IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 October 2022, accepted 11 November 2022, date of publication 18 November 2022,
date of current version 30 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3223426

RESEARCH ARTICLE

A Novel Approach Based on Multi-Level Bottleneck Attention Modules Using Self-Guided Dropblock for Person Re-Identification

MUHAMMAD USMAN YASEEN[1], MOUSTAFA M. NASRALLA[2], (Senior Member, IEEE), FAIZA ASLAM[1], SYED SOHAIB ALI[3], AND SOHAIB BIN ALTAF KHATTAK[2]

[1]Department of Computer Science, COMSATS University Islamabad, Islamabad 45550, Pakistan
[2]Department of Communications and Networks Engineering, Prince Sultan University, Riyadh 12435, Saudi Arabia
[3]CGG Services (UK) Ltd., RH10 9QN Crawley, U.K.

Corresponding author: Muhammad Usman Yaseen (muhammadusmanyaseen@gmail.com)

This work was supported by Prince Sultan University (for paying the Article Processing Charges (APC) of this publication).

ABSTRACT Person re-identification has inspired a lot of interest due to its significance in intelligent video surveillance. It is a difficult task due to the presence of critical challenges such as changes in appearance, misalignment, occlusion and background noise. Batch drop block layer (BDB) has been used recently in person re-identification by exploiting the feature erasing procedure. However, BDB drops a block of features randomly, resulting in the loss of contextual information, which makes the model difficult to train. Also, due to the random dropping of features, large area of discriminative information may lose during training, resulting in low efficiency and performance. To address this problem and to improve the model representation power, we propose a novel, lightweight, self-adaptive bottleneck attention module with a self-attention branch to improve the model performance by reducing the parameter overhead with negligible computation cost. The proposed approach entails bottleneck attention module (BAM) which is incorporated between ResNet layers to remove the background noise and to nominate the high-level semantic part. Further, dilated convolutions with batch normalization are used to tackle the contextual information loss problem and to avoid overfitting. In addition, an informative global branch is used which captures the global representation of the network, and the attention branch entails the multiscale local salient information. Two types of loss functions including softmax and batch hard triplet are used in the training process for each branch, forcing the network to encapsulate the common attribute within the similar identity and to maintain distance between distinct individuals. Compared with BDB, our network improves the mAp to 88.1%, and Rank-1 gets 96.3% for the market-1501 dataset. The results on Cuhk-03-Detected dataset showed 79.2% mAp score, with 81.4 %, Rank-1, whereas on Cuhk-03-labelled dataset, a mAP score of 81.3% and a Rank-1 score of 83.3% is achieved. Experiments reveal that ResNet model with addition of multiple BAM layers performs consistently over the state-of-the-art datasets using softmax and batch hard triplet loss.

INDEX TERMS Multiscale feature extraction, lightweight self-attention, bottleneck attention module.

I. INTRODUCTION

Person re-identification [1] aims to identify the pedestrians of interest captured by numerous non-overlapping cameras across different times. The objective is to identify the corresponding individual candidate across the probe image in the gallery setting, which is a challenging task in the computer vision community [2], [3], [4]. The re-identification process is utilized for the image retrieval task, and it returns a similar identity with a ranked list. In the re-identification process, if the location of the camera is known, then it becomes easier to track the person's location with the help of path tracing from one place to another. However, to follow the target person using various cameras, the individual's identity has to

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

FIGURE 1. A probe image finds the matching candidate in the gallery set.

be retrieved from the second camera based on the detailed information acquired from the first camera. Re-identification is relatively correlated with a recognition problem [5]. The target person's probing image is provided, and all of their associated identities in a gallery collection are searched. More precisely, the identification pipeline looks for a person in one camera and extracts a set of features using a deep learning model. When the same person walks through a different camera, the identification pipeline compares them to the learned features and identify them as a corresponding person. A similar scenario can be seen in Figure 1, where several people pass through different camera feeds; and a specific person present in the query image is identified in the probe images present in the gallery set.

In person re-identification, various problems are inherently challenging, like intra-class variation and inter-class variation. Intra-class variation problem deals with matching an individual across a particular scene. On the other hand, an inter-class variation problem deals with a situation in which a different person has similar appearance across camera views. Also, the appearance of people [6], [7] varies significantly due to the variation in viewpoint [8], illumination condition [9], [10] and due to the low resolution of images [9], [10], [11]. Recent studies revealed that the re-identification problem faces dissatisfying results due to the influence of self-occlusion [12], [13] pose variations [14], and background inference occlusion [11] in pedestrian images.

Deep neural networks (DNNs) appear to be promising in terms of feature extraction required for person re-identification. Feature extraction for person re-identification can be divided into two primary categories: Re-ID global feature representation, which leverages the global feature learning [15], [16], [17], [18], [19] and partial feature representation learning [20], [21], [22], [23], [24]. Global feature learning aims to identify the most relevant appearance clues in defining identities and distinguish them from others. Partial feature learning mainly considers the part of the feature instead of recognizing the full features of the person.

The existing deep neural networks performing person re-identification has two major shortcomings. Firstly, these systems solely employ single-level deep layer characteristics. Using multilevel features from different layers in deep learning methods is inherently tricky. The max-pooling technique causes feature maps from various levels to have varying sizes. Apart from that, a single-level feature might not be sufficient. Secondly, existing deep models are trained via a single loss function like softmax. Considerable intra-class variation and inter-class similarity across different views are limitations of softmax loss.

In deep neural networks, the drop block approach [25] proposed recently seemed to be promising to learn attentive local features and to extract rich features for person re-identification task. However, the random selection of dropping features looses a lot of background information and some important features as well that significantly impact the performance. Batch drop block (BDB) [25] is the modified technique of drop block which tries to overcome the limitation of drop block, but it still has some flaws and does not produce optimal/satisfactory results.

One of the major drawbacks is that the probability of discarding a feature at random drops the performance [26], and it should be small enough to ensure the convergence of the model during training. This makes it very hard to uncover more diverse features. The random dropping of a block of features makes the model difficult to train because the vast area of discrimination may be removed in this case. Also, the network is easily misguided to pay attention to the discriminative region if we do not have an explicit regularizer to drive the attention in feature learning. Therefore, carefully designing the dropping module could improve the model performance. Moreover, current deep learning models [10], [26], [27] rely on the deep layer's single-level feature while ignoring the shallow layer's detailed low-level part.

To cope with all these challenges and to improve the model's representation power, there is a need for an effective mechanism to preserve the contextual information because features at lower layers have a small receptive field due to the occlusion and background noise. In this context, we propose a novel, lightweight, self-adaptive bottleneck attention

module with a self-attention branch to improve the model performance. The proposed method is composed of four core components. Firstly, the bottleneck attention module BAM is incorporated between ResNet layers. This helps to remove the background or texture feature and nominate the high-level semantic part. The attention mechanism can regulate the weight of extracting significant aspects of pedestrians, enhancing the semantic information of a high-level feature map.

Secondly, the model extracts features with dilation factors of 4. To tackle the contextual information loss, dilated convolution is used. The dilated convolution gets intrinsic information sequences by expanding the receptive field size. It extracts more recognizable features and broadens the range of feature sets. The dilation value increases the receptive field, which is beneficial to preserve the contextual information. In this way, low-level features of all scales are combined to get the optimal results with the help of a dilation factor.

Thirdly, the informative global branch is introduced which captures the global representation of the network, and the attention branch entails the multiscale local salient information all are concatenated to a one-dimensional feature vector. Lastly, two types of loss functions including softmax and batch hard triplet are used in the training process for each branch, forcing the network to encapsulate the common attribute within the similar identity and keep distance between distinct individuals.

The contributions of the proposed person re-identification system are as follows:

- To solve the problem of occlusion, background or deformation, we propose a lightweight BAM model, which is integrated for multilevel feature representation after each Res-Convolution block.
- We adopted a bottleneck attention map to remove the background or texture feature and nominate the high-level semantic part. The attention mechanism regulates the weight of extracting significant aspects of pedestrians, enhancing the semantic information of high-level feature maps.
- In ResNet-50, for each block, dilated convolution is added to increase the receptive field while keeping the image resolution unchanged with multiple kernels of various sizes. This network can extract multiscale features of the image without the loss of image resolution.
- To provide the solution of scale variation and misalignment issue, a self-attention layer is incorporated to drop the mask efficiently, emphasizing on the most discriminative salient features considering a self-guided attention map instruction. It facilitates strong spatial semantic information for the foreground region and enhances the feature matching accuracy.

The rest of the paper is organized as follows: Section II reviews the state-of-the-art of person re-identification and identifies potential research gaps. Section III details the proposed person re-identification approach with in depth explanation of each component. Experimental setup is detailed in Section IV and Section V depicts the results obtained from the proposed approach with an analysis of the different parameters of the model. Section VI concludes the paper with a glimpse on the future work.

II. LITERATURE REVIEW

Extensive research has been conducted in recent years to improve accuracy and efficiency of person re-identification systems. We divide and describe existing research into four paradigms in terms of metric learning methods, hand-crafted feature learning, deep learning methods, and attention based methods.

A. METRIC LEARNING

Previous studies reported that metric learning [28], [29] used the distance metric methods to compute the matching score between the pair of images. The research has been conducted in two phases, Unsupervised learning [30], [31], [32] and Supervised-learning [33], [34], [35]. Supervised-learning employed with labelled images to compute the distance metric function. Metric learning methods involve offline training where data is given in positive (same person in two different cameras) and negative (different person in different cameras) pairs. On the other hand, discriminative methods are trained online in real-time. Metric learning methods that improve feature representation do so through direct appearance modelling [28] or indirect appearance modelling through feature mapping [21]. Some techniques improve matching performance using the distance metric learning approach.

The metric learning task relies on distance metric learning, which reduces the intra-class distance between the actual sample of positive image pairs while extending the inter-class distance between an opposing pair of images [9]. After that, the discriminative ranking method was employed to optimize the distance between a couple of pictures [10]. Subsequently, the author formulates the statistics approved relative distance-based metric learning approach for the correct matching of a pair of images [11].

