VAC-Net: Visual Attention Consistency Network for Person Re-identification

Weidong Shi
Northeastern University
Shenyang, Liaoning, China
shiweidong1003@gmail.com

Yunzhou Zhang*
Northeastern University
Shenyang, Liaoning, China
zhangyunzhou@mail.neu.edu.cn

Shangdong Zhu
Northeastern University
Shenyang, Liaoning, China
zhushangdong@gmail.com

Yixiu Liu
Northeastern University
Shenyang, Liaoning, China
liuyixiuasd130@gmail.com

Sonya Coleman
University of Ulster
York Street, Belfast, United Kingdom
sa.coleman@ulster.ac.uk

Dermot Kerr
University of Ulster
York Street, Belfast, United Kingdom
d.kerr@ulster.ac.uk

ABSTRACT

Person re-identification (ReID) is a crucial aspect of recognising pedestrians across multiple surveillance cameras. Even though significant progress has been made in recent years, the viewpoint change and scale variations still affect model performance. In this paper, we observe that it is beneficial for the model to handle the above issues when boost the consistent feature extraction capability among different transforms (e.g., flipping and scaling) of the same image. To this end, we propose a visual attention consistency network (VAC-Net). Specifically, we propose Embedding Spatial Consistency (ESC) architecture with flipping, scaling and original forms of the same image as inputs to learn a consistent embedding space. Furthermore, we design an Input-Wise visual attention consistent loss (IW-loss) so that the class activation maps(CAMs) from the three transforms are aligned with each other to enforce their advanced semantic information remains consistent. Finally, we propose a Layer-Wise visual attention consistent loss (LW-loss) to further enforce the semantic information among different stages to be consistent with the CAMs within each branch. These two losses can effectively improve the model to address the viewpoint and scale variations. Experiments on the challenging Market-1501, DukeMTMC-reID, and MSMT17 datasets demonstrate the effectiveness of the proposed VAC-Net.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Computer vision**; **Computer vision problems**; **Object recognition**;

KEYWORDS

Person re-identification, Viewpoint change, Scale variations, Visual attention

ACM Reference Format:

Weidong Shi, Yunzhou Zhang, Shangdong Zhu, Yixiu Liu, Sonya Coleman, and Dermot Kerr. 2022. VAC-Net: Visual Attention Consistency Network for Person Re-identification. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR ’22), June 27–30, 2022, Newark, NJ, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3512527.3531409

1 INTRODUCTION

Person re-identification (ReID) is the task of recognizing the same person across multiple non-overlapping cameras. Although significant progress has been made in the past few years, many challenges still exist. Due to the different viewpoints and the distances of the person from the cameras, there are diverse viewpoint variations and scale variations in images, which affects the recognition performance of the person ReID model. To improve the problem caused by viewpoint variations, [4, 21] and [35] proposed a rotational convolution network and a rotational invariant network, respectively, to extract viewpoint invariant features. However the above works have limited capability in richer viewpoint variations. Furthermore, based on class activation maps (CAMs) [27, 45] methods, [41] proposed the Siamese network to promote consistent attention regions between pairs of the images from the same identity. Nevertheless, we argue that model should have consistent attention regions for different transforms of the same image and a little difference between pairs of images from the same identity.

To deal with the issue caused by scale variations, [1, 14, 23, 46] designed multi-scale CNN architectures. In particular, [23] leveraged multiple convolutional feature streams with different receptive fields to generate multi-scale features, yet it contained a large number of model parameters. Considering that, [46] designed a lightweight omni-scale network (OSNet) with pointwise and depthwise convolutions. Although OSNet is a lightweight network and excellent performance can be achieved, it is overly computationally complex and time-consuming. Unlike the above approaches which only use single-scale images, [3, 34] adopt multi-scale images as inputs to train the network. Specifically, [3] proposed a Deep Pyramid Feature Learning (DPFL) architecture in which different sizes of the same image were fed into different branches to acquire scale-specific discriminative information. [34] send the full scale, half scale and top half (cropped image comprising of the persons identity above the waist) of the person images into three separate convolutional streams to obtain a larger number of highly discriminative features. However, both [3] and [34] concatenate multi-scale

