The Visual Computer (2023) 39:4087–4102
https://doi.org/10.1007/s00371-022-02577-0

ORIGINAL ARTICLE

Per-former: rethinking person re-identification using transformer augmented with self-attention and contextual mapping

N. Pervaiz[1] · M. M. Fraz[1] · M. Shahzad[1,2]

Accepted: 6 June 2022 / Published online: 9 July 2022

Abstract

Person re-identification (re-id) is an autonomous process that uses raw surveillance images to identify a person across multiple non-overlapping camera views without requiring any kind of hard biometrics like fingerprints, retina patterns or the facial images. The CNN-based deep architectures are most frequently used to solve the person re-id problem. Generally these CNN architectures capture the attentive regions of a person at local neighborhood level with increased focal view at the deeper levels of the network. However these do not learn the self-attentions among distant parts of a person's image, which can play a vital role in person re-id especially to handle the inter-class and intra-class variances. In this work, we propose a novel person re-id approach to learn the self-attention among different parts of a person image whether these lie within local proximity or at the far distant regions for robust re-identification. We adapt the vision transformer architecture with a lightweight self-attention module, which learns the global associations among the distinct attentive regions having similar context within a person image. Further to this, we escalate the baseline model by acquainting it with a self-context mapping module, which coalesces the contextual embeddings into the self-attention learning for the neighboring and the distant image regions. It helps to capture the globally associated salient regions of a person to get the holistic view at the initial network layers. The proposed self-attention-based re-id architecture outperforms the vanilla CNN counterparts for both of the re-id performance measures, i.e., accuracy and mean average precision. The re-id accuracies are improved 5.5%, 4.6% and 17% for Market1501, DukeMTMC-ReID and MSMT-17 datasets, respectively, as compared to the vanilla CNN-based re-id architectures. The implementation and trained models are made publicly available at https://git.io/JLH2S.

Keywords Person re-identification · Visual surveillance · Vision transformer · Self-attention · Self-context mapping

1 Introduction

Person re-identification is a computer vision-based research problem. It aims to identify a person as a same entity when he/she appears in different cameras of a surveillance network or in the same camera at different times. Figure 1 shows an autonomous person re-identification system that tends to identify a child captured by four different non-overlapping cameras. Where the huge variations in cameras' views, illumination conditions and poses make it more challenging to re-identify the child in all different cameras' views [1]. The major applications of person re-id include general public security, lost child recovery, theft prevention and the road harassment deterrent.

Over the years, multiple deep learning architectures are emerged; the convolutional neural networks (CNNs) are the most popular deep architectures [2–4] to solve vision-based research problems including person re-identification. Numerous CNN variants have been developed over the years which have improved the performance of person re-id to a great extent [5–9]. However, the CNN-based architectures have two major limitations: (1) The convolutional layers capture local neighborhood based attention regions of an image with increasing receptive field along the depth of network. Due to the small receptive field, the initial layers of the network are unable to capture the holistic view of an image and

✉ M. M. Fraz
moazam.fraz@seecs.edu.pk

N. Pervaiz
nazia.perwaiz@seecs.edu.pk

M. Shahzad
muhammad.shahzad@tum.de

[1] National University of Sciences and Technology (NUST), Islamabad, Pakistan

[2] Technical University of Munich (TUM), Munich, Germany

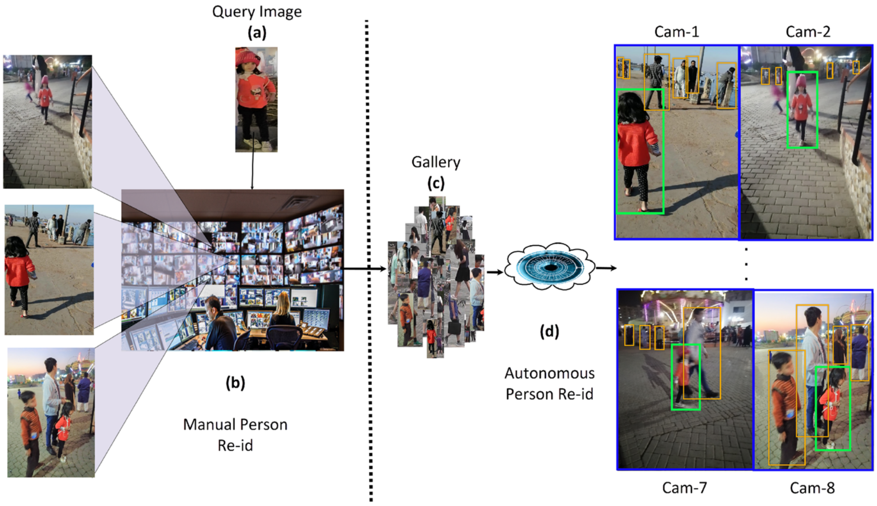


Fig. 1 Person re-identification: **a** A query image to be re-identified over the complete surveillance network; **b** Current practices of manual security surveillance; **c** A pool of gallery images; **d** An automated process to re-identify a person correctly across multiple non-overlapping camera views on the basis of his/ her appearance

the deeper layers rely on the learning of preceding layers for global visualization of the image. Hence the global view of an image is only attainable at the deeper layers and not accessible at the initial layers of the network. (2) The CNN models do not capture the associations among far-distant attentive regions of an image at any particular level of the network [10,11].

To address the first limitation of vanilla CNN models, various re-id architectures propose additional branches or streams in CNN architectures to integrate the explicit spatial or channel-wise attention learning with CNN embeddings [12–21]. In these multi-branch networks, one branch computes vanilla CNN embeddings, while the other focuses on explicit attention computation and are jointly optimized. However most of these architectures are not the generalized ones and learn only the specified attention patterns. Moreover the second limitation remains unaddressed, i.e., lack of self-attention learning among far off positions of an image [22,23].

The self-attention among distant parts of a person image provides global view of the image. For instance, if a network layer visualizes the entire image to learn its all attention regions and builds the associations among distant attention regions, it can easily re-identify a person wearing a red bag and white shoes despite poses variations and illumination changes.

Transformer is a deep learning model that learns self-attentions among a given sequence of inputs and is currently state of the art to solve the language problems [10]. However, due to the large number of pixels in an image, it is not practical to use a standard transformer architecture for the vision-based classification tasks. A standard transformer encoder intakes a sequence of one-dimensional data, it learns self-attentions among the given sequences and embeds this self-attention information into the final representations of the data. For images, one potential way is to take the pixels sequences as one-dimensional input data. However, typically a large number of pixels in images makes the self-attention computation across the whole set of pixels computationally very expensive and thus consequently limits the use of transformers particularly for vision applications.

We propose a re-id approach that follows the transformer architecture to learn the globally associated attentions of an image in each layer of the model. However, instead of using the individual pixels for self-attention computation, we divide a person image into a predefined number of patches during the image to sequence conversion phase. The patches are then fed sequentially to the network. As compared to the pixel-wise sequence input, the patch-wise sequence input reduces the computations exponentially. It is important to mention that the image to sequence conversion process is performed using the vision-based variant of transformer [24]. The proposed work is one of the first few demonstrations of transformers in the context of person re-identification. We extend the proposed baseline model by incorporating a self-context mapping (SCM) module in it. The SCM augments the self-attention learning process with the self-contextual mapping and results in more generalized person representations. It enhances the network inference performance by robustly tackling the intra-class variations. The proposed baseline architecture (i.e., without SCM) tends to learn only the global associations among the attentive regions/patches that may be spatially far within an image. The inclusion of the SCM module further improves the global self-attention mapping by encompassing the self-contextual embedding even among the less attentive distant regions within an image. Hence, the incorporation of the latent representation learning from the comparatively inconsiderate regions augments the proposed network's discriminative ability. The proposed extended model outperforms the baseline model and the existing CNN-based re-id models. The contributions presented in this work are summarized as follows:

1. We adapt the transformer architecture with an enhanced self-attention module, which learns the global associations among the distinct attentive regions having similar context within a person image, for robust re-identification.
2. We also escalate the baseline model by acquainting it with a self-context mapping module, which coalesces the contextual embeddings into the self-attention learning for the neighboring and the far-flung regions in an image.
3. We evaluate the proposed baseline and its extended variant on three public re-id benchmarks, i.e., Market1501, DukeMTMC-Reid, and MSMT-17 and attain significant improvements over the CNN-based re-id approaches.