The metric learning method can improve discrimination by projecting features into a subspace with reduced intra-class distance and increased inter-class distance by introducing a margin between intra-class and inter-class. When combining feature extraction with metric learning, there are still certain limitations. The feature extraction method, in general, is still unable to adapt adaptively to the needs of metric learning. Two of the components are still distinct. Therefore, when individuals are subjected to such significant changes in their environment, these pre-designed features may not distinguish between persons who seem identical. Person recognition systems are separated due to the lack of interaction between feature extraction and metric learning. There is limited interactivity between feature extraction and metric learning, which characterizes the person identification task [21].

FIGURE 2. Possible challenges in person re-identification.

B. HAND-CRAFTED FEATURE LEARNING

Authors [36] proposed a conventional system that uses hand-crafted features for person re-identification. A multi-scale CNN approach was developed to identify the detail of a person across multiple overlapping cameras and optimized the network parameter efficiently under various illumination variations, occlusions and pose variation scenarios. Partial occlusion, variation in viewpoint, and misalignment are considered as the most common challenges in the person identification domain. A spatial channel pipeline was constructed to deal with partial occlusion. This methodology employs critical local and global contextual information to create a person's discriminative look in the event of occlusion and misalignment [37].

Authors in [60] proposed an Ad-hoc feature vector that incorporates head and shoulder anthropometric texture of feature information for person reidentification. They exploited three depth feature vectors and three-intensity feature vectors of various positions including front overview, overhead view, and leave view from the top view images. In this way, depth and intensity information was collected, which increased the robustness of the proposed method against lightning conditions.

The proposed model provided satisfactory results over the small training samples using similarity loss. By merging local and global features in one branch diminished the detailed information. We may add or remove the local branches in a multi granularity network. The hand-crafted feature learning has its own limitations; when individuals are subjected to such drastic changes in their surroundings, these pre-designed traits may be unable to discriminate between people who appear to be identical.

C. DEEP LEARNING

Various deep learning methods [38], [39], [40] were exploited to extract the discriminative features in order to represent the person's appearance and to overcome possible challenges involved in person re-identification including illumination changes, occulusion, and view point variation as shown in Figure 2. Currently, most of the deep learning approaches take benefit from the state of the art deep architectures including ResNet-50 [41],DenseNet-201 [42], GoogleNet, and VGG-Net [43] to extract important features for person identification. An improved deep learning Siamese architecture proposed by [44] explores two novel layers; i.e., cross-neighborhood difference layer and followed layer after that across patch difference, which is calculated using a softmax function to check the similarity of a pair of images.

In another study [45], a novel LoopNet architecture was proposed for person identification with the most brutal sample mining techniques based on the listwise ranking. Multiple loss was introduced to resolve the issues of the listwise ranking approach. The proposed model achieved the best results by global hard sample mining and semihard sample mining in a listwise ranking model instead of computing the mining sample locally in mini-batches. The authors, in [45] proposed a multi granularity network (MGN) technique that concatenates on a single branch's local and global features. The gradient-based method was used in this paper to improve matching accuracy for the person identification problem.

Domain generalizability count as a significant challenging issue in the person reidentification task. [46] constructed a learnable voting network that is the modified version of a meta-learning process trained over the alignment loss to cope with the domain generalizability issue. Relevance aware mixture of experts (RaMoE) algorithm was used to receive the complementary detailed information from the source domain and then forwarded to the destination domain. Reference [47] proposed an approach based on the decorrelation loss to preserve the diverse and discriminatory features of the source domains. The proposed scheme adapts the source domain features and then aggregates the feature to boost the model generalizability in the target domain.

Neural architectural search(NAS) was proposed in [48] based on attention module to learn the spatial and channel attention feature map. These two feature maps were combined to improve the feature representative ability without pretraining the model. It is a proven fact that attention effectively deals with inference, background changes, misalignment of body parts among the pedestrian feature map.

However, attention search space(ASS) with a hybrid optimization scheme determines where the attention should be placed in the re-identification module to improve efficiency [48].

The majority of deep learning methods have covered the drawbacks of hand crafted feature learning, but still, it has two shortcomings. On the one hand, these systems solely employ single-level deep layer characteristics. Using multi-level features from different layers in deep learning methods is inherently tricky. The max-pooling technique causes feature maps from various levels to have varying sizes. Apart from that, a single-level feature might not be sufficient. The softmax loss function is typically used in deep learning methods. Their performance is adequate, but there are some areas where they may improve, such as intraclass and interclass distance. The single loss function, on either hand, is insufficient for the person reidentification task.

D. ATTENTION LEARNING

Attention module was designed to extract local features and global feature to preserve the identity appearance, shape, and pose representation which helped to match the image with the target person. For this purpose, semantic consistency loss retains the semantic information between the conditional image and the generated pose image [49]. Attention plays a critical role in both aspects of channel-wise attention and spatial dimensions, where the emphasis is on extracting discriminative features for efficient feature representation.

The attentive discriminative feature Learning (ADFL) module proposed in [50] focuses on attention and includes a skip connection to improve model adaption and generalizability. Only the source domain was used to train the model. ADFL strategy effectively performs in the cross-domain, employing attention module. Spectral normalization adopted for the training process has less computation cost, and there is no need for extra hyperparameter to tune the model [34].

To extract an informative localized region from the input image, an informative attention-based algorithm was designed, composed of 2 subnetworks running in parallel: channel and spatial attention pipeline [39]. However, channel-wise attention mainly relies on obtaining the most informative part from the given input image. Spatial attention considers the positional information; it decides where the significant area is prominent in the input image.

Occlusion is the most common challenge present in the person re-identification domain. It is crucial to create a suitable framework for obtaining a distinguishing feature from the non-occluded area. The transformer-aware technique is employed for the occluded person identification problem. It is composed of a pixel-wise encoder transformer and a prototype-based decoder transformer approach.

The salience weekend approach and five attention branches are exploited for efficient feature refinement [51] to get the desperate local features by removing background details. The proposed technique provides the stability of the network and extracts entire useful features by the salience weakening method instead of erasing them directly. It would be best practice to incorporate the salience weekend approach with the temporal attention framework.

In the person re-identification domain, ResNet does not assist in identifying the exclusive feature of the person, but still, it precludes the background noise information. Attention emerges with the multi-branch network to address the above problem. It adopts a filter network to reduce the background information and encourage the model to acquire the exclusive feature of a person. The addition of an attention mechanism increased the number of parameters and training time [38].

The Self Attention and Channel Attention approaches were combined into a unified framework to cope with the misalignment issue and to reduce background noise and occlusion. The feature representation is differentiated using multiple classifiers, and the similarity score is improved along with self-attention. Self-channel attention facilitates the matching score to address the misalignment, occlusion, and noise challenges using salient feature representation and strong feature representation.

III. PROPOSED APPROACH FOR PERSON RE-IDENTIFICATION

In this section, we present the detail of our proposed person re-identification system. The proposed system is based on a light-weighted BAM network, which is applied on each ResNet50 stage and induces the model to learn the entire region of the object. BAM generates the self-attention map from the input feature map and produces the drop mask and drop map. Both have different roles and are computed via self-attention maps. The drop mask penalizes the most discriminating portion for inducing the pattern to cover a significant part of the object. Moreover, the self-attention map extracts the most discriminative region to increase the discriminatory power of the model. During training, drop mask is chosen under the guidance of a self-attention map for each iteration. Thus, the selected one is applied over the input feature map. In addition, it does not provide any trainable parameter when it is implemented over the multiple feature map simultaneously. Furthermore, with the BAM approach, the most discriminative region is identified by viewing the lower level detailed information and erased efficiently. However, it analyses that BAM generates negligible overheads and efficient performance.

The flow of the proposed system is depicted in Figure 3. The input images produce a feature map with three dimensions, i.e., channels, Height and width as represented in (1).

$$\mathcal{F} \in \mathbf{F}^{C \times W \times H} \tag{1}$$

The lightweight BAM network is applied to each ResNet-50 stage and causes the model to learn the entire object region by extracting the most efficient feature from each layer using BAM. As indicated in Figure 3, the feature maps are transferred to ResNet-stage 1 which are then further passed to dilation block. After that, spatial multiplication is performed using the BAM architecture. The output of the first feature