[*] Corresponding author

ICMR ’22, June 27–30, 2022, Newark, NJ, USA

ACM ISBN 978-1-4503-9238-9/22/06…$15.00
https://doi.org/10.1145/3512527.3531409

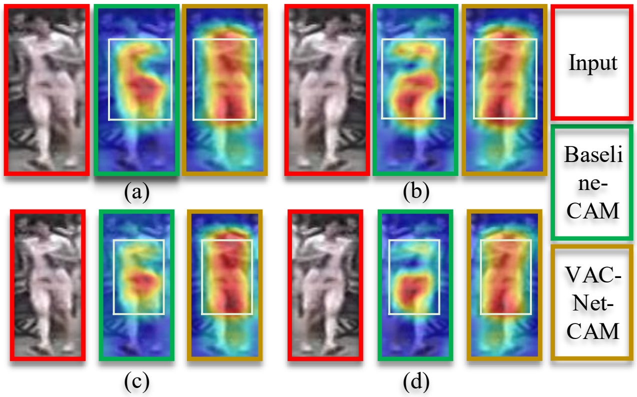


Figure 1: Visualization of CAMs. (a) Original images; (b) Flipped images; (c) Scaled images; (d) Flipped and Scaled images. The red, green and brown borders represent the inputs, the class activation maps of baseline and our proposed method, respectively.

specific features as final hyper-descriptors to achieve better performance, which incurs a high computation cost in the retrieval process.

Unlike the above works, we design a simple yet effective method named VAC-Net to address the viewpoint changes and scale variations at the same time. From the green border of the Figure. 1, we can easily observe that the simple model (i.e., baseline) is unable to focus on the same regions of interest among different transforms of the same image. Thus, we argue that the viewpoint changes and scale variations may be solved if the model have consistent feature extraction ability (i.e., brown borders) among different transforms of the same image. To the end, we design an Embedding Space Consistent (ESC) architecture comprised of three identical branches with three transforms (i.e., origin, flipping and scaling) from the same image as inputs. Attribute to ESC architecture, the model is capable of exploiting more consistent embedded features among the three transforms. In addition, we design an input-wise visual consistency loss to promote the advanced semantic information consistency among the three transforms. Finally, we propose a layer-wise visual consistent loss to further enhance the consistency of the different stages in each branch. The above two losses can further improve the consistency of the embedded feature space among the three branches, as shown in the brown border of the Figure. 1.

Therefore, the main contributions are as follows:

- We propose an embedding spatial consistency architecture by learning a consistent embedding feature space for different transforms of the same image to better handle person viewpoint changes and scale variations.
- We propose an input-wise visual consistent loss to promote the model by aligning the consistency of the last advanced semantic features among the three transforms;
- We propose a layer-wise visual consistency loss to guide the model to focus on the consistency of feature regions extracted from different stages of each branch.

2 RELATED WORK

Most existing person ReID methods focus on two important aspects: discriminative feature representations [12, 33] and effective similarity metrics [18, 37]. The viewpoint changes and scale variations make it important to extract visual attention consistent features for person ReID. Then it is expected that an effective similarity metric will measure the similarities among persons using the given embedded features.

Deep learning was first applied to person ReID by [16, 38], and the rapid development of CNNs has greatly boosted the progress of person ReID in recent years. Most works pay more attention to learning discriminative features, effective distance measurements or both in combination. For feature learning methods, [28] designed a two-stream network architecture to generate appearance representations and part representations, which were further aggregated to generate the part-aligned features. In [31] a Part-based Convolutional Baseline (PCB) was proposed to extract part-level features. Considering the relations between individual body parts and the entire body, a relation network for person ReID was proposed in [22] to increase the accuracy of the model. For the similarity metrics, [5] first applied the triplet loss [26] which was proposed for face recognition to improve the performance in ReID. Further, [3] analyzed the relationship of inter- and intra-class distances and proposed a deep quadruplet network. In the most recent metric learning work, [30] designed a circle loss to generate an unified formula for triplet loss and softmax cross-entropy, i.e., learning with class labels and pair-wise labels. In this paper, we also propose a visual loss to improve visual attention consistency and guide the model to achieve excellent performance.