2 Related work

Over the last couple of years, with the advancement of hardware architectures and the availability of large-scale datasets, the attention-based deep architectures are being explored extensively. Attention mechanisms get intuition from the human vision system and focus on the most attentive regions of the image for decision making and suppress the less attentive regions.

Since the beginning of the deep learning era, the attention regions of the images are captured at small receptive fields of the input image and are propagated towards deeper layers of the deep networks to have their aggregated perception [25–27]. This cumulative attention information provides good intuition about the regions of the images, which can play an important role of person re-id. But it bounds the deeper layers of CNNs to view whatever the initial layers of a deep architecture have forwarded to them hence, losing the global context to a great extent. One of the primitive vanilla CNN-based re-id architecture is ID-discriminative embedding (IDE) [28] that integrates the person detection confidence in the person re-id scores and provides initial insights into how weakly labeled detection data helps improve re-id accuracy.

The invention of the skip connection [2] and its subsequent variants handle this global information loss elegantly, however as each block of the network uses only its preceding block's output in learning of its activation map and results in losing global information. The residual networks are among the most famous CNN architectures, where residual connections are used to reduce the information loss while progressing data towards next layers in a deep network. The ResNet-50 [2] is used as a standard CNN baseline to build further customized and specialized deep CNN architectures to solve many classification and retrieval problems, including person re-id. The dense connections-based densenet architecture [29] handles this issue and inputs all previous block outputs (including input) into the all succeeding blocks to keep maximum information along the way. However, these dense connections need huge computational resources, especially at large scale. Moreover, the convolutional layers of both networks [2] and [29] focus on learning attentions at the local neighborhood level and do not establish mapping of attentions at distant positions of a given image at each layer, w hereas, in the proposed work, we learn the self-associations among attentive regions of a person's image to perform person re-id.

TriNet [20] is a metric learning-based CNN architecture that learns the optimal feature space by looking at both the positive and negative anchor images. The singular vector decomposition (SVD) deep network [30] is built upon ResNet-50 baseline and iteratively integrates the orthogonality constraint in CNN training. The attribute augmented person re-id [31] scheme, which is a little closer to our baseline architecture in their average precision, needs additional requirements of triplet formation or the ground truth of person attributes.

By using variants of CNNs as a backbone, several customized re-id networks are designed that explicitly embed higher level attention besides the local attentions learnt by CNNs base architectures. Most of these attention learning networks opt multi-stream structures to learn the regional/ spatial attention along with deep convolutional representations [32,33]. In addition to learning the spatial attentions, the attentive but diverse re-id model [34] and critical attention learning mechanism [35] also focus in learning channel wise attentions in order to extract the most significant channels only as all the channels do not contain significant information. Besides attentive channel information, [34] learns correlation among the attentive channels as well, however all these multi-stream attention learning mechanisms need huge computational resources [36]. Moreover these networks do not learn the associations among far-distant attention regions of an image which are apparently dispersed all over the image but can provide better intuition to re-identify a person. To this end, we proposed a mechanism for smart association of the dispersed attentive and contextual regions with each other, which gives more generalized and holistic representation of a person in the image.

The multi-task and attention-aware learning network [37] proposed by Chen et al. demonstrated an explicit holistic attention branch to learn global attention along with a partial attention learning branch to learn local attention. However, the network has additional requirements of the key points for its local attention branch and learning self-attention within sparse parts of the image is not the scope of this work.

The abovementioned methods learn hard-level attention to perform person re-id, but the computation of only hard attention makes these networks less generalizable. Harmonious attention networks [12] handle multi-level attention, i.e., both of the hard attention (regional saliency) and the soft attention (pixel level saliency). It keeps the generalization of CNNs as well, but it still cannot to capture the relationships among

distant attentive regions within an image. Taking inspiration from the state-of-the-art self-attention methods for natural language problems solution, the researchers explored the self-attention impact for person re-id task. In [38], the multi-scale convolutions are applied to the entire image and to the predefined three local parts of the image, i.e., upper, middle and bottom. The latent parts localization is performed by using spatial transformer networks to learn the self-attention. However, this work needs the positions of local parts and the value range constraint on the scale parameter as prerequisites. Luo et al. first time use the transformer-based network for person re-id. [39] integrates the deep convolutions-based re-id module with the pair-wise spatial transformer networks (STN) module to perform the partial person re-id. The spatial transformer networks module samples an affined image (a semantically corresponding patch) from the holistic image to match the partial image. The re-id module learns the embedding of holistic, partial and affined images, the STN module performance is influenced by the re-id module.
One of the major limitation of all these convolution-based attention networks is that these networks learn the dependency among immediate neighborhoods both at the initial layers of the network and at the deeper layers as well [10]. The structure and working mechanism of CNNs do not exhibit the learning of attention dependency at distant positions of an image or feature maps at a particular level. This arises the need to learn self-attention or intra-attention at each learning level (layer). The self-attention learns the associations among different parts of images and embed this information in the global representations. In this work, we learn the self-attention patterns of a person's image to solve the person re-id problem. Transformer encoder [10] is a deep architecture that learns the self-attentions among the sequence of inputs to represent the whole input.

In the re-id domain, the self-attention-based additional structures [38,39] are used to supplement the learning of the core convolutions-based deep network. However, the sole use of transformers to develop a person re-identification solution is challenging because of the computationally expensive pixel-wise self-attention operations of the transformer architecture.

In this work, the transformer encoder-based person re-identification model Per-former$_{base}$ is proposed and is further enhanced by integrating self-context mapping module in it. The mappings of self-attention and associated context provide a better holistic understanding of the image. We compare the extended version of our model (Per-former$_{SCM}$) results with our baseline model (Per-former$_{base}$) and with the various convolutions-based vanilla architectures. The proposed model significantly improves the person representation. A detailed comparison is given in the Results section.

3 Methodology

The proposed work is one of the first demonstrations of employing vision-based transformer in the context of person re-identification application. We propose a vision-based transformer architecture for person re-identification (Per-former$_{base}$) and extend it by introducing a sophisticated self-context mapping module. The image to sequence module (ISM), converts the images into a sequence of image patches to avoid the computationally expensive pixel-wise operations, which are then used as the input of Per-former$_{base}$. Next, the self-attention across all the patches of an image is computed by the multiple-head self-attention module (MSAM), which exponentially reduces the self-attention computation complexity as compared to the pixel-wise computations. The position-wise attention module (PWAM) handles the translational variances in the image patches. Finally, the self-context mapping module (SCMM) augments the contextual associations among the image patches into the baseline model, enhances its learning and results in our extended model, i.e., Per-former$_{SCM}$.

The proposed network is illustrated through Fig. 2. Following the standard practices of self-attention-based deep architectures, i.e., the first paper of transformer [24] and its extended researches [40,41] etc, we choose the depth of layers as 12-layers architecture in our work. Each layer contains three sub-modules; multi-head self-attention module, position wise activation module and self-context mapping module in addition to the residual connections to avoid the information loss along the depth of the network. All 12 layers follow the same architecture, i.e., the repetition of sub-modules. The operational details of each module/ component is given in respective subsection.

3.1 Image to sequence module (ISM)

Instead of processing pixel-wise information of the pedestrian images, we extract a sequence of smaller patches from the large size person images. For this purpose, we divide each image into a fixed number of patches. Each two-dimensional patch is then flattened through all the input channels (RGB) to form a 1D vector featuring the complete image information. The flattening process involves the conversion of a 2D patch of size 16 × 16 pixels into a 1D embedding vector of length 768. Each pixel of a patch contains RGB values, i.e., 3 values per pixel, each 16 × 16 patch comprises 256 pixels, hence a 16 × 16 patch contains a total of 768 values (256 × 3) which are extracted from red, green, blue channels of the patch. As it is crucial to keep the location information of the local discriminative parts of an image thus, we attach a learnable position vector with the embeddings vector of each patch and aim to learn the positions of the attentive regions in an image as shown in Fig. 3.