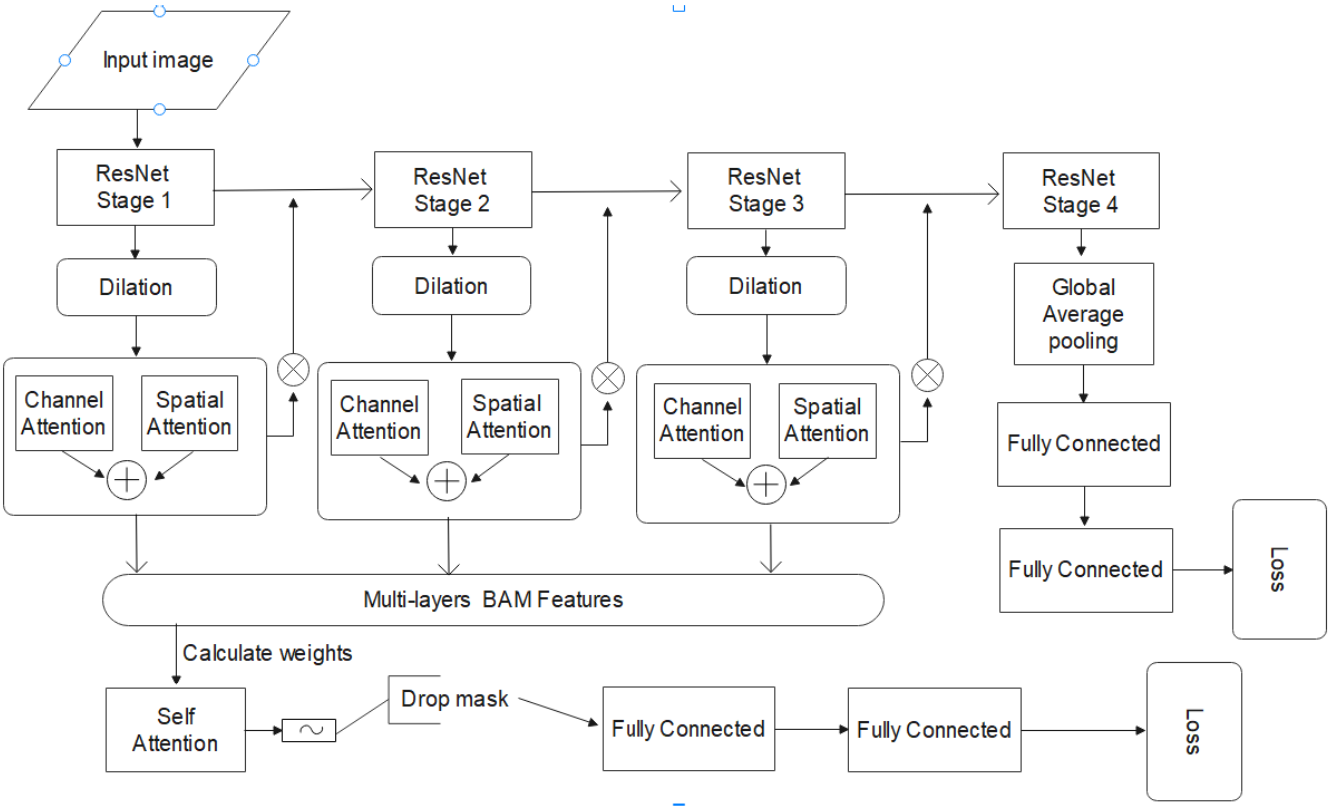

FIGURE 3. Workflow of the proposed person re-identification system.

map is transferred as an input to the second ResNet-stage2, in which the same procedure is applied as the ResNet-stage1. This process continues till ResNet-stage4. The output of each ResNet layer produces feature attention map1, feature attention map2, feature attention map3 and feature attention map4 with a dilation factor of 4. All of these are then concatenated into the multilevel feature attention maps (6).

$$\mathcal{F}_{att}(map1) = (DCONVF1(\mathcal{F}_C + \mathcal{F}_S) \quad (2)$$

$$\mathcal{F}_{att}(map2) = (DCONVF2(\mathcal{F}_C + \mathcal{F}_S) \otimes F_{att}(map1)) \quad (3)$$

$$\mathcal{F}_{att}(map3) = (DCONVF3(\mathcal{F}_C + \mathcal{F}_S) \otimes F_{att}(map2)) \quad (4)$$

$$\mathcal{F}_{att}(map4) = (DCONVF4(\mathcal{F}_C + \mathcal{F}_S) \otimes F_{att}(map3)) \quad (5)$$

$$\mathcal{M}_{sf}(map) = F_{att}(map1) + F_{att}(map2) + F_{att}(map3) + F_{att}(map4) \quad (6)$$

After that, the input passes through the two branches, i.e., global branch and multilayer Attention branch. The global branch facilitates global feature representation via global average pooling, and the output from this branch is passed to two fully connected layers. The first fully connected layer generates the feature vector Fc1, whose dimensions are reduced in the Fc2. After that, the loss is calculated separately for each branch.

The multilayer attention branch applies the pooling operation instead of max pooling on the multilevel attention feature, and the output is passed as an input to the self-attention. The self attention calculates the weights of feature maps and produces the drop mask via threshold factor, which helps to learn the local attentive feature robustly. The drop mask penalizes the most discriminating portion for inducing the pattern to cover a significant part of the object.

The self-attention map extracts the most discriminative region to increase the discriminatory power of the model. During training, a drop-mask is chosen under the guidance of a self-attention map for each iteration. Thus, the selected one is applied over the input feature map. In addition, it does not provide any trainable parameter when it is implemented over the multiple feature map simultaneously. The BAM approach helps to identify the most discriminative region by viewing the lower level detailed information. This means that BAM generates negligible overheads due to the reduction value of r = 16, which increases the receptive field at less parameter with efficient performance.

A. ResNet-50 MODEL

We adopted the modified version of ResNet 50 model because of its competitive performance in recent peron re-identification systems. It is the most common concise architecture with relatively negligible overhead compared to the dense architecture, which increases the model complexity

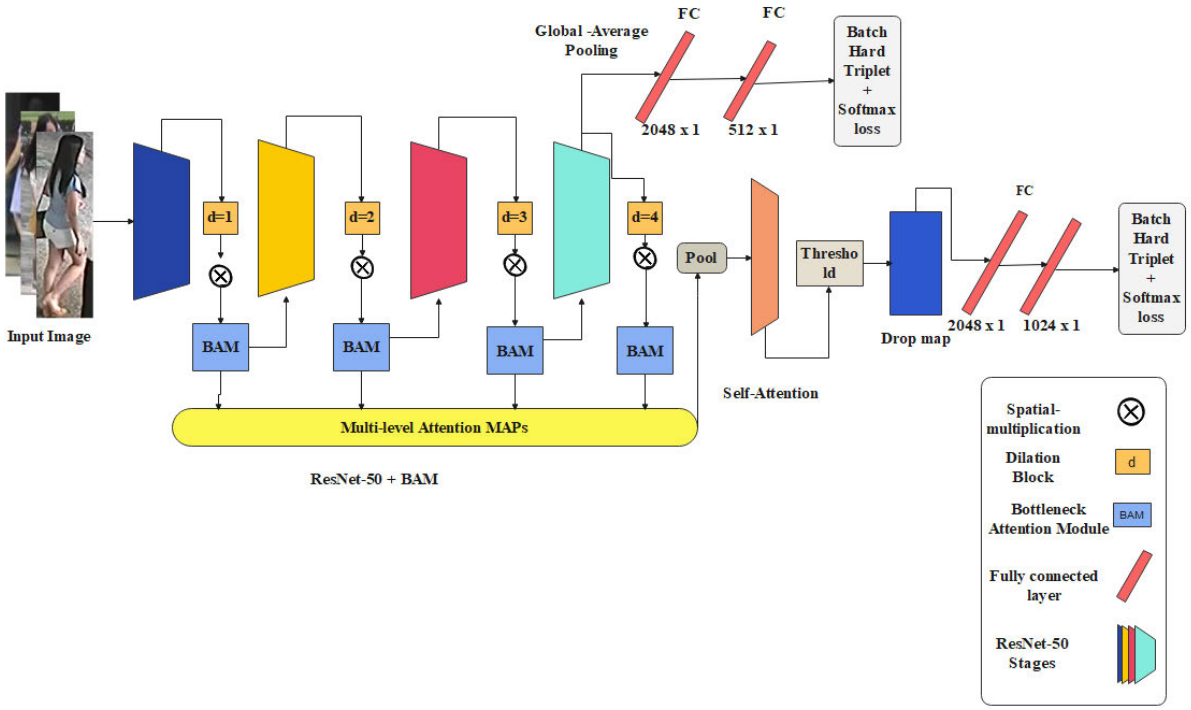

FIGURE 4. Architecture of the proposed person re-identification system.

due to the deep layers. ResNet50 model consists of four Residual-Convolution blocks, stage1, stage2, stage3, and stage 4, as illustrated in Figure 4. The ResNet 50 architecture as proposed in earlier studies requires four stages to efficiently perform identity convolution and other operations on the input image. These stages are required to perform the initial convolution and max-pooling operations with different kernel sizes. The proposed modification in the original ResNet50 block is as follows: Firstly, we remove the down sampling operation in the fourth ResNet block to preserve the large area of the receptive field to enable the local detail of features or body parts. Secondly, the bottleneck attention module (BAM) is integrated after each Res-Convolution block to achieve the multi-scale information using dilatation. It may focus on the salient part, considering the detailed content of each image. And the incorporation of the bottleneck module after every stage of Res-Conv-block constitutes a BAM.

The pooling operation is performed after each BAM block to combine the attention feature and to obtain the final person feature. The proposed system extracts the multi-level detailed information using the bottleneck attention module and generates the self-attention map. More specifically, after each stage of the ResNet-50 model, the bottleneck module is incorporated with a dilated convolution layer. Then the features of all stages of the ResNet block are pooled using spatial-wise multiplication and concatenated into the final feature map. After that, self-attention maps are produced with the help of channel-wise pooling from these previous layers.

B. BOTTLENECK ATTENTION MODEL

Inspired by the studies of [52] and [53], we adopt the BAM, which diagnoses low-level features such as background texture feature at an early stage. It usually focuses on the exact target, which has high-level semantic information. To highlight the local detailed information of pedestrian images, an efficient BAM-based module is designed to erase the most discriminative part of the feature map. BAM is a self-contained adaptive module that dynamically suppresses or erases the feature map through the attention module.

$$\mathcal{F} \in \mathbf{R}^{W\times H\times C} \tag{7}$$

$$\mathcal{F}_{att}(map) \in \mathbf{R}^{W\times H\times C} \tag{8}$$

$$\mathcal{F}' = (F + \mathcal{F} \otimes F_{att}(map)) \tag{9}$$

The proposed system dramatically reduces the parameter overhead compared to the pyramid approach utilized earlier for the re-identification task [50]. BAM is incorporated in the bottleneck before performing the down sampling operation. In this case, only global average pooling (GAP) was used to get the statistics on the feature map in spatial and channel dimension, whereas CBAM also considers using the max pooling and average pooling. Max pooling generates the most salient features from the feature map and compensates the GAP output, which encodes global statistics softly. In the case of BAM, Convolution operation is performed using a dilation value of 4 to increase the receptive field. At the depth of the network, the CBAM uses a large filter size, and typically a convolution layer is used with d = 1 to incorporate the same procedure.