3 METHOD

The technical details about VAC-Net will be presented in this section. Firstly, we simply describe the architecture of the backbone in 3.1. Then, we introduce the overall proposed architecture for the Visual Attention Consistency Network in 3.2. Finally, we elaborate on the proposed input-wise visual consistent loss and layer-wise visual consistent loss in 3.3 and 3.4, respectively.

3.1 Architecture Overview

The overall architecture of the proposed network is illustrated in Figure. 2. We utilize ResNet50 [9], which is extensively used in previous ReID works [19, 39, 42], as the backbone model. The selected model includes some modifications in line with recent work [36]. We remove the last fully connected (FC) layer and add a batch normalization (BN) layer and a 1×1 convolutional layer in front of the global average pooling layer. Analogous to [7, 31], the last spatial down-sampling operation is also removed to increase the spatial resolution. It is worth noting that the three branches share weights and we only employ the features from the original branch as the person feature descriptors for the subsequent Re-ID testing task.

3.2 Embedding Spatial Consistency

In practical application scenarios, the cropped person images have diverse scale variations and viewpoint changes due to different viewpoints and distances of the person from the cameras, which presents

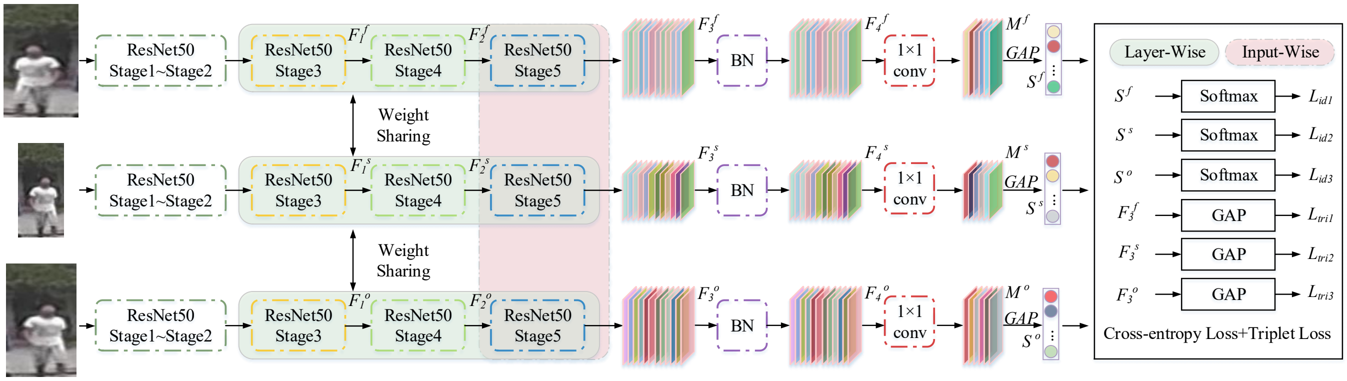


Figure 2: Architecture of VAC-Net. The branches from top to bottom represent flipping branch, scaling branch and original branch, respectively. The three transforms of the input images are simultaneously send into the network to calculate loss and update the model in the training process. However, in the testing process, the output of the original branch with the original image as input is generated as the final feature for the subsequent retrieval task.

