to make a fair comparison with our method. The quantitative performance comparison of our method with the other methods is given in Table V. As shown in this table, our method outperforms all other methods across all datasets in both rank-1 accuracy and mAP evaluation metrics. For instance, the LAGA-Net outperforms the MBA-Net by 0.81% in rank-1 accuracy and 0.73% in mAP on HD dataset.
### TABLE IV
Quantitative performance comparison of our method (LAGA-Net) with existing state-of-the-art body-based person Re-ID methods on Market-1501 [19], DukeMTMC-Re-ID [20], CUHK03 [5] and MSMT17 [3] datasets. The results are shown in rank-1 accuracy (%) and mAP (%). Best and second best results are shown in red and blue, respectively. * denotes unsupervised person Re-ID methods.

| Method         | Market-1501 | DukeMTMC | CUHK03 (L) | MSMT17 |
|----------------|-------------|----------|------------|--------|
|                | rank-1 mAP  | rank-1 mAP | rank-1 mAP | rank-1 mAP |
| MGCAM [25]     | 83.79       | 74.33    | -          | -       |
| DGNet [6]      | 94.8        | 86.0     | 86.6       | 74.8    |
| Interpreter 50 [36] | 94.74 | 87.11    | 87.84      | 75.27   |
| ONSNet [37]   | 94.8        | 84.9     | 88.6       | 73.5    |
| MGN [10]      | 95.7        | 86.9     | 88.7       | 78.4    |
| PCB+RR [9]    | 93.8        | 81.6     | 83.3       | 69.2    |
| PDC [8]       | 84.14       | 63.41    | -          | -       |
| BagTicks [4]  | 94.5        | 85.9     | 86.4       | 76.4    |
| ABD-Net [12]  | 95.60       | 88.28    | 89.00      | 78.59   |
| RGA-Net [13]  | 96.10       | 88.40    | -          | 81.10   |
| MHN [11]      | 95.1        | 85.0     | **89.1**   | 77.2    |
| IANet [8]     | 94.4        | 83.1     | 87.1       | 73.4    |
| RLCC* [23]    | 90.8        | 77.7     | 83.2       | 69.2    |
| IICS* [24]    | 88.8        | 72.1     | 80.0       | 64.4    |
| LAGA-Net (Ours) | **96.18** | **88.76** | **89.71** | **78.92** |

### TABLE V
Quantitative performance comparison of our method (LAGA-Net) with existing state-of-the-art hand-based person Re-ID methods (GPA-Net [16], MBA-Net [39], RGA-Net [13] and ABD-Net [12]) on right dorsal (D-r) of 11k, left dorsal (D-l) of 11k, right palmar (P-r) of 11k, left palmar (P-l) of 11k and HD datasets. The results are shown in rank-1 accuracy (%) and mAP (%). Best and second best results are shown in red and blue, respectively.

| Method         | D-r of 11k | D-l of 11k | P-r of 11k | P-l of 11k | HD |
|----------------|------------|------------|------------|------------|----|
|                | rank-1 mAP | rank-1 mAP | rank-1 mAP | rank-1 mAP | rank-1 mAP |
| GPA-Net [16]   | 94.80      | 95.72     | 94.87      | 95.93      | -  |
| MBA-Net [39]   | 97.45      | 97.98     | **96.71**  | **97.41**  | 98.05 |
| RGA-Net [13]   | 94.77      | 95.67     | 95.30      | 95.98      | 92.66 |
| ABD-Net [12]   | 95.89      | 96.76     | 94.26      | 95.34      | 96.21 |
| LAGA-Net (Ours)| 97.56      | 98.11     | **96.82**  | **97.53**  | 98.16 |

instance, our method outperforms the RGA-Net by 5.50% in rank-1 accuracy and 4.96% in mAP on palmar right (P-r) of 11k dataset.

### F. Qualitative Re-ID Results
The qualitative Re-ID results of our proposed method is shown in Figs. 3 and 4. As can be observed on Fig. 3, our proposed method (LAGA-Net) has an improved performance on Market-1501 dataset over the baseline (the global without attention component of the LAGA-Net) in retrieval performance. While the LAGA-Net retrieves all top-5 correct results (top row for each query), the baseline only retrieves few results (bottom row for each query). This indicates that the LAGA-Net learns more robust and discriminative deep feature representations.

### V. Discussion
The proposed multi-branch deep network architecture, LAGA-Net, is a compound approach for end-to-end discriminative feature representations learning for person Re-ID based on both body and hand images. We have shown in the experiments section (see Section IV) that extensive evaluation of this method has demonstrated excellent performance not only on body datasets but also on hands datasets, as shown in Table IV and Table V, respectively.

The different branches in the proposed multi-branch deep network capture different but complementary information to boost the performance of the network as shown in ablation analysis section (see Section IV-D) on both body and hands datasets, particularly in Table III. Each branch has its own importance in the proposed network. The attention branches, channel attention branch and spatial attention branch, focus on the relevant features of the image while suppressing the irrelevant backgrounds. To maintain translation equivariance, relative positional encodings is integrated into spatial attention module of the spatial attention branch. The global and local branches intends to capture global context and fine-grained information, respectively. By properly integrating these branches...
in our proposed network, we have demonstrated that it is possible to effectively learn robust and discriminative feature representations for person Re-Id based on both body and hand images. The main importance of the proposed method is the excellent performance it has shown on both body and hand images. The elegant approach of integrating many different components into our proposed compound approach has demonstrated learning of robust and discriminative feature representations which is helpful in overcoming the many challenges the person Re-Id is facing such as pose variations, occlusion, view point changes, lighting changes, background clutter, noisy labels, etc.

The proposed method has various applications. For instance, the person Re-ID based on body can be used for intelligent video surveillance in less controlled or uncontrolled environments. The hand-based person Re-Id can be used for criminal investigation in uncontrolled environments, for instance, for recognizing or re-identifying perpetrators of serious crime such as sexual abuse in case only hand images of the perpetrators are available, which is very crucial in assisting international police forces.

VI. CONCLUSION

In this work, we introduce a compound approach for end-to-end discriminative deep feature learning, the Local-Aware Global Attention Network (LAGA-Net), for person Re-Id based on both body and hand images. The LAGA-Net is a multi-branch deep network architecture consisting of channel and spatial attention modules in branches in addition to global (without attention) component of the LAGA-Net (baseline). The green and red bounding boxes denote the correct and the wrong matches, respectively. Feature embeddings from our proposed method (LAGA-Net) give better retrieval performance.

Fig. 4. Some qualitative results of our method on 11k [21] and HD [22] datasets using query vs ranked results retrieved from gallery. From top to bottom row are right dorsal of 11k, left dorsal of 11k, right palmar of 11k, left palmar of 11k and HD datasets. The green and red bounding boxes denote the correct and the wrong matches, respectively.

REFERENCES

[1] Nathanael L. Baisa, “Occlusion-robust online multi-object visual tracking using a GM-PHD filter with CNN-based re-identification,” Journal of Visual Communication and Image Representation, vol. 80, pp. 103279, 2021.
[2] Ergys Ristani and Carlo Tomasi, “Features for multi-target multi-camera tracking and re-identification,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, June 18-22, 2018, pp. 6036–6046.
[3] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian, “Person transfer GAN to bridge domain gap for person re-identification,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