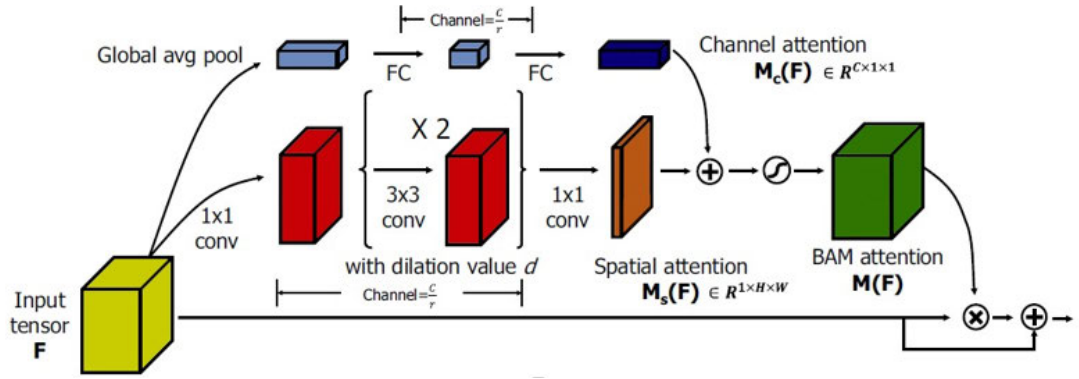

FIGURE 5. Architecture of bottleneck attention module.

Global average pooling tends to identify the whole extent of an object and forces the network to identify most of the discriminative parts. On the other hand, global max pooling focuses on only one discriminative part. As per details of prior work by Zhou et al. [54], when averaging of a map is performed in global average pooling, the resultant value is maximized by finding all discriminative parts of an object as all low activations reduce the output of the particular map. In case of global max pooling, since only max operation is being performed, the low scores of all parts of the image except the most discriminative one do not impact the score.

In BAM, spatial and channel attention maps are generated in parallel, which are later added to produce the final attention map. CBAM also uses a similar approach; first, channel-wise attention is placed, and then the spatial attention module is added. BAM contains two types of attention, i.e., channel-wise attention and spatial-wise attention. Both have their own set of objectives.

The channel-wise attention is used to estimate the most important property of the target item. In channel attention, the spatial axis-wise GAP is squeezed, and then we regress the channel attention using two fully connected layers. Spatial attention mainly chooses the necessary spatial region, rather than distributing the part of the image equally. And it significantly reduces the dimension of channel-attention. In our architecture, we designed BAM, which consists of two kinds of attention modules, i.e., spatial attention and channel-wise attention. Both modules have different functionality, the channel-wise attention module focuses on key regions against each channel, whereas spatial attention focuses on spatial attributes of an image. In each channel, to aggregate the feature maps, a feature vector to encode global information in each channel

$$\mathcal{M}_{Catt}(\mathcal{F}) \in \mathbf{R}^{C} \tag{10}$$

is produced after performing global average pooling operation on feature maps (10). Moreover, a multi-layer perceptron with one hidden layer is used to estimate attention across channels. The output size of the MLP layer is set to C /r, where r is the reduction ratio to reduce the number of parameters. For scale adjustment, the batch normalization operator is used with spatial attention output as represented in (11).

$$\mathcal{M}_{Catt}(\mathcal{F}) = \mathcal{BN}(mlp(Ap(\mathcal{F})) \tag{11}$$

For channel attention, spatial axis-wise GAP is squeezed, and then channel attention is regressed using two fully connected layers. The necessary spatial regions are chosen in spatial attention, rather than distributing the part of the image equally. The spatial attention module pays attention to the image's position information, allowing the model to determine which feature maps have more spatial weight.

$$\mathcal{M}_{Spatt}(\mathcal{F}) \in \mathbf{R}^{W\times H} \tag{12}$$

$$\mathcal{M}_{patt}(\mathcal{F}) = \mathcal{BN}(conv_3^{1\times1}(conv_2^{3\times3}(conv_0^{1\times1}(\mathcal{F})))) \tag{13}$$

From the four convolutional layers, two-layers has a convolution kernel of size 1×1 to minimize the dimension of feature maps as represented in (13). Using the convolution of size 1×1 with channels, the input tensor $F \in SC$ results in a reduced dimension map $s(F) \in S$, and using the contextual information effectively two dilated convolutions are performed.

$$\mathcal{F}_{att}(map) = sig(\mathcal{M}_{Catt}(\mathcal{F}) + \mathcal{M}_{Spatt}(\mathcal{F})) \tag{14}$$

$$\mathcal{BN}(W_1(W_0 Ap(\mathcal{F}) + b_0) + b_1)$$

$$where: W_0 \in \mathbf{R}^{\frac{C}{R}\times C}, b_0 \in \mathbf{R}^{\frac{C}{R}}, W_1 \in \mathbf{R}^{C\times\frac{C}{R}}, b_1 \in \mathbf{R}^{C} \tag{15}$$

C. GLOBAL BRANCH

Global features mainly consider the primary body part and ignore other features such as feet and waist, while the local branch prefers particular points. The global branch is used to embed the global feature representation. It also supervises the feature dropping branch's training and generates the self drop block layer, which is applied to the well-learned feature map. The final compact feature vector of the global branch is of size 2048×1 which is reduced to 512×1 dimension of the feature vector.

D. DILATION LAYER

Convolution with arbitrary kernel size is known as dilation convolution. The idea of dilated convolution is to increase the input space or gap with a dilation factor.The benefit of expanding the receptive field by expanding the kernel is that it allows us to receive intrinsic information at multiple spatial scales without increasing the parameter cost. Intrinsic sequence information can be captured initially using dilated convolution by expanding the receptive field to resolve the problem of loss of contextual information. It extracts more recognizable features and broadens the range of feature sets.

To tackle the contextual information loss, dilated convolution with dilation factor of 4 is used. The dilated convolution gets intrinsic information sequences by expanding the receptive field size. It extracts more recognizable features and broadens the range of feature sets. The dilation value increases the receptive field, which is beneficial to preserve the contextual information. In this way, low-level features of all scales are combined to get the optimal results with the help of a dilation factor.

Park et al. [52] performed multiple ablation studies with different dilation rates of 1, 2, 4 and 6, based on the ResNet50 architecture. The dilation value determines the sizes of receptive fields in the spatial attention branch. It was revealed that the performance improves with larger dilation values, though it is saturated at the dilation value of 4. This phenomenon can be interpreted in terms of contextual reasoning. As mentioned earlier that dilated convolutions results in an exponential expansion of the receptive field in the spatial attention branch which enables the proposed system to aggregate contextual information. It is also to be noted that the dilation value of 1 is equivalent to standard convolution operation and results in low accuracy. The dilation value of 2 means skipping one pixel per input and the dilation value of 4 means skipping 3 pixels. This demonstrates the effectiveness of a context-prior for inferring the spatial attention map.

E. SELF-ATTENTION MAP

The self-attention model operates directly on the feature map, which finds the pixel-based contextual information. In that sense, every pixel in the feature map has a corresponding weight or value. The knowledge of the feature map is determined by the related weight of the pixel point. However, the weight of a feature determines how the corresponding feature point affects the overall task. It solves the problem of over-reliance on local features by combining the diversity of global and local features.

Self-attention [55], [56] focuses on the global correlation and considers the global information, just complementary to local correlation. Self-attention additively computes the correlation of each position on the feature map and mainly focuses on the essential discriminative parts such as cloth hair and bag based on the global correlation of the original map. So the noise from the background will be the weekend.

The Self-attention layer is incorporated to keep the complete fledge information of the entire image and to obtain a pixel context feature map; after that, similarity is calculated between the feature maps, which enables the background clutter problem to be addressed robustly. We computed the multiscale feature vector with the concatenation of BAM based residual dilated convolution layers. The output of the multiscale feature map passes to the self-attention layer, which calculates the average weight metrics and applies the threshold over the attention score to generate the drop mask.

However, the drop mask is also generated to lessen the background effect of information. This attention mechanism extracts the foreground region from the given feature map by calculating the weight of the attention map from different layers. Self-adaptive threshold-based drop block techniques motivated from [57] are adopted that erased the selected random feature guided by the self-attention mechanism or their guided region. Therefore, different dropping ratios are utilized to achieve the desire results rather than rely only on the horizontal stripes. To ensure the channel-wise correlation among the features, self-attention was introduced to identify the correct feature map. Self-attention framework enhance the feature matching score to find out the similar region in different image locations as described in Algorithm 1.

F. LOSS FUNCTION

In general, the performance of a person identification system is determined by the dataset's nature and the outcome of loss functions. The objective of the loss function is to ensure that the images whose attributes are close to each other should have a small distance between them [33]. On the other hand, the pedestrian images with different features keep the space more prominent between them in the re-identification task to measure the similarity.

Several loss functions have contributed to metric learning to beat the performance of image retrieval, like contrastive loss [12], triplet loss [13], [14], quadruplet loss [26], and batch hard triplet mining [22] are exploited to optimize metric learning. Triplet loss is suitable for the metric learning task. However, it is optimized to achieve the performance of various loss functions. With triplet losses, high computational efficiency was attained in the context of visual attention [58]. The literature has demonstrated that efficient sampling of data by selecting complex samples improves the performance of the proposed architecture.