huge challenges for the person ReID task. Unlike [35, 41, 46], who designed specific network architectures and [3, 34], who concatenated specific features, our approach is based on the idea that humans have a consistent ability to perceive different transforms of the same target. We firstly design an embedding spatial consistency (ESC) architecture to guide the CNN to align embedded features consistently. Specifically, I_f, I_s, I_o are represented the flipping transform (simulating viewpoint variance), scaling transform (simulating scale variance) and the original image, respectively. During the training stage, the three input images I_f, I_s, I_o are simultaneously passed through the the backbone network to generate the features $F_3^f, F_3^s, F_3^o \in \mathbb{R}^{B\times D\times H\times W}$ (see Figure. 2) and B, D, H, W are the number of mini-batches, channels, height and width, respectively. Then, we feed the tensors F_3^f, F_3^s, F_3^o into the batch normalization layer (BN layer in Figure. 2) to generate $F_4^f, F_4^s, F_4^o \in \mathbb{R}^{B\times D\times H\times W}$. Meanwhile, the tensors F_3^f, F_3^s, F_3^o are simultaneously input into a global average pooling (GAP) layer to generate the feature vectors $e \in \mathbb{R}^{B\times D}$ (for clarity, we omit the f, s and o superscripts hereafter) which are then applied to calculate the triplet loss [10] as illustrated in the box on the right of Figure. 2 and defined as:

$$L_{tri}=\frac{1}{B}\sum_{i=1}^{P}\sum_{a=1}^{K}[m+\max_{p=1...K}\left\|e_a^i-e_p^i\right\|_2-\min_{\substack{n=1...K\\ j=1...P\\ j\neq i}}\left\|e_a^i-e_n^j\right\|_2]_+ \tag{1}$$

where e_a^i, e_p^i, e_n^j are features of anchor, positive and negative samples respectively, P and K are the number of identities in each mini-batch and the number of images for each identity, respectively and m is the margin. Then, we send F_4 to the 1×1 convolution operation to acquire the class activation map $M \in \mathbb{R}^{B\times C\times H\times W}$, which can be formulated as:

$$M = W \otimes F_4 \tag{2}$$

where $\otimes$ is a convolution operation, $W \in \mathbb{R}^{1\times 1\times C\times D}$ is the convolutional kernel weights, and C is the number of person classes. Finally, a GAP operation is applied to obtain the class score of the tensor M.

The class score $S \in \mathbb{R}^{B\times C}$ is sent to a softmax layer to calculate the identification loss L_{id} as follows:

$$L_{id} = -\sum_{b=1}^{B}\log\frac{\exp^{(S_{l_b})}}{\sum_{c=1}^{C}\exp^{(S_c)}} \tag{3}$$

where, $l_b{}_{b=1}^{B} \in 1, ..., C$ is the identity label of each image and S_{l_b} is the output of class l_b from the classifier.

Intuitively, the ESC architecture has the ability to guide the network to generate more consistently embedded features. However, we argue that only relying on the ESC architecture is weak as it lacks a suitable supervision signal and hence we further design two simple yet effective visual supervised losses.