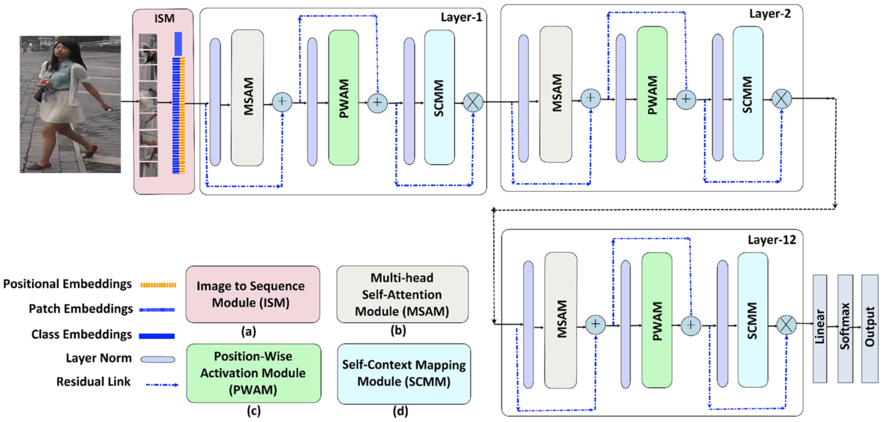


Fig. 2 The Proposed Architecture—Per-former$_{SCM}$. Detailed architecture of each module (**a**–**d**) is separately illustrated in the respective diagram

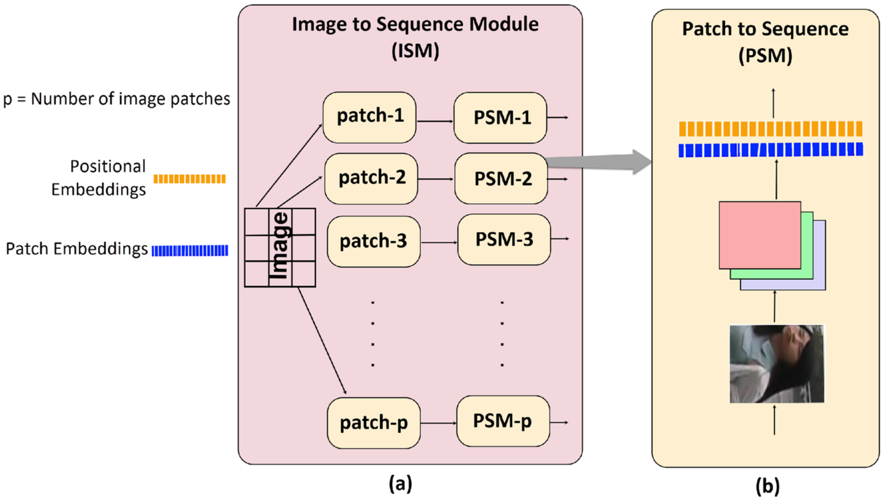


Fig. 3 Image to sequence module (ISM) involves the generation of image patches and respective positional vectors

For each image, along with its patches' embeddings vectors, an additional learnable feature vector is introduced which learns the class representations for classification purposes. Eventually, for an image, all one-dimensional vectors, i.e., patches' embeddings vectors, along with positional vectors and the class representation vector, make the input of Per-former as shown in equation 1.

$$
\begin{aligned}
Emb_{T(i)} = {} & (E_{class}, P_1E, P_2E, ...P_{n-1}E, P_nE) \\
& + Emb_{pos}
\end{aligned}
\tag{1}
$$

where $\mathrm{Emb}_{T(i)}$ is the flattened embeddings of all patches of an image i which are fed into the Per-former. Each Per-former layer comprises two major modules, i.e., multi-head self-attention module and a position-wise activation module. Additionally the layer normalization and residual connection are used to minimize the information loss.

3.2 Multi-head self-attention module (MSAM)

The multi-head self-attention module (MSAM) comprises of multiple self-attention heads (SAH) as shown in Fig. 4

The resultant embedding vectors received from the image to sequence module are fed into a self-attention module for self-attention learning among all the patches of a given image. In contrast to convolutional layers, the attention heads perceive a global view of an image at each level, where the heads attend all positions of an input layer or an image itself in case of the first layer. The heads start learning the dependencies among far distant parts of an image, right from the starting layers of the network.

3.2.1 Self-attention

The attention operations map the queries (Q) and a set of key values (K, V) to the output vector. The queries, keys and values are the matrices and make the input for a set of patches for which attention is computed simultaneously. We use the most commonly used multiplicative attention function, i.e., scaled dot product attention, instead of the additive attention method to compute the computationally efficient attention matrices. The additive attention function is computationally expensive and uses a feed-forward network with a single hidden layer, i.e., 3-layer network (input + hidden + output),

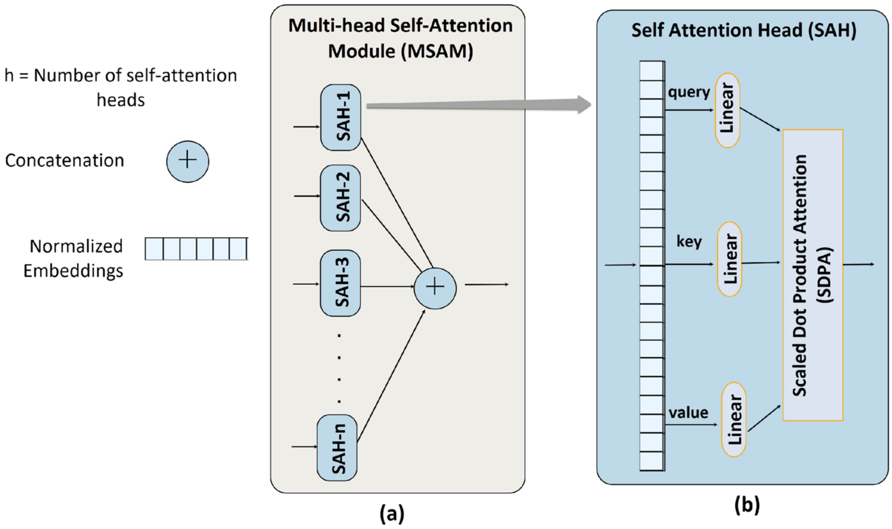

h = Number of self-attention heads
Concatenation
Normalized Embeddings
Multi-head Self-Attention Module (MSAM)
SAH-1
SAH-2
SAH-3
SAH-n
(a)
Self Attention Head (SAH)
query
key
value
Linear
Linear
Linear
Scaled Dot Product Attention (SDPA)
(b)

Fig. 4 Multi-head self-attention module (MSAM) computes scaled dot product attention and establishes associations among attentive regions among all image patches

which involves one multiplication of the input vector by a matrix, then by another matrix, and then the computation of resultant vector. In contrast, the smart implementation of the scaled dot product attention computation does not break out the whole matrix multiplication algorithm and basically is a tight, easily parallelized operation. However to avoid the vanishing gradient problem of softmax due to larger product values, we scale the product by a scale factor of $1/\sqrt{d_k}$ [10]. The scaling factor is taken as the factor of the dimensions of the model (d_m) and the number of heads (h) as shown in the Eq. 2.

$$d_k = \frac{d_m}{h} \tag{2}$$

The computation of the multiplicative self-attention (SA_{mul}) is given in Eq. 3.

$$SA_{mul}(Q, K, V) = \left(\frac{\exp(QK^T)}{\sum_j \exp(QK^T)_j} * \frac{1}{\sqrt{d_k}} \right) V \tag{3}$$

We use multiple parallel self-attention heads to run multiple parallel attention functions. Our main objective is to learn attention from different positions and representation spaces and jointly update learnable parameters of each attention head as shown in Eq. 5. This mechanism efficiently learns the self-attention and associations among local parts at distant positions in an image. The weights of the attention are set on the basis of pairwise similarity among the patches' sequences. These learnt self-attentions are further refined through the depth of transformer layers.

$$h_i = SA_{mul}\left(QW_i{}^Q, KW_i{}^K, VW_i{}^V \right) \tag{4}$$

where h_i is the attention computed by ith head with resultant trainable parameter matrices, i.e., $W_i{}^Q$, $W_i{}^K$, $W_i{}^V$, as shown in Eq. 4

$$MSA = F_x \left(\sum_i h_i \right) W^o \tag{5}$$

Finally, the multi-head self-attention module integrates the attention computed by each head using the function (F_x) of concatenation. It limits the information loss through a residual connection, i.e., W^0, which contains the normalized output vector of the preceding layer. The W^0 is integrated with a layer's output before submitting it to the next modules. The attention learning mechanism aims to learn the attention and its interdependence in a global manner.