We have adopted an offline hard mining technique [28] to train the network. The feature vector from the global and attention branch is combined to form the final embedding feature for person identification. For the re-identification problem, triplet loss [22] significantly increased the network's performance and has great potential to achieve the desired results. We verify our proposed scheme lightweight self-adaptive bottleneck attention module network on the metric learning loss, which combines softmax loss and soft margin batch hard triplet loss as represented in (16). Each

branch in the network separately computes the loss function. The objective of this loss is to increase The matching score between the probe image and the target image, as well as the distance between the anchor and the positive point of the image, are minimized, while the distance between the anchor and the negative point of the image is increased.

$$\mathcal{L}\wr\int\int = \mathcal{L}_{softmax} + \mathcal{B}.\mathcal{H}L_{Triplet} \tag{16}$$

- Batch Hard Triplet Loss: Considering the drawback of considerable training time [12], it is necessary to mine the hard triplet. We adopted the typical strategy used in [13]. We randomly selected the 5,000 samples of images for each epoch and computed the corresponding bottleneck feature vector to calculate the pairwise dissimilarities. Then for each of the 5,000 query images, we randomly selected a positive sample among three ones with an enormous discrepancy and the negative example among the ten ones that have the less dissimilarity as represented in (17). This is the simplest way to reduce the computational overheads by employing the efficient strategy, which takes less than 30 sec.

$$\mathcal{B}.\mathcal{H}L_{Triplet}(0, X) = \begin{cases} +maxE_D(\mathcal{F}_\theta(1_a), \mathcal{F}_\theta(ip)) \\ -minE_D(\mathcal{F}_\theta(i_a), \mathcal{F}_\theta(i_n)) \end{cases} \tag{17}$$

- Softmax Loss: In an attempt to improve generalizability, identification loss(softmax) represented in (18) is used, which aids in learning representative features. These representative features, express common characteristics of the same person from different scene view [20]. The softmax loss attempts to divide the embedding space into distinct subspaces using a hyperplane feature vector. Further, Label smoothing regularization is used, which is an excellent approach in the person identification domain to remove the overfitting problem [15] in the classification task and to achieve adequate performance.

$$\mathcal{L}_{softmax} = \sum_{i=1}^{j} \rho_i \log(q_i) \tag{18}$$

IV. EXPERIMENTAL SETUP

This section details the experimental setup used to implement the proposed person re-identification system. We first describe the datasets used in this study, then the evaluation parameters used to evaluate the proposed system. We also describe the parameter configuration of the model used in this study.

A. DATASETS

From the state of the art, it has been observed that for the person identification problem, some datasets are widely used for image retrieval tasks such as Market-1501 [22], Cuhk-03 [59], and Duke MTMC [46].

- Market-1501: The Market-1501 dataset comprises 32,668 images that are split into two parts, training and testing set. All the images were captured by six different overlapping cameras. The training set has 751 person or identities, which includes 12936 images, and the test set has 750 people, which contains 19732 images.
- Cuhk-03: The Cuhk-03 dataset is gathered at a Chinese University using ten multiple cameras of Hong Kong City. We divide the dataset into a training set with 767 identities and a test set of 700 identities. So, a total of 1467 people has comprised of 14097 images in this dataset. Dataset is partitioned into 767 identities for training and 700 identities for testing. The dataset's labeled specification comprises training instances 7,368; gallery images include 5,328, and 1400 query images from the test set. The detected dataset has 7365 images from the training set, 5332 images from the gallery set, and 1400 images from the testing set.

B. EVALUATION CRITERIA

To evaluate the proposed deep learning model, various assessment metrics have been utilized that vary to the corresponding problem. Deep learning used different performance evaluation parameters for image retrieval tasks over the benchmark publically available datasets. For the person re-identification problem, four primary metrics are adopted, including Commutative Matching Characteristics (CMC), Precision, Mean Average Precision, and Top1 accuracy.

- CMC precludes Rank-n, which presents the similarity score equivalent to the number of correct matching probes divided by the total number of the probe. Mostly rank1, rank5, rank10 are the most commonly utilized method to visualize the performance of the proposed model.
- Precision defines a specified threshold value of rank K. In this case, only the number of correct matches of the probe selected from the top K rank and the below Rank k values from the threshold is ignored.
- MAP can be defined as we check the matching correspondence against each query image from the gallery images. If the correct matching probe never gets retrieved, precision correspondence to the gallery image is zero. These metrics are designed to validate the prediction of the proposed model. Some researchers used individual metrics to evaluate the performance, and some used a combination of metrics.
- Top1 accuracy is the conventional accuracy of the first retrieval object, but multiple authentic images will be recovered in a person recognition task.

C. PARAMETER CONFIGURATION

Our proposed model is implemented in the PyTorch framework. The parameter configuration is summarized in Table 1. We performed the training experiment in google colab. Experiments evaluated with 4 × GTX-1080 GPUs. We used the pre-trained model ResNet50 as a backbone network in our

TABLE 1. Parameter configuration.

Parameters	Rate
Batch size	128 Training/Testing
Epochs	800
Learning rate	0.0001
Gamma	0.1
Image size	384 * 128
Optimizer	Adam
Weight decay	0.0005
Loss	Batch hard triplet + Softmax
Margin	none
Workers	4
Mode	Retrieval

proposed scheme, with an inclusion of BAM layer after each ResNet block. In the training phase, self attention based feature dropping branch drop the feature map via threshold the drop mask. In the testing stage, all the features are combined from both branches as the final embedding of the feature vector of the input image.

1) Pre-processing: In the training phase, the input image is resized to 384 × 128 to capture the detailed information from the pedestrian images and ten paddings. For the data augmentation step, the selected resized image is flipped horizontally and vertically. Right-left image flipping is also utilized in the testing phase. By default, the testing images are resized to the same training phase 384 × 128 with normalization.
2) Batch Generation: Our proposed method is trained over the mini-batches that randomly sampled p identities by selecting k images. For each person, we then test k images. Every picture in the training set fulfills the desired requirement of triplet loss. For example, if the P=32 and K=4, the batch size is 128 used for the model training. Each identity contains four instances of person images. We then incorporated the batch hard triplet loss with softmax loss. We used the Adam optimizer with B=0.9 and B2 = 0.99. Our model is trained over 800 epochs in total. Learning rate decay with a parameter 0.1, and at the early stage, it keeps 3.5 epower-5. The dimension of the fully connected layer to enlist the person feature is fixed set 1024.
3) Loss Function: In the training time, we adopted the default setting of parameter and hyperparameter and resized the image to 384 × 128 during the training and testing time. Data augmentation is applied with horizontal flips randomly which is followed by normalization step. The baseline architecture uses the batch hard triplet loss and softmax loss, respectively. The performance of triplet loss is similar to the classification loss when data is significant.

TABLE 2. Result on market-1501 dataset.

Method	mAp	Rank-1	Rank-5
BDB	86.3	94.6	97.4
SaTDB	86.7	95.2	97.9
OsNet	85.8	94.9	98.3
RBMF	88.5	92.5	98.5
Proposed Model	88.7	96.3	97.3

TABLE 3. Result on Cuhk-03-detected dataset.

Method	mAp	Rank-1	Rank-5
BDB	73.0	76.0	88.8
SaTDB	76.2	83.2	84.6
OsNet	67.8	72.3	74.3
RBMF	74.3	76.7	79.8
Proposed Model	79.2	81.4	84.6

V. RESULTS AND ANALYSIS

This section details the results of our proposed person re-identification system. Results reveal that BAM is more effective in terms of parameter overhead or accuracy trade off as compared to state of the art. It means that the proposed network achieves better accuracy with little overhead. It usually seems that deeper networks with significant parameters have achieved better results.

Although, BAM added few extra layers to the architecture, but with negligible overhead. Extensive experiments showed that the proposed system based on BAM increases the accuracy and performance and still has less overhead than naively putting extra layers in the network. The improvement is not merely due to the increased depth, but also because of the feature refinement.

1) RESULTS ON MARKET-1501 DATASET

The results of the proposed system on market-1501 dataset are depicted in Table 2. All experiments are performed in a single probe setting. It is evident from the table that the proposed method outperforms all the other schemes by a large margin. More precisely, the proposed system achieved 88.7% mAp and 96.3% Rank-1 results and outperformed most of the state-of-the-art algorithms. The addition of BAM layers in the ResNet architecture along with the employment of dilation mechanism played an important role to achieve these performance improvements. The improvement from the benchmark scheme is around 4.1% in a mAp and 2.1 % in Rank-1. Compared with BDB [60], the proposed system improves the mAp from 86.3% to 88.7% and Rank-1 improves from 94.6% to 96.3%. Hence, it verifies that the efficiency of our proposed model with comparative approaches.

2) RESULT ON CUHK-03 DETECTED AND LABELED DATASET

We further evaluate our model on Cuhk-03 Detected dataset. As evident from Table 3, the proposed network achieves an

TABLE 4. Result on Cuhk-03-labeled dataset.

Method	mAp	Rank-1	Rank-5
BDB	76.0	79.0	89.8
SaTDB	80.0	79.3	81.6
OsNet	68.8	69.9	70.3
RBMF	78.3	81.1	83.3
Proposed Model	81.3	83.3	85.1

TABLE 5. Comparison between the methods against model complexity parameters.

Method	Parameters	Time (hrs)
BDB	30643420	6.3
SaTDB	34.8 M	7.24
Top-DB	23,508,032	2.6692
OsNet	2,193,616	0.9789
DDB	49,846,720	7.4425
Proposed Model	32.92M	6.3

accuracy map of 79.2 % and the rank score is 81.4% and 84.6% against Rank-1 and Rank-5 respectively. Similarly, it can be observed from Table 4, that the proposed system outperforms state of the art on Cuhk-03 Labeled dataset as well. It achieved 81.3 % mAp, and 83.3% and 85.1% accuracy against Rank-1 and Rank-5 respectively. Further, we verify the effectiveness of our proposed network as a comparison with BDB[32], which also employs the dropping block strategy. The proposed system improved the map accuracy on Cuhk-03 detected dataset from 73% -79%, and the Cuhk-03 Labeled dataset improved the 76.0% to 81.3% accuracy map.

It is also interesting to observe that the proposed model has less number of parameters as compared to most of the existing approaches. As a result of which the overall execution time is also lower than compared to others which is evident in Table 5.