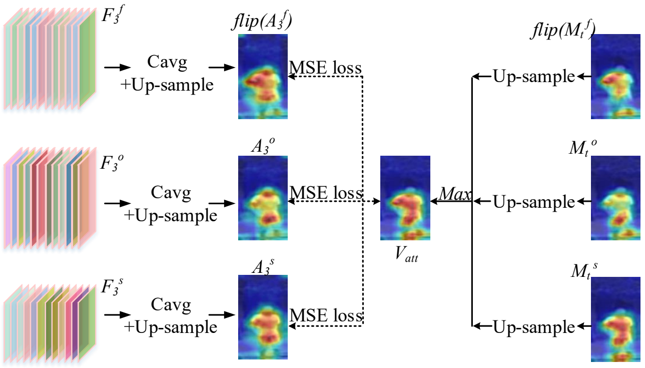


Figure 3: Illustration of the input-wise visual consistent loss

3.3 Input-Wise Visual Consistent Loss

Although ESC achieves a performance improvement by utilizing an identical architecture and shared parameters, the performance gains are very limited under supervision of the classifier. Therefore, we further design a Input-Wise consistent loss (IW-loss) to supervise the advanced visual consistency of the different transforms in the same image. Considering the class activation map presents the model's attention regions, we select a class activation map $M_t \in \mathbb{R}^{B\times1\times H\times W}$ for each transform according to the label t. Specifically, we obtain a more effective supervised signal $V_{att} \in \mathbb{R}^{B\times1\times H\times W}$ by selecting the maximum response from the three CAMs (i.e., M_t^f, M_t^s, M_t^o), which can be formulated as:

$$V_{att} = \max(US(flip(M_t^f)), US(M_t^s), US(M_t^o)) \tag{4}$$

where the $\max(\cdot)$ selects the maximum spatial response value along the channel dimension and $flip(\cdot)$ and $US(\cdot)$ are the flipping operation and up-sample operation, respectively. To achieve better embedded spatial consistency, V_{att} is adopted to supervise the output(i.e., F_3^f, F_3^s, F_3^o in Figure 3) generated from the last stage of the backbone network. In detail, we utilize average pooling along the channel dimension (i.e., $Cavg$ in Figure. 3) to obtain A_3^f, A_3^s, A_3^o, as follows:

$$A_3^o(x,y) = US(Cavg(F_3^o)) = US(\frac{1}{D}\sum_{d=1}^{D} F_{3_d}^o(x,y)) \tag{5}$$

A_3^f and A_3^s are calculated in the same way as A_3^o. The input-wise visual consistent loss L_{IW} can be formulated as:

$$L_{IW} = \frac{1}{3}(||V_{att} - flip(A_3^f)||_2 + ||V_{att} - A_3^s||_2 + ||V_{att} - A_3^o||_2) \tag{6}$$

Under the supervision of the IW-loss, the model can focus on more consistent visual regions of interest.

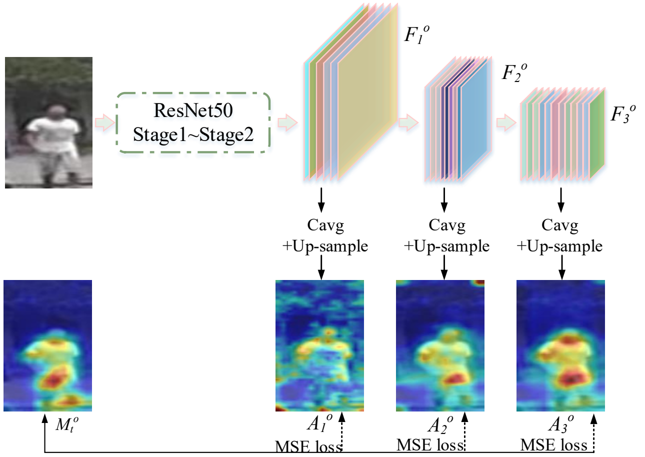


Figure 4: Illustration of the layer-wise visual consistent loss

3.4 Layer-wise Visual Consistent Loss

Unlike IW-loss which concerns the visual consistency among the different transforms of the same image, we think the network should pay attention to consistently discriminative regions at different semantic stages in each branch, which has two advantages: 1) it can guide the network to focus on more effective regions of a person's body; 2) it is beneficial to further align the embedding feature space. Therefore, we propose the Layer-Wise visual consistent loss (LW-loss) as illustrated in Figure 4. Specifically, we apply Eq. 5 to the output of stage3~stage5 in each branch to generate A_1^f, A_2^f, A_3^f, A_1^s, A_2^s, A_3^s and A_1^o, A_2^o, A_3^o. Then the layer-wise visual consistent loss can be defined as:

$$L_{LW} = \frac{1}{|G|}(\frac{1}{N}\sum_{j\in G}\sum_{i=1}^{N}(||M_t^j - A_i^j||_2)) \tag{7}$$

where, $G = \{f, s, o\}$ is the set of the flipped, scaled and original transforms and N=3. The experimental results show that the layer-wise visual consistency loss has a more positive effect on optimizing the model.