The proposed encoder consists of 12 layers of parallel self-attention heads. Each image is divided into 16 × 16 patches. By default each patch has 3 dimensions/ channels, i.e., RGB. We linearly project the patches into 768 dimensional space (i.e., patch-width × patch-height × 3). Instead of providing absolute or relative position information (as followed in CNN architectures), we use alike dimensions learnable positional vectors of the patches and add them to the respective patch embeddings to make a sequence of input for the transformer. During the self-attention learning process, each position of a self-attention layer attends all the input positions coming from the previous layer, hence efficiently learns the globally associated attention among the patches.

3.3 Position-wise activation module

In addition to the multi-head attention module, each transformer layer comprises of a position-wise activations module (PWAM) as illustrated in Fig. 5, which handles the translational variance of the input at local level. The network learns the local discriminative positional embeddings and the local associations which play an important role in overall learning

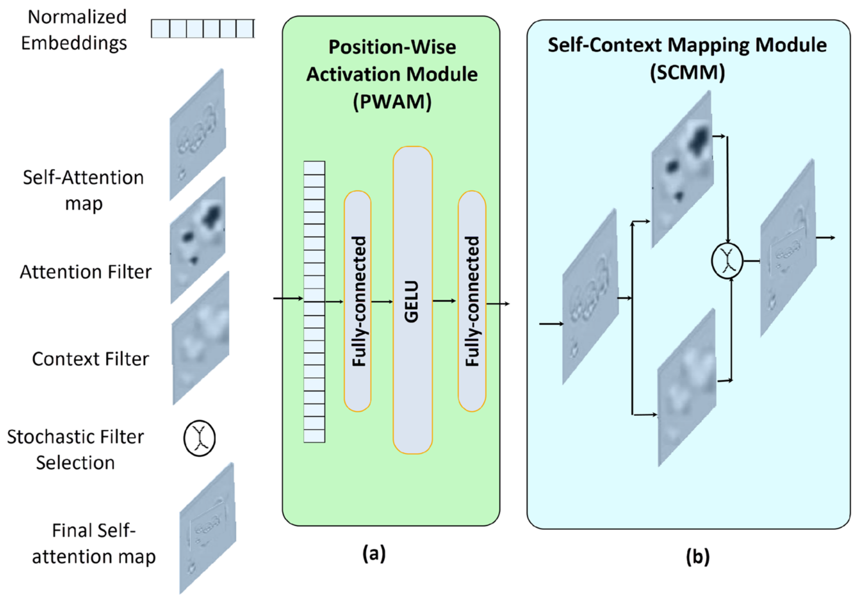


Fig. 5 **a** Position-wise activation module (PWAM) handles the translational variance of the input. **b** Self-context mapping module (SCMM) learns associations among less attentive regions of image patches in addition to the most attentive regions

of the class representations. The PWAM comprises a multilayer perceptron layer which contains two fully connected layers with the Gaussian error linear unit (GELU) nonlinearity [42], the equation 6 explains the GELU activation. The input feature dimensions for PWAM module is the same as the output of MSA module, the hidden layer dimensions are four times the input and the output features dimensions are set equivalent to the PWAM input dimensions. The Layer-norm is applied before every block and residual connections after every block. Although the same activation function is used in each layer, different sets of parameters handle the incoming position-wise attention uniquely.

$$GELU(x) \quad = \quad x\phi(x) \tag{6}$$

where x is the input vector received from the multi-head self-attention module and the $\phi(x)$ is the Euler's phi function and defines the cumulative distribution of the Gaussian distribution. The GELU nonlinearity weights inputs by their value, rather than gates inputs by their sign as in ReLUs $(x\mathbf{1}_{x>0})$. The positional embeddings do not contain the explicit position information of the attention regions, rather these are learnt during the training process.

Layer normalization is applied before each sub-module of the transformer. As we use multiple layers transformer architecture, the information loss is prevented by using residual connections. A residual connection after each module integrates the previous layer's output into the current layer's output.

3.4 Self-context mapping module

In addition to learning the highly attentive regions within images, we aim to learn the less attentive regions as well. We achieve this by removing the most attentive regions of the image patches in an efficient way. For this purpose, the most attentive parts of the images are used to create two complementary filters. One filter retains the most discriminative information in it and is known as the attention filter, while the other filter keeps all the contextual information which excludes the most attentive regions and is called the context filter.

During the training process the attention filter learns the most salient regions of the patches and increases the learning focus of the model as shown in Eq. 7. On the other hand, the contextual filter suppresses the most salient regions and the model focuses to learn the contextual information implicitly as given in Eq. 8. Both of the filters are selected stochastically during the training, Eq. 9 enables the model to effectively learn both the most attentive parts of the image and the contextual information, the final output is shown in equation (10).

$$Z' \Rightarrow \top = Att_{filter} \tag{7}$$

$$Z' \Rightarrow \bot = Con_{filter} \tag{8}$$

where Z' is the input fed into the SCMM module, the $\top$ and $\bot$ functions use a threshold function to generate the Att_{filter} and Con_{filter}, respectively.

$$M = Attention_{mask} \quad || \quad Context_{mask} \tag{9}$$

$$Output_{SCM} = Z' \quad \otimes \quad M \tag{10}$$

In contrast to the multi-stream architectures [26,29,32,33] which explicitly integrate the additional attention streams in the base architecture of convolutional networks to capture the attention regions, the SCM module is a lightweight structure that does not cause any overhead in terms of trainable parameters. The SCM module is an elegant architecture to learn the attention along with contextual information without increasing the computational overhead.

We plug-in the SCM module in each transformer layer such that the output of the SCM module of one layer becomes the input of the subsequent layer. Therefore, in addition to the learning of self-attentions among the sequence of input patches of an image, the SCM module enables the learning of self-contextual relationship among the given set of patches/sequences. The residual connections employed at various stages of the network minimize the information loss. We use Imagenet pre-trained weights to train the proposed model and the cross-entropy loss function to optimize the proposed model. We visualize the activations of SCM endorsed self-attentions which precisely focus the associated attentive regions of an image and are significant for person re-id. Self-attention-based activations are absolutely different from the convolutions-based activations. Detailed qualitative and quantitative results are presented in the results and discussion section.

4 Experiments

4.1 Materials

We choose three standard medium- to large-scale person re-identification datasets for analytical study of the proposed self-attention-based re-id network along with the other deep vanilla architectures. The datasets include Market-1501 [43], DukeMTMC-reID [44] and a recently proposed large-scale re-id benchmark MSMT17 [45]. Details of all three datasets are given in respective sections:

4.1.1 Market1501

Market1501 is a medium-scale image-based person re-identification dataset comprising a total of 32,668 person image crops which are captured by 6 cameras. It contains 1501 unique person identities, where 751 unique ids are included in the training set and the remaining ids are part of the gallery set. The query set consists of 3,368 person crops which are searched from the gallery images.

4.1.2 DukeMTMC-reID

DukeMTMC-reID is another medium-scale image-based person re-identification dataset comprising a total number of 36,411 person crops captured by 8 different cameras. It contains 1812 unique person identities, out of which 702 unique identities are included in the training set and the remaining non-overlapping 1110 identities are part of the gallery set. The query set consists of 702 ids which are searched from the gallery images.

4.1.3 MSMT17

MSMT17 is a large-scale person re-identification dataset comprising 124,068 images captured by 15 cameras with 4101 unique identities in it. The twelve cameras from a total of fifteen cameras capture indoor images and the remaining three cameras capture outdoor images. The training set includes 1041 unique identities with a total of 30,248 image crops. The rest of 93,820 images contain 3060 unique identities and form the gallery and query sets. The query set consists of 11,659 person crops and the gallery set consists of 82,161 person crops. MSMT17 is comparatively a complex dataset with excessive variations in the images background, illumination and the poses.

Table 1 summarizes the stats of datasets which are used to evaluate the proposed method.

4.2 Evaluation metrics

For person re-identification, the most widely used evaluation metrics are the cumulative matching characteristics (CMC) and mean average precision (mAP) [28]. We choose the same to evaluate the proposed network. Details of each metric are given in respective subsection.