3) ANALYSIS AND VISUALIZATION

To further analyze the performance of proposed person re-identification system, some sample results and their visualization is illustrated in this section. In addition, to show the significance of our proposed scheme, various techniques are examined critically to see the behavior of images. As shown in Figure 6, different activation attention maps compute the discriminative region by extracting the feature of the image followed by various techniques that are well suited and unique in terms of their functionality. Few techniques such as BAM, Self-attention BAM, and Drop mask generated by the self BAM can be visualized in Figure 7. Therefore, attention maps are developed to find the most discriminative part of the target image to improve feature learning from the existing approach BDB [60].

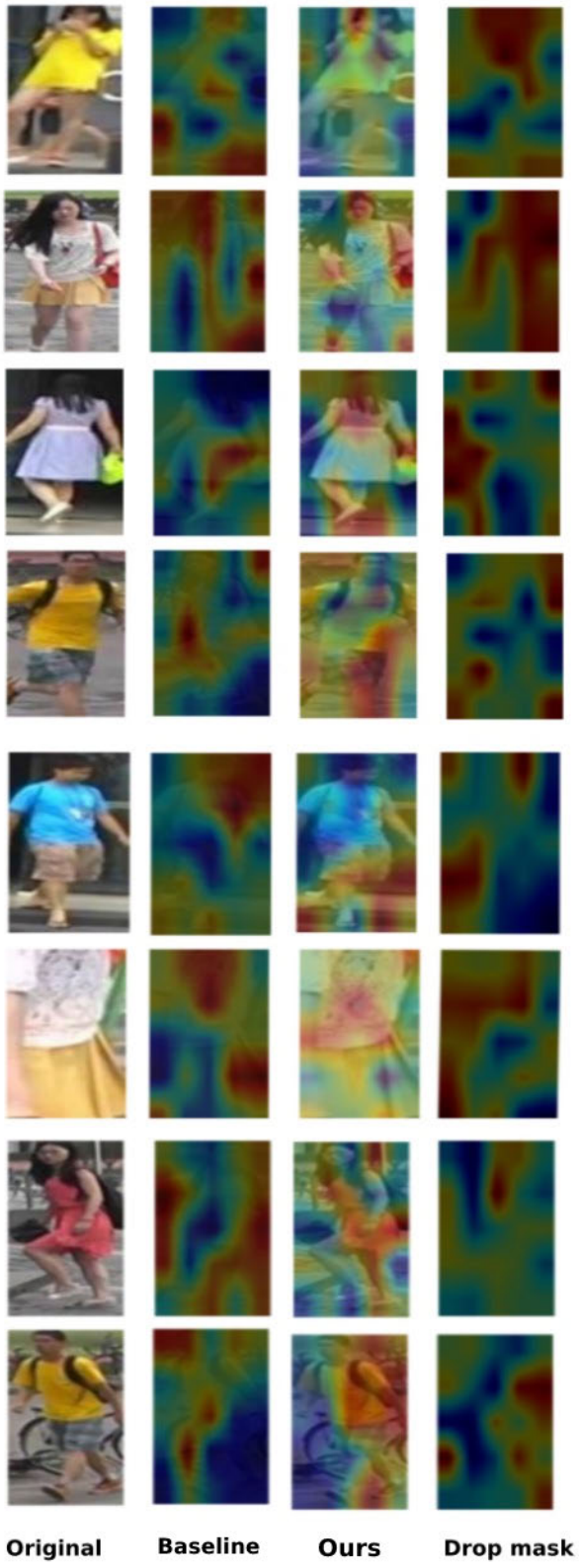

FIGURE 6. Visual result on the dataset market 1501.

The class activation map is indicating the most discriminative regions and spatially distributed features which are being used by the proposed model to re-identify persons. The

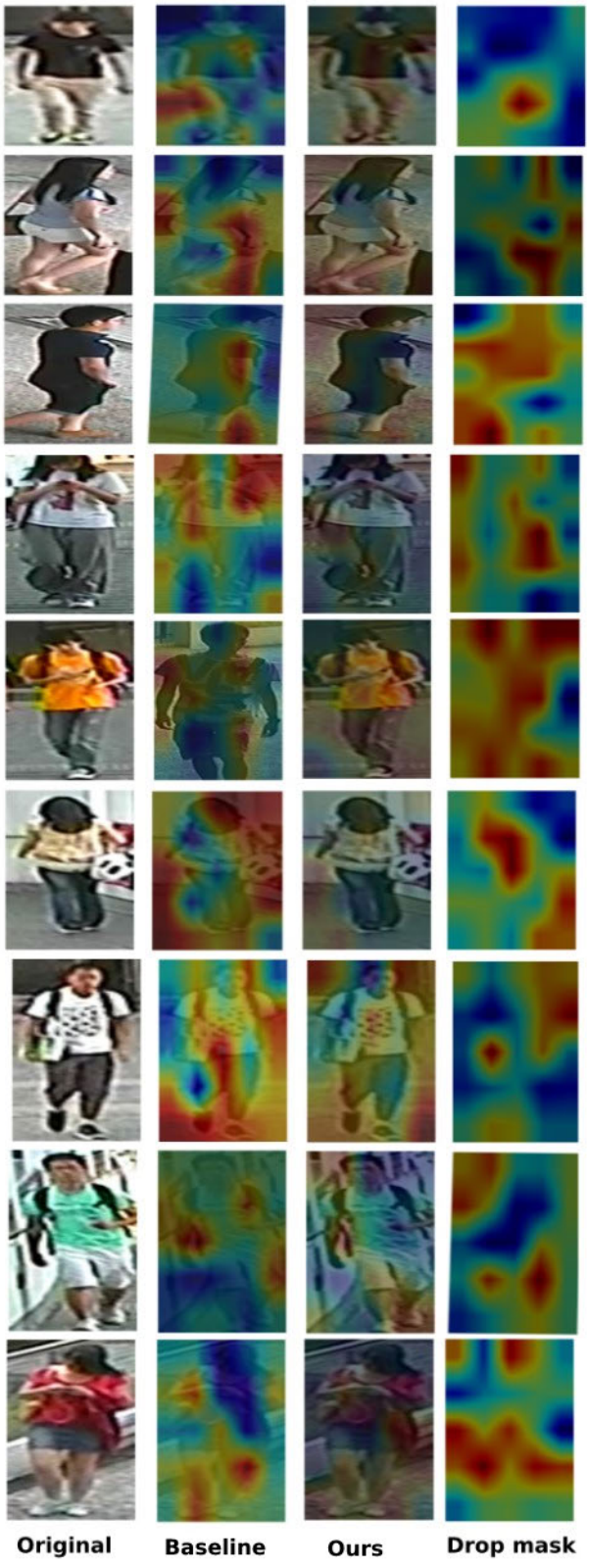

FIGURE 7. Visual result on the dataset Cuhk-03.

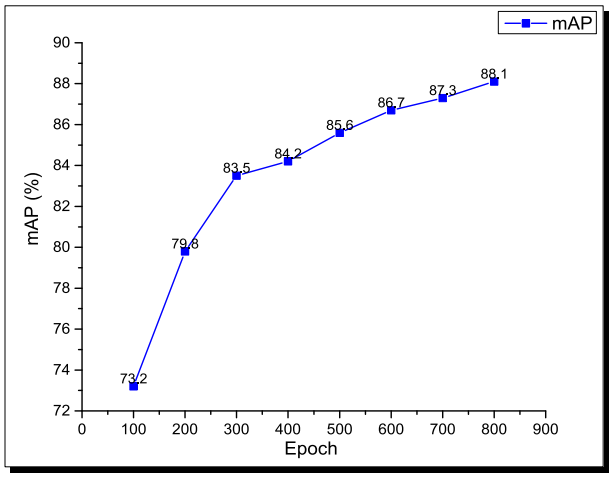

FIGURE 8. Evaluation on the dataset market 1501 mAP curve with label.

proposed model entails a structure which ensures simple connectivity with the subsequent layers with the help of which the most discriminative regions of an image can be identified properly. This is achieved by using the parameters of the output layer and the convolutional feature maps projected on each other. This generates a class activation map representing the weighted sum of the feature maps which is more comprehensive in nature and helps to achieve insights of the learning process involved in the proposed model. The class activation map seems to highlight some non-discriminative regions in few cases but also focuses on more attentive region features. Also, visual inspection of the salient representations from the BDB reveals that the contours of the persons are more clear and accurate. Another intuitive explanation of focusing on non-important regions is that, the reinforcement of the attentive feature learning on all parts of a person with semantic correspondences is ensured by blocking the roughly aligned regions.

It was also observed from experiments that the proposed ResNet-Dilated convolution block with self BAM strategy achieves the fine-grained local feature learning in a robust manner as compared to osNet [11] and CBAM [15]. The first column in Figure 6 and 7 represents the original images and the remaining columns shows the visualization of baseline and drop mask. When self-attention is incorporated with the fusion of BAM, attention weights are calculated to compute the high-level semantic detail content of the feature to represent the pedestrian with full feature representation power.

Figure 8 and 9 depicts that despite the challenges present in the publicly available re-identification datasets, the performance of our proposed system is higher. Even in the presence of background clutter, and the low resolution of images, the proposed model achieve 88.3 % accuracy map and 96.1 % accuracy of Rank-1 computed against 800 number of epochs. In contrast, the benchmark schemes are not performing well visually and quantitatively. Therefore, the proposed method has a strong feature discriminative ability to extract features with a higher Rank-1, and map score.

Additionally, the proposed system is compared by using a bar chart generated among the accuracy of map and Rank-1

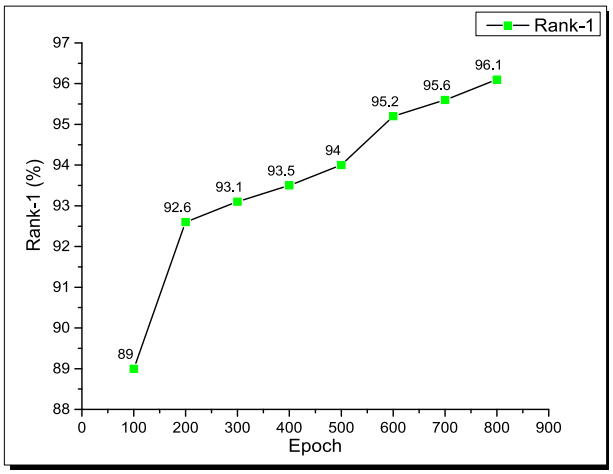

FIGURE 9. Evaluation on the dataset market 1501 Rank-1 curve with label.