The overall training loss of the proposed approach is formulated as follows:

$$L = \frac{1}{N}\sum_{n\in G}(L_{id}^n + L_{tri}^n) + \lambda(L_{IW} + L_{LW}) \tag{8}$$

The hyperparameter λ is used to balances the terms in Eq. 8.

4 EXPERIMENTS

In this section, we evaluate the performance of the proposed method using three large public image-based person ReID datasets. We first introduce the datasets and implementation details, respectively. Then we present an ablation study of our method. Finally, we compare the proposed VAC-Net method with state-of-the-art methods.

4.1 Datasets and Evaluation Metrics

We select three large-scale person ReID datasets to evaluate our approach, including Market-1501 [40], DukeMTMC-reID [43] and MSMT17 [32].

Market-1501 contains 32,668 images of 1,501 identities captured by one low-resolution camera and 5 high-resolution cameras. All images utilize the deformable part model (DPM) [6] as a person detector. The dataset is divided into a training set and testing set where 12,936 images from 751 persons are used as the training set and the remaining 750 persons are adopted to generate the query and gallery sets. The query and gallery sets contain 3,368 and 19,734 images respectively. To increase the difficulty of the ReID problem, 2,798 images with just body parts or background are additionally added into the gallery set.

DukeMTMC-reID consists of 16,522 training images of 702 identities and the remaining 702 identities are used as the testing set with 2,228 query images and 17,661 gallery images. All images extracted from the DukeMTMC [25] tracking dataset are captured from 8 high-resolution cameras. As there are large variations within the same identity and high appearance similarity across persons, the dataset is extremely challenging.

MSMT17 is composed of 126,441 training images from 4,101 identities, which is a recently released large-scale person ReID dataset. The images are captured by 3 indoor cameras and 12 outdoor cameras and persons are detected using Faster CNN [24] algorithm. Initially we consider 1,041 identities, and apply 30,248 images from these identities as the training dataset and the remaining images are used to form the validation dataset. The other 3,060 identities are selected as the testing set with 11,659 query images and 82,161 gallery images. It is worth noting that the training set does not contain the images of the 3,060 identities. Due to MSMT17 containing more images and identities with more camera views, more complex backgrounds and differing illumination changes, etc, it is considered to be a more challenging dataset than Market-1501 and DukeMTMC-reID.

Following the majority of similar works on person ReID, we utilize two common evaluation metrics, Cumulative Match Characteristic (CMC) [40] and mean Average Precision (mAP) [8] to appraise the performance of each proposed method. It is worth noting that all experiments are conducted with the single query setting.

4.2 Implementation Details

We employ ResNet50 [9] pre-trained on the ImageNet dataset as our backbone network. The original, flipped and scaled images are resized to 384 × 128, 384 × 128 and 336 × 112, respectively. We choose images corresponding 16 individual pedestrians, 4 images for each, as the inputs in a mini-batch. Note that at the testing process, we only use the original branch to obtain the final feature and apply the cosine distance to evaluate the similarity of the query image and gallery image. In the training process we utilize the Adam optimizer [13] with a weight decay of 5×10^{-4} to update the network. Following the approach in [20], the warm-up adjustment strategy is applied to fine-tune the classifier parameters for 10 epochs with the learning rate gradually increased from $3.5e^{-6}$ to $3.5e^{-4}$, and another 50 epochs with an initial learning rate $3.5e^{-4}$ dropped to 0.1× at epoch 30 and 50, respectively. The parameter λ is set to 0.01. Experiments are conducted on one NVIDIA TITAN GPU, and our method is implemented using Pytorch.