4.2.1 Cumulative matching characteristics (CMC)

For given query images, the cumulative matching characteristics (CMC) computes the ranks of gallery images on the basis of their similarity with the query images. Rank-1 accuracy counts all the correct gallery images which are matched with the query image on the top rank, i.e., rank-1 when a particular method is used. We use rank-1 up till rank-20 accuracy to evaluate the proposed method.

4.2.2 Mean average precision (mAP)

For the multiple non-overlapping query instances of a unique identity, we compute the average precision of the correct matching of all query instances. It is termed as the mean average precision.

4.3 Experimental settings

We use publicly available person re-id benchmarks and opt the self-attention-based deep architecture for person re-identification. We resize the person images height and width to 224 × 224 with the patches of various sizes, i.e., 16 × 16, 32 × 32 and 64 × 64. We use the pretrained weights of Imagenet-21k provided by [24]. For data augmentation we

Table 1 Specifications of person re-identification datasets which are used to evaluate the proposed Per-former re-id models

Datasets	Cameras	Images	Total IDs	Train IDs	Test IDs
Market1501	06	32,668	1501	751	750
DukeMTMC-ReID	08	36,411	1812	702	702
MSMT-17	15	124,068	4101	1041	3060

opt the standard approaches of horizontal flip and random crop which are most appropriate for the pedestrian datasets. We apply standard Imagenet normalization on the images. We use recently proposed *rectified adam optimizer (radam)* [46] and choose the default settings of hyper-parameters as $beta_1 = 0.9$ and $beta_2 = 0.999$ in this work. It is adaptable to a wide range of learning rates (starting from 0.1 to 0.003) with slightly increased performance output as compared to the vanilla Adam optimizer. We set the starting learning rate from 0.0003 which is decayed with the factor of 0.1 at every tenth and twentieth epochs. We train all the models on NVIDIA Tesla P40 GPU with 24GB graphic RAM size, we choose the batch size of 128. We train all the models for the maximum of 150 epochs for all of the re-id datasets.

Table 2 Comparison of proposed methods with existing vanilla architectures for Market-1501 dataset (All results are measured in the percentages)

Method	mAP	R1	R5	R10	R20
IDE [28]	46	72.54	–	–	–
SVDNet [30]	62.1	82.3	-	-	–
DF [47]	63.4	81	–	–	–
PDF [47]	63.41	84.14	–	–	–
DJL [48]	65.5	85.1	–	–	–
IDE+Camstyle [49]	65.87	85.66	–	–	–
ResNet-50 [2]	65.9	83	92.7	95.2	97
A3M [31]	68.97	86.54	–	–	–
TriNet [20]	69.14	84.92	94.21	–	–
Per-former$_{base}$	72	88.5	94.6	96.6	98.2
Per-former$_{SCM}$	**74.5**	**89.4**	**95.6**	**97.6**	**98.5**

Bold value indicates the superior quantitative performance of the respective methodology

Table 3 Comparison of proposed methods with existing vanilla architectures for DukeMTMC-ReID dataset (All results are measured in the percentages)

Method	mAP	R1	R5	R10	R20
IDE [28]	51.83	72.31	–	–	–
ResNet-50 [2]	55.96	73.2	–	-	–
DenseNet-121 [29]	55.08	73.16	-	–	–
TriNet [20]	53.5	72.44	–	-	–
TriNet+RE [20]	56.6	73	–	-	–
SVDNet [30]	56.8	76.7	–	–	–
IDE+CamStyle [49]	57.61	78.32	–	–	–
Per-former$_{base}$	57	77.8	86.6	89.5	91.7
Per-former$_{SCM}$	**61.6**	**81.1**	**89**	**91.5**	**93.2**

Bold value indicates the superior quantitative performance of the respective methodology

5 Results

We compare the results of proposed architecture Per-former$_{SCM}$ with existing convolutions-based vanilla architectures as well as with our baseline network, i.e., Per-former$_{base}$. In this work, instead of using the convolutions-based feature maps we propose an entirely transformer-based person re-id solution, we call it the baseline (or vanilla architecture) for the purely transformer-based re-id solutions. We compare this work with more than five CNN-based vanilla architectures for each of the person re-id benchmarks.

Quantitative results of the proposed architecture for all of three re-id benchmarks are given in respective tables and graphs. The results show that even our baseline/vanilla self-attention-based re-id model Per-former$_{base}$ brings a significant performance improvement over existing CNN-based vanilla architectures. The proposed extension of baseline model, i.e., Per-former$_{SCM}$ further improves its performance in respect of all re-id performance metrics.

Table 2 shows the performance comparison for re-id benchmark Market-1501. A significant performance improvement is seen for Per-former-based re-id models over vanilla CNN-based re-id models. The proposed Per-former$_{SCM}$ outperforms existing vanilla architectures and the proposed baseline method for both evaluation measures, the rank-1 accuracy and the mean average precision. The CNN vanilla architectures, compared with the proposed work are discussed in the related work section.

A comparison of proposed work and existing relevant architectures for person re-id dataset DukeMTMC-TeID is given in Table 3. The proposed method consistently improves the performance of the baseline model and surpasses all vanilla CNN re-id architectures in both performance metrics.

MSMT-17 is a pretty large person re-id benchmark with high complexity of cluttered background and poses variations. It can be seen from Table 4 that vanilla deep architectures—particularly the ResNet-50 and the

Table 4 Comparison of proposed methods with existing vanilla architectures for MSMT17 dataset (All results are measured in the percentages)

Method	mAP	R1	R5	R10	R20
ResNet-50 [2]	22.25	46.88	–	–	–
DenseNet-121 [29]	21.5	46.32	–	–	–
DLCE [50]	31.58	60.48	–	–	–
ShuffleNetV2 [51]	15.7	36.8	–	–	–
MobileNetV2 [52]	18.62	42.53	–	–	–
Per-former$_{base}$	35.6	64.1	77	81.9	85.9
Per-former$_{SCM}$	**39.4**	**67.7**	**79.7**	**84**	**88**

Bold value indicates the superior quantitative performance of the respective methodology

shuffleNet—that perform comparatively well on other person re-id benchmarks (Market-1501 and DukeMTMC-ReID), but still do not achieve good results for MSMT-17 dataset. The experimental results of proposed models outperforms all CNN-based vanilla architectures with a significant difference for all performance measures.

Figure 6 shows the results of top ten similar matches sorted out by the proposed system for given query images and it clearly shows that the system works remarkably well even for very difficult poses and low-resolution images.

6 Discussion

6.1 Results visualization

We visualize the globally associated attention captured by proposed transformer-based architecture and compare it with the local attention captured by CNN-based architectures. Figure 7 shows the qualitative comparison of both activation maps. Per-former$_{SCM}$ focuses on crisp self-attention regions and maintains their global relationships. All the cases shown in the figure depict the intelligent learning of self-attentions by Per-former$_{SCM}$ with certainly high accuracy.

Although both CNN-based and transformers-based architectures tend to highlight the salient regions of the images, the Per-former$_{SCM}$-based attention maps given in the columns (b) and (d) of the figure illustrate a clear difference from the convolution-based attention maps shown in the columns (a) and (c) for respective input images.

6.2 Ablation study

In the proposed architecture, we choose the depth of layers as per standard architectures of sequence-based deep models, i.e., the first paper of transformer [24] and its extended researches [40,41] etc. We use 12-layers depth in all variants of the proposed models for experimentation and evaluations. However, in the ablation study, we perform detailed experimentation to analyze the impact of different number of layers for person re-id task. The experimental results show that the 12 superimposed layers of the transformer-based re-id model supersedes the other variants in terms of speed and accuracy. The experimental results and discussion are given in Table 5. In later ablation study experiments, we choose 12-layer model to analyze the significance of SCM module at different layers of the network and the impact of different patch size for person re-id.

We employ the SCM module at various positions/ layers of the network to analyze the impact of self-contextual learning at different levels of the model. Additionally, we use different rates of keeping attentive regions and dropping them for learning of the contextual information along with the attention regions.

The experimental results show that due to the inherent capability of the transformers to learn multiple self-attentions among the sequence of inputs (in this case the sequence of patches) the proposed scheme covers the contextual information learning in addition to the attention learning. Therefore, we do not need to drop a bigger ratio of attention in SCM module. Only suppressing top 10% attentive regions with the threshold value of 0.8 works perfect to improve the context learning and to generalize the model.