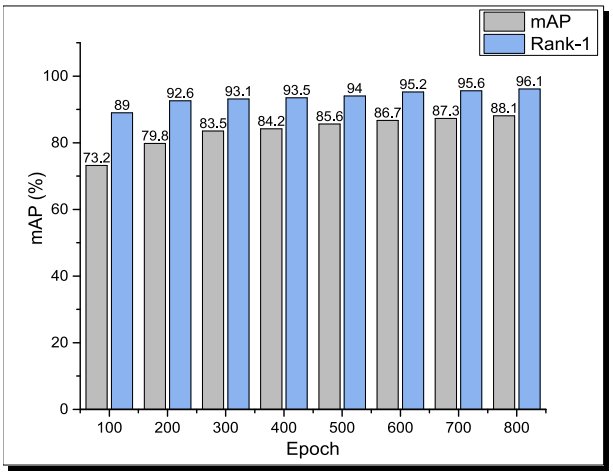

FIGURE 10. Evaluation on the dataset market 1501 mAP and Rank-1.

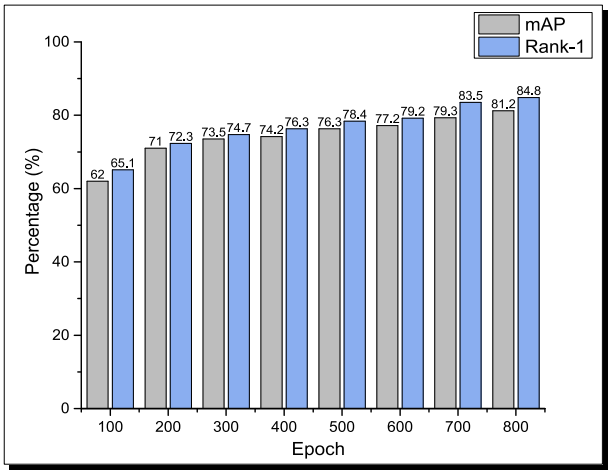

FIGURE 11. Evaluation on the dataset Cuhk-03 mAP, Rank-1.

against different numbers of epochs. Figure 10 shows the evaluation on the dataset Market 1501 mAP and Rank-1 while Figure 11 shows the evaluation on the dataset Cuhk-03 in

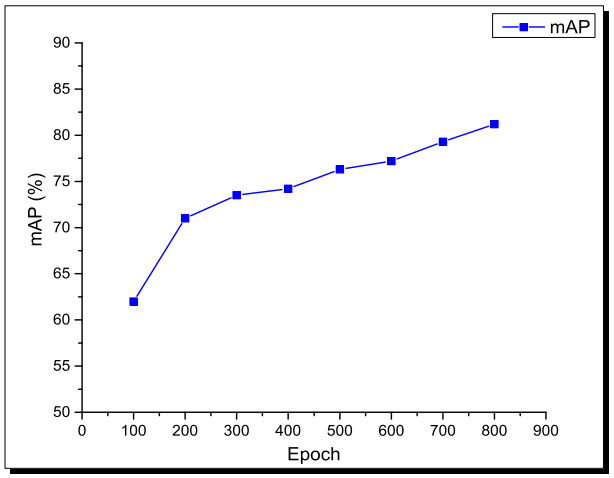

FIGURE 12. Evaluation on the dataset CUHK-03 mAP.

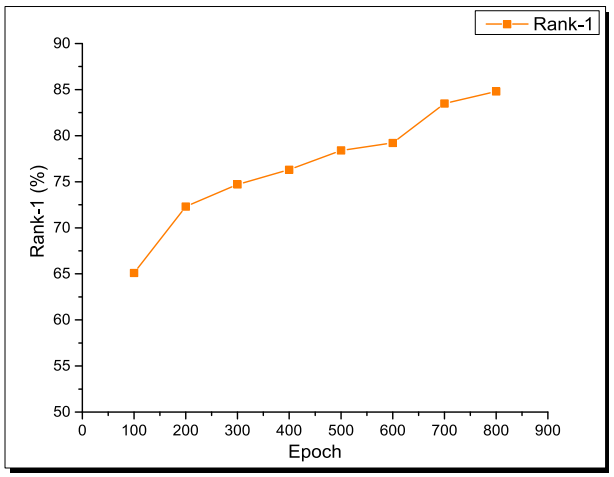

FIGURE 13. Evaluation on the dataset CUHK-03 Rank-1 curve.

terms of mAP and Rank-1. It chooses various aspect's ratio in BAM such as dilation factor, reduction ratio and dropping ratio of the image with the dimension of 384 *128. Self BAM uses the same training procedure as of BDB. It uses lower and upper feature maps for person identification because lower feature maps improve semantic information and capture precise information from the input image. It uses a single network for multiscale prediction by utilizing features from different layers.

We further configure the parameters for the training to obtain the results from the combined loss (Softmax + Batch Hard Triplet). Figure 12 and 13 shows the training curves of mAp and Rank-1 on the CUHK-03 dataset. It can be observed that the value of the map and Rank-1 consistently grow higher as the number of epochs increases. This method is still robust to the baseline approach, which increases the map from 76.0% to 81.0% and Rank-1 from 79.0% to 83.3 % to extract the multiscale feature robustly.

Figure 14 shows the training loss against the number of epochs. It can be observed that as the number of epochs

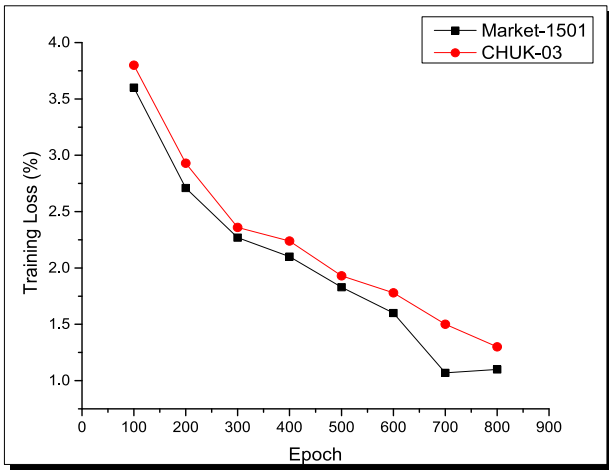

FIGURE 14. Evaluation on the dataset loss of market-1501 and CHUK-03.

increases, the training loss decreases. The proposed system adopts the combined loss function to achieve higher performance because a single loss has not had enough capacity to deal with the person re-identification challenges. The proposed loss function consists of Softmax loss and Batch hard triplet loss, motivated from the baseline strategy. The results revealed that softmax performance is higher than the sigmoid loss, that is why it is incorporated in the proposed system. The objective of using batch hard triplet in our research is to mine the efficient triplet to compute the matching correspondence between the anchor and positive sample of images. It also helped to reduce the distance between different examples of anchor and negative images.

VI. CONCLUSION

This paper proposes a self-attention and BAM-based person re-identification system by incorporating dilated convolutions. The proposed system is capable of learning discriminative multiscale feature representation with different dilation factors. It also learns the discriminative local detail deduced from various feature maps required for efficient person re-identification. The corresponding LSBAM network adopted two branches; i.e., global branch which used global average pooling to represent the global feature representation and the self-attention based feature dropping branch which helped to learn the detailed multiscale low-level feature.

The proposed system effectively tries to grasp what and where to focus or suppress, and it refines intermediate features and locates spatial information directed by the feature attention map. Inspired by the self attention-based dropout layer, we suggest and empirically test the selection of an attention module at the bottleneck of a network, which is the most critical point of information flow. The proposed study reveals that an efficient feature extraction scheme preserves the contextual information that achieves the multiscale feature representation without any change of image resolution. As a result, the proposed framework will concentrate on the lower levels' target regions. At the higher levels, the network can effectively deal with target misalignment and cluttered backgrounds. The superior performance on person identification suggested that the self attention-based BAM algorithm is of broad interest in targeting the visual recognition tasks. We also demonstrated that the proposed network effectively analyzes the regularization effect of drop masks using the combined softmax and batch hard triplet loss. Extensive experiments on two public datasets reveal that, despite its lightweight module, the LSBAM network achieves state-of-the-art performance compared to the existing approaches in image retrieval tasks.

In the future, it would be the best practice to incorporate the proposed scheme in other object recognition tasks and use the re-ranking algorithm. We also intend to improve the proposed system so that it can be used on more challenging large scale video based person reidentification datasets such as MARS in which individuals has variations in poses, colors and illuminations.

ACKNOWLEDGMENT

The authors would like to acknowledge Prince Sultan University and Smart Systems Engineering Laboratory for their valuable support.