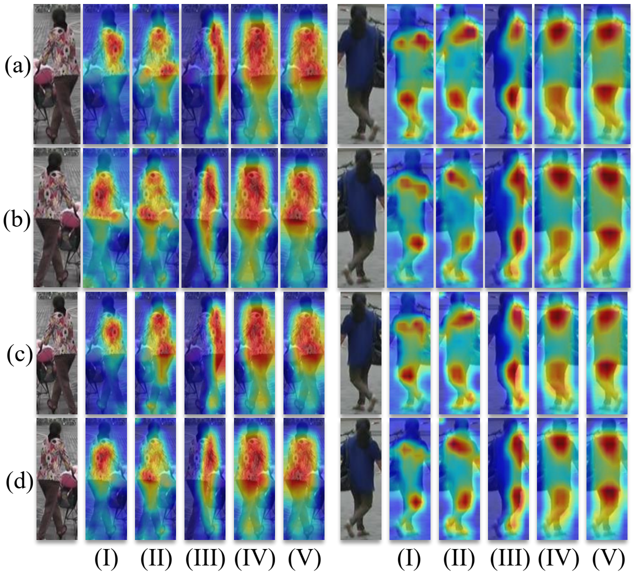


Figure 5: Visualization of CAMs. (I): Baseline; (II): ESC; (III): ESC+IW-loss; (IV): ESC+LW-loss; (V): ESC+IW-loss+LW-loss. (a),(b),(c),(d) are the original, flipped, scaled and flipped+scaled input, respectively. As shown in column (V), VAC-Net spans more diverse attention regions over the whole person.

4.3 Ablation Study

4.3.1 Qualitative Analysis. As shown in Figure 5, compared with the baseline (I), the ESC architecture (II) has more diverse discriminative regions than baseline; which is the reason why ESC architecture achieves better performance than baseline. When we add IW-loss (III) and LW-loss (IV), respectively to the ESC architecture, the network pays attention to more abundant regions than ESC(II), and the LW-loss is more powerful in guiding the network to focus on consistent visual regions. Finally, when we simultaneously add IW-loss and LW-loss to the ESC architecture (V), the discriminative regions are more diverse and visual attention consistency is more obvious, e.g., the legs of the person in the blue cloth, which implies that better visual attention consistency leads to better consistent embedding feature space. In addition, as shown in Figure. 6, when inputting the original image as the query, the baseline has two incorrect results while our method achieves correct results in person retrieval. It is worth noting that when inputting the transforms of the original image, our incorrect results are ranked at the end and our query results have higher overlap, which proves VAC-Net is an