The significance of SCM module is clearly seen in both the scenarios, i.e., whether it is employed at the final activation map or within each layer of the baseline. In both cases, it improves the model learning and results in an improved average precision and accuracy. However employing self-context mappings at each layer of the baseline architecture provides the best results when compared to the variants having SCM module only at the final activation layer. The comparative results are shown in table 6.

We use Imagenet pretrained weights to start the training process of our model with the compulsion of using predefined patch sizes in our experimentation. However we infer from the visual and quantitative results that the re-id results can even be boosted further by employing different sizes of images and the patches, especially if multi-scale patches are used in an integrated manner. We find that the integration of SCM module in each layer of the baseline architecture consistently improves the generalization of the re-id models for all datasets. As per the best of our knowledge, our baseline model and its extension are the very first vanilla architecture that could achieve such high level performances for person re-id.

For position-wise activation module, we choose hidden layers dimensions four times to the size of initial patch dimensions. We experiment with different size of image patches and attain the best results for the 16 × 16 size patch. A comparison of different patch sizes used is given in Fig. 8. For 16 × 16

Fig. 6 Re-id Results of the proposed network. Top 10 closest matches are shown for respective query images. Green boundaries show the correct matches and the red boundaries show the incorrect matches

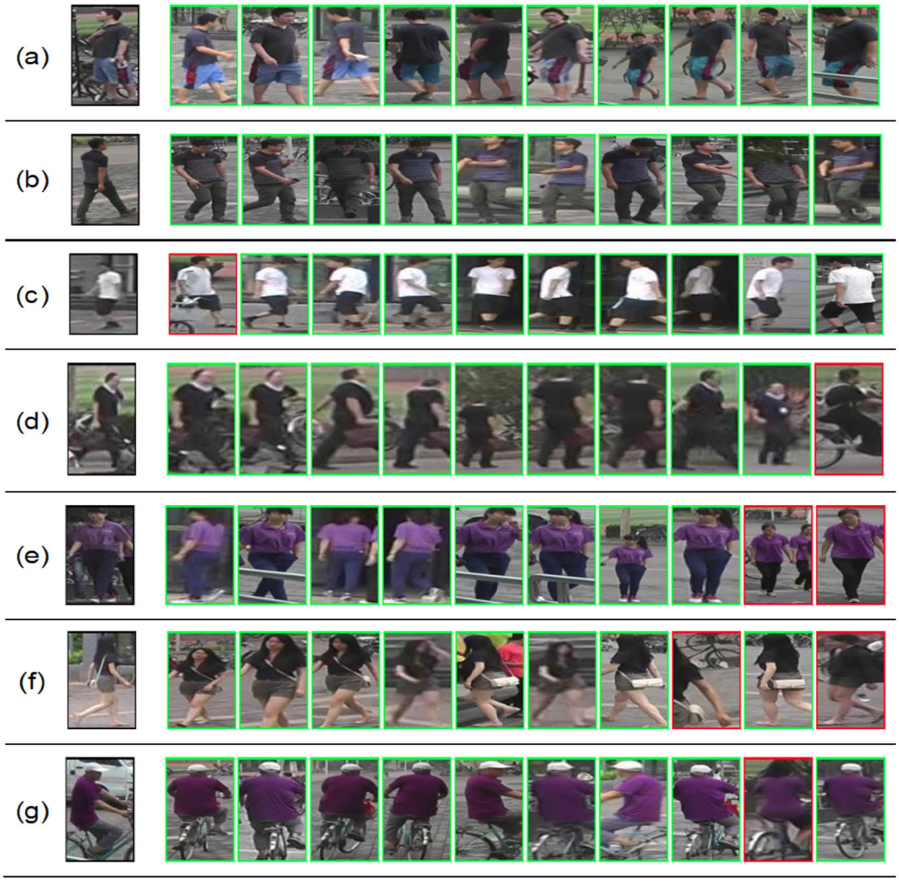


Fig. 7 Class Activation Maps of the Proposed Model. Column **a** and **c** show CNN-based activations of Resnet50. Column **b** and **d** show Per-former$_{SCM}$ activations of the same input image (best viewed in color)

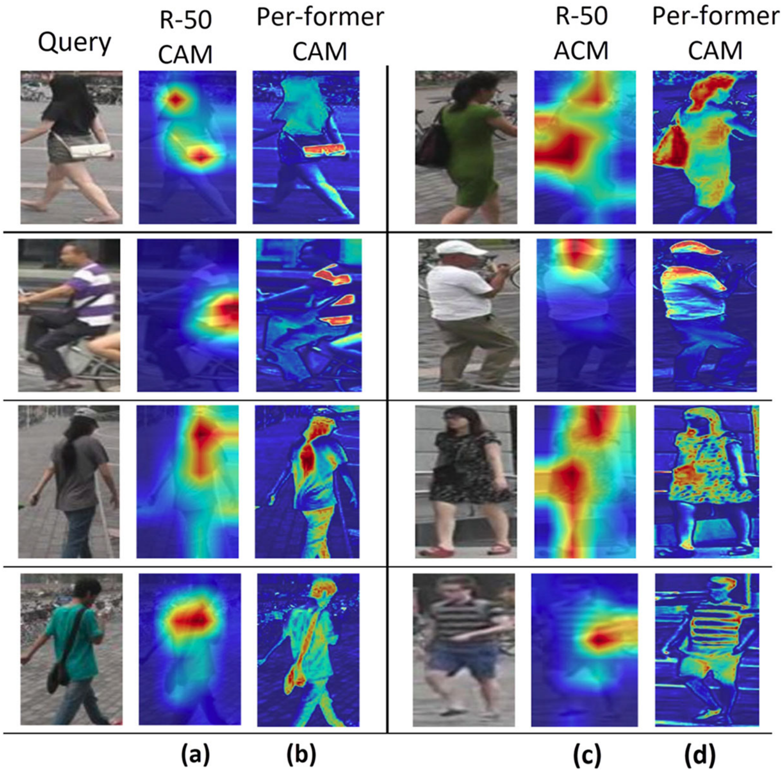


Table 5 Per-former variants are the variants of Per-former$_{base}$ model with different number of layers, where the subscript numbers 4, 8, 10 and 12 show the respective number of layers

Per-former variants	No. of Params	Epoch-2		Epoch-4		Epoch-6		Epoch-8	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1
Per-former$_4$	29.67M	32.7	54.1	46.7	68.1	49.5	71.2	51.1	73.5
Per-former$_8$	58.1M	52.3	73.8	61.1	81.4	62.5	81.7	65.2	84.5
Per-former$_{10}$	72.2M	63.2	81.3	67.9	85.4	67.4	85.9	68.9	86.5
Per-former$_{12}$	86M	**68.1**	**84.8**	**68.5**	**85.7**	**71.3**	**87.6**	**72**	**88.5**

Bold value indicates the superior quantitative performance of the respective methodology

Table 6 Impact of self-context mapping module at various layers of the network—Market-1501 (FLAM = Final layer's activation map; ELAM = Each layer's activation map; TH = Threshold; Attn Drop = Attention Dropout Rate)

Per-former variants	TH	Attn Drop	mAP	R1	R5	R10	R20
Per-former$_{base}$	–	–	72	88.5	94.6	96.6	98.2
Per-former$_{SCMatFLAM}$	0.8	25	63.8	84.8	94.2	96.3	98.0
Per-former$_{SCMatFLAM}$	0.8	10	73.2	88.7	95.9	97.7	98.4
Per-former$_{SCMatELAM}$	0.8	10	**74.5**	**89.4**	**95.6**	**97.6**	**98.5**

Bold value indicates the superior quantitative performance of the respective methodology

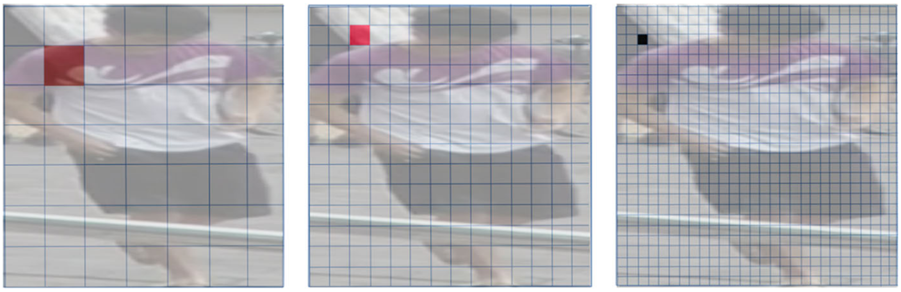


Fig. 8 Visual illustration of image patches sizes used in the proposed model. **a** 32 × 32 pixel patches, **b** 16 × 16 pixel patches, **c** 8 × 8 pixel patches

patch size and 768 embedding dimensions, the hidden layer dimensions are 3072. We choose different patch sizes to find out the optimal patch size for re-id task. Choosing the patches size bigger than 16 × 16 or smaller than 16 × 16 did not work well for person re-id, the experimental results using different patch sizes are given in Table 7, while keeping embedding/hidden dimensions intact does not aid in re-id performance. Moreover, we analyzed the impact of different size of patches and the number of layers upon the performance of person re-id and time complexity. We found that the smaller size of patches and the less number of layers did not perform well due to loss of the contextual information in small patches. The graphical view of the experimental results are shows in the graph 9. The bigger size patches do not encompass the small-scale local attention associations in an image, therefore all proposed experimental results are based on the patch size of 16 × 16 with hidden layers dimensions of 3072. However the impact of SCM module is consistent for all sizes of image patches and improves the generalization of re-id network over baseline network as given in Table 7.