REFERENCES

[1] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1367–1376.
[2] S. Zhai, S. Liu, X. Wang, and J. Tang, "FMT: Fusing multi-task convolutional neural network for person search," *Multimedia Tools Appl.*, vol. 78, no. 22, pp. 31605–31616, Nov. 2019.
[3] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3415–3424.
[4] A. Hazarika, S. Poddar, M. M. Nasralla, and H. Rahaman, "Area and energy efficient shift and accumulator unit for object detection in IoT applications," *Alexandria Eng. J.*, vol. 61, no. 1, pp. 795–809, Jan. 2022.
[5] S. A. Kumar, I. García-Magariño, M. M. Nasralla, and S. Nazir, "Agent-based simulators for empowering patients in self-care programs using mobile agents with machine learning," *Mobile Inf. Syst.*, vol. 2021, pp. 1–10, Nov. 2021.
[6] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1179–1188.
[7] M. G. Martini, C. Hewage, M. M. Nasrall, O. Ognenoski, C. Chen, P. Chatzimisios, T. Dagiuklas, and L. Atzori, "QoE control, monitoring, and management strategies," in *Multimedia Quality of Experience (QoE): Current Status and Future Requirements*. New York, NY, USA: Wiley, 2016, pp. 149–167.
[8] J. Hu, J. Lu, and Y.-P. Tan, "Sharable and individual multi-view metric learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2281–2288, Sep. 2018.
[9] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5551–5560.
[10] R. Quispe and H. Pedrini, "Top-DB-Net: Top DropBlock for activation enhancement in person re-identification," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2980–2987.
[11] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3702–3712.
[12] D. Tao, L. Jin, Y. Wang, and X. Li, "Person reidentification by minimum classification error-based KISS metric learning," *IEEE Trans. Cybern.*, vol. 45, no. 2, pp. 242–252, Feb. 2015.

[13] N. Martinel, C. Micheloni, and G. L. Foresti, "Saliency weighted features for person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 191–208.
[14] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 653–668, Mar. 2013.
[15] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.
[16] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[17] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.
[18] T. D'Orazio and G. Cicirelli, "People re-identification and tracking from multiple cameras: A review," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1601–1604.
[19] I. U. Rehman, D. Sobnath, M. M. Nasralla, M. Winnett, A. Anwar, W. Asif, and H. H. R. Sherazi, "Features of mobile apps for people with autism in a post COVID-19 scenario: Current status and recommendations for apps using AI," *Diagnostics*, vol. 11, no. 10, p. 1923, 2021.
[20] Y. Guo and N.-M. Cheung, "Efficient and deep person re-identification using multi-level similarity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2335–2344.
[21] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4004–4012.
[22] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*.
[23] D. Wu, C. Wang, Y. Wu, Q.-C. Wang, and D.-S. Huang, "Attention deep model with multi-scale deep supervision for person re-identification," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 1, pp. 70–78, Feb. 2021.
[24] M. M. Nasralla, "Sustainable virtual reality patient rehabilitation systems with IoT sensors using virtual smart cities," *Sustainability*, vol. 13, no. 9, p. 4716, Apr. 2021.
[25] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "DropBlock: A regularization method for convolutional networks," 2018, *arXiv:1810.12890*.
[26] R. R. Varior, M. Haloi, and G. Wang, "Gated Siamese convolutional neural network architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 791–808.
[27] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE TIP*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.
[28] W. Ge, "Deep metric learning with hierarchical triplet loss," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 269–285.
[29] A. Munir, N. Martinel, and C. Micheloni, "Self and channel attention network for person re-identification," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 4025–4031.
[30] P. Peng, T. Xiang, Y. Wang, M. Pontil, S. Gong, T. Huang, and Y. Tian, "Unsupervised cross-dataset transfer learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1306–1315.
[31] H.-X. Yu, W.-S. Zheng, A. Wu, X. Guo, S. Gong, and J.-H. Lai, "Unsupervised person re-identification by soft multilabel learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2148–2157.
[32] L. Zhang, L. Zhang, B. Du, J. You, and D. Tao, "Hyperspectral image unsupervised classification by robust manifold matrix factorization," *Inf. Sci.*, vol. 485, pp. 154–169, Jun. 2019.
[33] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2019.
[34] J. Bi and C. Zhang, "An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme," *Knowl.-Based Syst.*, vol. 158, pp. 81–93, Oct. 2018.
[35] M. Geng, Y. Wang, T. Xiang, and Y. Tian, "Deep transfer learning for person re-identification," 2016, *arXiv:1611.05244*.
[36] M. M. Nasralla, I. García-Magariño, and J. Lloret, "MASEMUL: A simulation tool for movement-aware MANET scheduling strategies for multimedia communications," *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–12, Mar. 2021.
[37] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 732–748.
[38] X. Bai, M. Yang, T. Huang, Z. Dou, R. Yu, and Y. Xu, "Deep-person: Learning discriminative deep features for person re-identification," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107036.
[39] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5157–5166.
[40] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 994–1003.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[44] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.
[45] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 403–412.
[46] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 274–282.
[47] H. Sheng, Y. Zheng, W. Ke, D. Yu, X. Cheng, W. Lyu, and Z. Xiong, "Mining hard samples globally and efficiently for person reidentification," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9611–9622, Mar. 2020.
[48] Q. Zhou, B. Zhong, X. Liu, and R. Ji, "Attention-based neural architecture search for person re-identification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6627–6639, Nov. 2022.
[49] E. H. Adelson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *Org. RCA Eng.*, vol. 29, no. 6, pp. 33–41, Nov. 1984.
[50] N. Martinel, G. L. Foresti, and C. Micheloni, "Aggregating deep pyramidal representations for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1544–1554.
[51] X. Ning, K. Gong, W. Li, L. Zhang, X. Bai, and S. Tian, "Feature refinement and filter network for person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3391–3402, Sep. 2021.
[52] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*.
[53] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "A simple and light-weight attention module for convolutional neural networks," *Int. J. Comput. Vis.*, vol. 128, pp. 783–798, Apr. 2020.
[54] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[55] Y.-B. Liu, R.-S. Jia, Q.-M. Liu, X.-L. Zhang, and H.-M. Sun, "Crowd counting method based on the self-attention residual network," *Int. J. Speech Technol.*, vol. 51, no. 1, pp. 427–440, Jan. 2021.
[56] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. Wu, "Diverse part discovery: Occluded person re-identification with part-aware transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2898–2907.
[57] B. Jiang, S. Wang, X. Wang, and A. Zheng, "STADB: A self-thresholding attention guided ADB network for person re-identification," 2021, *arXiv:2007.03584*.
[58] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.
[59] J. Liu, Z.-J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei, "Multi-scale triplet CNN for person re-identification," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 192–196.
[60] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch DropBlock network for person re-identification and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3691–3701.

MUHAMMAD USMAN YASEEN received the B.S. degree in computer engineering from COMSATS University Islamabad, Pakistan, the M.S. degree in computer engineering from Dalarna University, Sweden, and the Ph.D. degree in machine learning from the University of Derby, U.K.

He is currently an Assistant Professor with COMSATS University Islamabad. He has published numerous research papers in peer-reviewed conferences and high-impact journals. His current research interests include machine learning, data analysis, video analytics, and management for large-scale computing, and scalability in high performance computing platforms.

MOUSTAFA M. NASRALLA (Senior Member, IEEE) received the B.Sc. degree (Hons.) in electrical engineering from the Hashemite University, Jordan, in 2010, and the M.Sc. degree in networking and data communications from Kingston University London, U.K., in 2011, where he received the Ph.D. degree from the Faculty of Science, Engineering and Computing (SEC). He was a member of the Wireless Multimedia and Networking (WMN) Research Group, Kingston University London. He is currently an Associate Professor with the Department of Communications and Networks Engineering, Prince Sultan University (PSU), Riyadh, Saudi Arabia. He is a fellow of the Higher Education Academy (FHEA). He is currently an Active Member of the Smart Systems Engineering Laboratory, PSU. He has solid research contributions in the area of networks and data communications which are proven with publications in reputable journals with ISI Thomson JCR. He has received several national and international funded projects, such as U.K. home office, EU FP7 CONCERTO, and 5G-enabled Smart City Development in Saudi Arabia. He has published over 40 papers in high impact factor journals and reputable conferences. His research interests include the latest generation of wireless communication systems (e.g., 6G, 5G, LTE-A, and LTE wireless networks), wireless sensor networks, network security, the Internet of Things (IoT), machine learning, radio resource allocation, telemedicine and video compression, and multimedia communications. He is a Senior Member of IEEE Young Professionals, IEEE ComSoc, and Association of Computing Machinery (ACM). He is an Organiser of the International Conference on Sustainability: Developments and Innovations and the 5G-Enabled Smart Cities Workshop in the 7th IEEE International Conference on Smart Cities. He served as an Active Reviewer and received several Distinguished Reviewer Awards from several reputable journals, such as IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, *Wireless Communications* (Elsevier), and *Computer Networks* (Elsevier). Currently, he is a Guest Editor of *Alexandria Engineering Journal* (Elsevier), *International Journal of Distributed Sensor Networks* (SAGE), and *Frontiers in Communications and Networks*.

FAIZA ASLAM received the B.Sc. degree in computer science from Bahauddin Zakariya University, Multan, Pakistan, in 2013, the M.C.S. degree from COMSATS University, Vehari, Pakistan, in 2016, the M.S. degree in computer science from COMSATS University Islamabad (CUI), Islamabad, Pakistan, in 2022. Her research interests include image recognition leveraging computer vision and deep learning techniques.

SYED SOHAIB ALI received the bachelor's degree in electronics engineering from International Islamic University, Islamabad, Pakistan, in 2007, the master's degree in electrical and telecommunication engineering from the National University of Sciences and Technology, Islamabad, in 2012, and the Ph.D. degree in computer science and engineering from the University of Genoa, Italy, in 2017.

He is currently working as a Geoscience Software Engineer with CGG services, U.K. He has number of research publication to his credit. His research interests include image registration, de-noising, image enhancement, and defogging.

SOHAIB BIN ALTAF KHATTAK received the B.S. degree in electrical engineering from COMSATS Institute of Information Technology, Pakistan, in 2014, and the M.S. degree in electrical engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2017. He is currently pursuing the Ph.D. degree with the Communication Research Center, Harbin Institute of Technology (HIT), China. Currently, he is affiliated with the Department of Communications and Networks Engineering, Prince Sultan University (PSU), Saudi Arabia, as a Visiting Ph.D. Student. He is a member of the Smart Systems Engineering Laboratory, PSU. He has also worked as a Research Assistant at the Lahore University of Management Sciences (LUMS), Pakistan, and Ilma University Karachi, Pakistan. His recent research interests focus on wireless communications, RF-positioning, and indoor localization.