Table 7 Impact of various sizes of image patches for all variants of Per-former networks—Market-1501

Per-former variants	Patch size	mAP	R1	R5	R10	R20
Per-former$_{base}$	32 × 32	66.6	85	92.6	95	96.7
Per-former$_{base}$	16 × 16	72	88.5	94.6	96.6	98.2
Per-former$_{base}$	8 × 8	31	67	80.7	85.5	88.7
Per-former$_{SCM}$	32 × 32	70.8	87.6	94.8	96.9	98.1
Per-former$_{SCM}$	16 × 16	**74.5**	**89.4**	**95.6**	**97.6**	**98.5**
Per-former$_{SCM}$	8 × 8	32.0	67.3	81.3	85.9	89.1

Bold value indicates the superior quantitative performance of the respective methodology

The proposed self-attention-based re-id models converge quickly when compared to the CNN-based re-id model, i.e., ResNet50. Both the Per-former$_{base}$ and Per-former$_{SCM}$ evaluation results attain around 84% rank-1 accuracy and 68% mean average precision right after the second epoch of the training while using the same batch size (i.e., 128), optimizer (i.e., radam) and learning rate (i.e., 0.0003). Per-former models' second epoch results are comparable to various vanilla CNN-based re-id models' peak results. However the extended model Per-former$_{SCM}$ improves its learning at later stages of the training and surpasses the performance of Per-former$_{base}$ for both performance metrics.

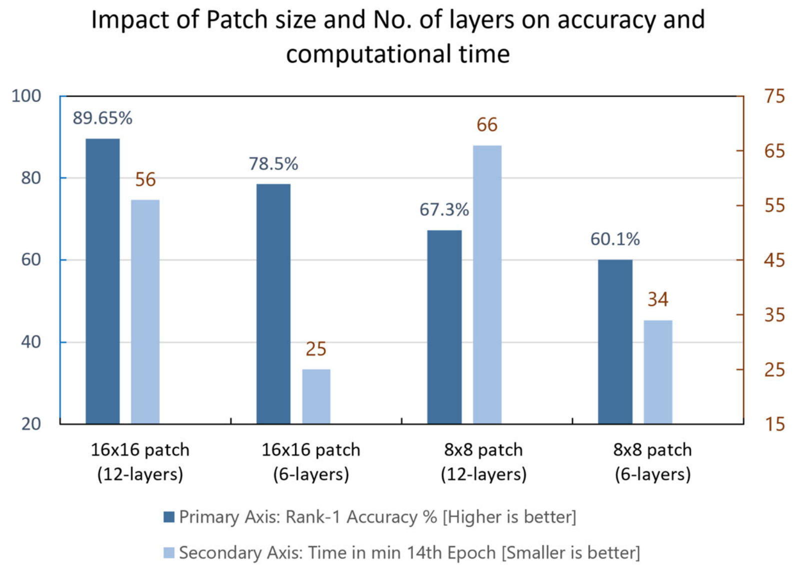


Fig. 9 Impact of size of patches and number of layers upon re-id performance and computational time—(For the patch size 8 × 8, the best re-id performance is recorded at 14th epoch, that is why the performance comparison is made for the same epoch for 16 × 16 patch size)

Table 8 Computational efficiency of the proposed model in terms of speed and scalability for Market-1501 Dataset. The transformer-based proposed models attain clearly higher performance than vanilla CNN model, i.e., ResNet-50

Model	Epoch-2		Epoch-4		Epoch-6		Epoch-8	
	mAP	R1	mAP	R1	mAP	R1	mAP	R1
ResNet-50	18.1	33.0	23	48.3	35.1	53.0	44.8	65.1
Per-former$_{base}$	68.1	84.8	68.5	85.7	71.3	87.6	72	88.5
Per-former$_{SCM}$	67.8	83.3	73.1	85.7	73.3	88.7	74.5	89.4

Bold value indicates the superior quantitative performance of the respective methodology

When the extended proposed model Per-former$_{SCM}$ is compared with the baseline model for Market1501 re-id dataset, it surpasses the performance of baseline architecture Per-former$_{base}$ by 2.5 points in average precision and around 1 point in re-id accuracy at rank-1.

Likewise for the DukeMTMC-ReID re-id benchmark, a substantial improvement of Per-former$_{SCM}$ is seen over the Per-former$_{base}$ network. An increase of 4.6 points in mAP value indicates that integrating self-context information into the self-attention learning to represent a person significantly improves the average precision of correct re-id of each query image. Rank-1 accuracy is also increased to around 4 point for DukeMTMC-ReID data set.

Consistency of the proposed re-id solution is seen for the large scale and complex dataset MSMT-17. The proposed extended model surpasses the baseline performance with around 4 points in both the rank-1 accuracy and mean average precision. Main reason for improved learning by our vanilla model and its proposed extension is that the SCM module learns the self-contextual information along with the global self-attentions of patches. The computational efficiency of the proposed model in contrast to the CNN baseline model ResNet-50 is given in Table 8. However, the learnable parameters of proposed Per-former models are 86 M that is much greater than the number of parameters of ResNet50, i.e., 23.5 M. However the proposed model converges very quickly in few epochs as compared to its CNN-based counterparts.

7 Conclusion

Since the shifting of re-id research from handcrafted algorithms to the deep learning techniques, the convolutional neural networks have been explored with a great dominance to develop the person re-id solutions. As the CNN-based architectures are bound to learn the distinctive features in a neighborhood region, these do not capture the associations among the discriminative features that lie at spatially distant regions within an image. However, for person re-identification problem the association among distant attentive regions plays an essential role. The human vision system identifies a person by associating multiple regions in a given image. For instance, associating the discriminative attributes of a person that are robust to illumination, pose variations and camera view changes. Taking inspiration from this, we imitate such human cognitive process in our baseline model Per-former$_{base}$ to build the attention associations among far-distant regions in the image of a person. We further extend the baseline model and introduce a self-context mapping module at each layer of the base network in Per-former$_{SCM}$. The extended model learns self-contextual mapping among far-distance regions (patches) of the images in addition of maintaining the self-attention mappings among the patches of an image. To the limit of our knowledge, this task is one among the first few application of transformer-based architectures in the context of person re-identification. The evaluation results and their visualization discussed in Sects. 5 and 6 show the superior performance of proposed models over existing vanilla CNN re-id models. A significant improvement in the person re-id accuracy and mean average precision is seen for all datasets used to evaluate this work, i.e., 5.5% accuracy improvement for Market1501, 4.6% for DukeMTMC-ReID and 17% for MSMT-17 dataset confirms the outstanding performance of the proposed vanilla architecture as compared to CNN vanilla models. Further, the ablation studies discuss the dominance of Per-former$_{SCM}$ over the baseline model Per-former$_{base}$. In the future, we also plan to enhance the proposed transformer-based model by applying specialized customization in this work [53,54]. We plan to explore the omni-scale image patches to learn multi-scale self-attention and self-context mappings for the supervised person re-id.

Declarations

Conflict of interest The authors declare no conflict of interests. All authors contributed to the study conception and design, material preparation, data collection and analysis. All authors read and approved the final manuscript.

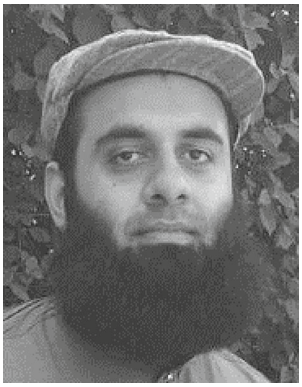
References

1. Zahra, A., Perwaiz, N., Shahzad, M., Fraz, M.M.: Person re-identification: A retrospective on domain specific open challenges and future trends. arXiv:2202.13121 (2022)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
3. Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep Learning, vol. 1. MIT Press, Cambridge (2016)
4. Jia, Z., Li, Y., Tan, Z., Wang, W., Wang, Z., Yin, G.: Domain-invariant feature extraction and fusion for cross-domain person re-identification. The Visual Computer, 1–12 (2022)
5. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)
6. Zhao, L., Li, X., Zhuang, Y., Wang, J.: Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3219–3228 (2017)
7. Bai, X., Yang, M., Huang, T., Dou, Z., Rui, Yu., Yongchao, X.: Deep-person: learning discriminative deep features for person re-identification. Pattern Recogn. **98**, 107036 (2020)
8. Perwaiz, N., Moazam, M., Shahzad, F.M.: Person re-identification using hybrid representation reinforced by metric learning. IEEE Access **6**, 77334–77349 (2018)
9. Batool, S., Zeeshan, M., Muhammad, Shahzad, A., Fraz, M.M.: End to end person re-identification for automated visual surveillance. In: IEEE International Conference on Image Processing, Applications and Systems, IPAS 2018, Sophia Antipolis, France, December 12–14, 2018, pp. 220–225 (2018)
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30**, 5998–6008 (2017)
11. Wei, D., Wang, Z., Luo, Y.: Video person re-identification based on rgb triple pyramid model. The Visual Computer, 1–17 (2022)
12. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2018)
13. Perwaiz, N., Fraz, M.M., Shahzad, M.: Hierarchical refined local associations for robust person re-identification. In: 2019 International Conference on Robotics and Automation in Industry (ICRAI), pp. 1–6. IEEE (2019)
14. Li, Yang, Huahu, Xu.: Deep attention network for rgb-infrared cross-modality person re-identification. J. Phys.: Conf. Ser. **1642**, 012015 (2020)
15. Si, J., Zhang, H., Li, C.-G., Kuen, J., Kong, X., Kot, A.C., Wang, G.: Dual attention matching network for context-aware feature sequence based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5363–5372 (2018)
16. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2017)
17. Mubariz, N., Mumtaz, S., Hamayun, M.M., Fraz, M.M.: Optimization of person re-identification through visual descriptors. In: Proceedings of (VISIGRAPP 2018) - Volume 4: VISAPP, Funchal, Madeira, Portugal, January 27–29, 2018, pp. 348–355 (2018)
18. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.,: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
19. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301 (2018)
20. Hermans, A., Beyer, L., Leibe, B.,: In defense of the triplet loss for person re-identification. arXiv:1703.07737 (2017)
21. Mumtaz, S., Mubariz, N., Saleem, S., Fraz, M.M.: Weighted hybrid features for person re-identification. In: Seventh International Conference on Image Processing Theory, Tools and Applications, IPTA 2017, Montreal, QC, Canada, November 28–December 1, 2017, pp. 1–6 (2017)
22. Faizan, R., Fraz, M.M., Shahzad, M.: Iab-net: Informative and attention based person re-identification. In: 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2), pp. 1–5. IEEE (2021)
23. Perwaiz, N., Fraz, M.M., Shahzad, M.: Stochastic attentions and context learning for person re-identification. PeerJ Comput. Sci. **7**, e447 (2021)
24. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Mostafa, Dehghani, Matthias, Minderer, Georg, Heigold, Sylvain, Gelly, Jakob, Uszkoreit, Neil, Houlsby: An image is worth 16x16 words: Transformers for image recognition at scale (2021)

25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
27. Ansar, W., Fraz, M.M., Shahzad, M., Gohar, I., Javed, S., Jung, S.K.: Two stream deep CNN-RNN attentive pooling architecture for video-based person re-identification. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications—23rd Iberoamerican Congress, CIARP 2018, Madrid, Spain, November 19-22, 2018, Proceedings, pp. 654–661 (2018)
28. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: past, present and future. arXiv:1610.02984 (2016)
29. Huang, G., Liu, Z., Van Der L., Kilian, M., Weinberger, Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
30. Sun, Y., Zheng, L., Deng, W., Wang, S.: Svdnet for pedestrian retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3800–3808 (2017)
31. Han, K., Guo, J., Zhang, C., Zhu, M.: Attribute-aware attention model for fine-grained representation learning. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 2040–2048 (2018)
32. Perwaiz, N., Fraz, M.M., Shahzad, M.: Smart visual surveillance: Proactive person re-identification instead of impulsive person search. In: 2020 IEEE 23rd International Multitopic Conference (INMIC), pp. 1–6. IEEE (2020)
33. Wang, G., Lai, J., Huang, P., Xie, X.: Spatial-temporal person re-identification. Proc. AAAI Conf. Artif. Intell. **33**, 8933–8940 (2019)
34. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.: Abd-net: Attentive but diverse person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8351–8361 (2019)
35. Guangyi Chen, Chunze Lin, Liangliang Ren, Jiwen Lu, Jie Zhou: Self-critical attention learning for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 9637–9646 (2019)
36. Chen, Z., Lv, X., Sun, T., Zhao, C., Chen, W.: Flag: feature learning with additional guidance for person search. Vis. Comput. **37**(4), 685–693 (2021)
37. Chen, Y., Wang, H., Sun, X., Fan, B., Tang, C., Zeng, H.: Deep attention aware feature learning for person re-identification. Pattern Recogn. **126**, 108567 (2022)
38. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. abs/1710.06555 (2017)
39. Luo, H., Fan, X., Zhang, C., Jiang, W.: Stnreid : Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification (2020)
40. Radford, A., Jeffrey, W., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
41. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.,: BERT: Pre-training of deep bidirectional transformers for language understanding, pp. 4171–4186 (2019)
42. Hendrycks, D., Gimpel, K.,: Gaussian error linear units (gelus). arXiv:1606.08415 (2016)
43. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J.,Tian, Q.: Scalable person re-identification: A benchmark. In: ICCV, IEEE Computer Society, pp. 1116–1124 (2015)
44. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.,: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision, pp. 17–35. Springer (2016)
45. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 79–88 (2018)
46. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.,: On the variance of the adaptive learning rate and beyond. arXiv:1908.03265 (2019)
47. Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q.,: Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3960–3969 (2017)
48. Li, W., Zhu, X., Gong, S.,: Person re-identification by deep joint learning of multi-loss classification. arXiv:1705.04724 (2017)
49. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camstyle: A novel data augmentation method for person re-identification. IEEE Trans. Image Process. **28**(3), 1176–1190 (2018)
50. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned cnn embedding for person reidentification. ACM Trans. Multim. Comput. Commun. Appl. (TOMM) **14**(1), 1–20 (2017)
51. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)
52. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)
53. Luo, H., Jiang, W., Fan, X., Zhang, C.: Stnreid: deep convolutional networks with pairwise spatial transformer networks for partial person re-identification. IEEE Trans. Multim. **22**(11), 2905–2913 (2020)
54. He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W.: Transformer-based object re-identification, Transreid (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

N. Pervaiz She has obtained her PhD degree in Computer Science from the School of Electrical Engineering and Computer science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan, and serves as Post Doc Research Assistant. Her research interests include Person Reidentification and Biometrics.



M.M. Fraz He has obtained PhD from Kingston University London, UK. After that he had been Post Doc research fellow at the Kingston University and the University of Warwick, UK. At present, besides working as an Associate Professor at the NUST-SEECS, Islamabad, he is Rutherford Visiting Fellow at The Alan Turing Institute, London, which is UK's National Center for Data Science and AI. His research interests include medical image analysis, automated visual surveillance, person reidentification and visual recognition.

M. Shahzad He has obtained PhD from Technical University of Munich Germany, and at present working as Associate Professor (W2) at the same university. His area of expertise is 3D Computer Vision, Point Cloud processing and Human Activity Recognition.