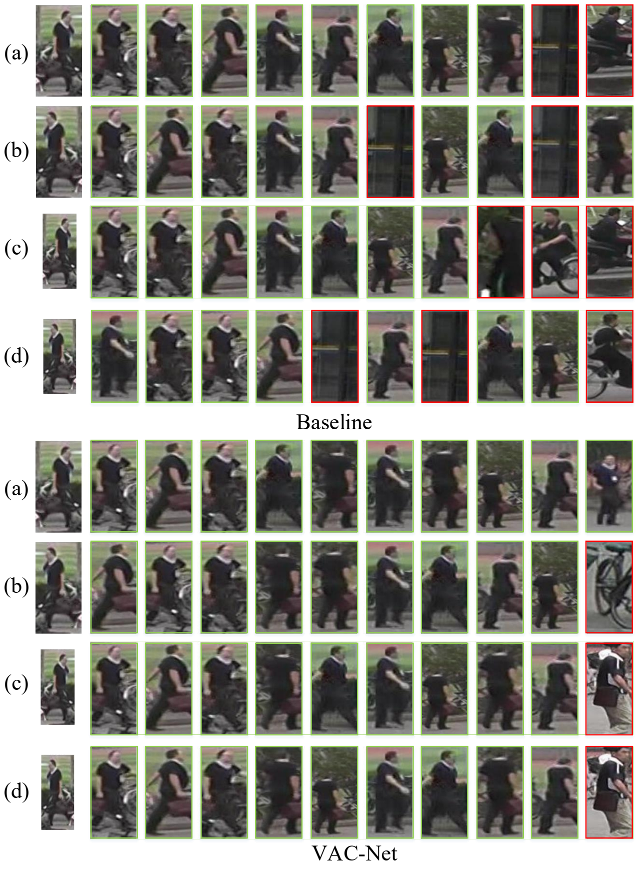
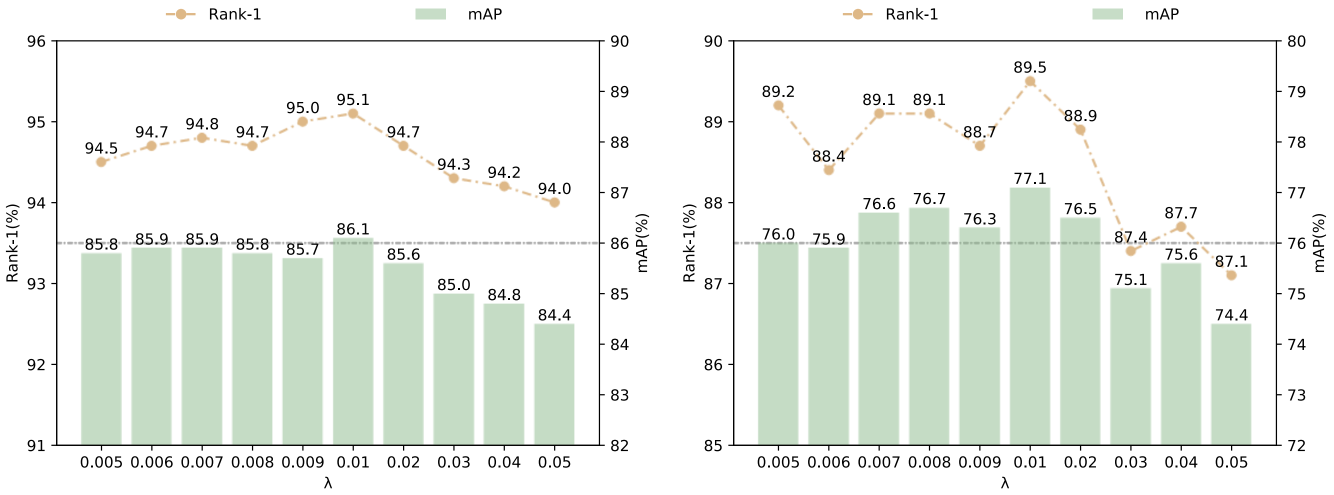


Figure 6: The top-10 results of person retrieval on Market-1501 dataset. Images in green boxes and red boxes are positive and negative results.((a): Original image; (b): Flipped image; (c): Scaled image; (d): Flipped+Scaled image)

Table 1: The ablation study of the proposed method on the Market-1501, DukeMTMC-reID, and MSMT17 datasets. The Rank-1 (R1) results and mAP accuracy are reported. Results are denoted as: B.+F. (original branch + flipping branch) and B.+S. (original branch + scaling branch). Best and second best results are denoted in red and blue respectively.

Method	Market-1501		DukeMTMC-reID		MSMT17	
	R1	mAP	R1	mAP	R1	mAP
Baseline	93.9	83.3	86.9	72.5	75.8	48.1
B.+ F.	94.3	84.2	87.0	73.9	78.9	51.7
B.+ S.	94.2	84.3	87.2	73.4	78.2	51.0
ESC(B. + S. + F.)	94.5	84.8	88.1	74.8	79.8	53.1
ESC + IW-loss	94.6	85.6	88.7	76.4	80.3	54.3
ESC + LW-loss	94.9	85.9	89.1	76.3	81.1	56.0
ESC+ IW-loss + LW-loss (VAC-Net)	95.1	86.1	89.5	77.1	81.3	56.3



(a) Market-1501

(b) DukeMTMC-reID

Figure 7: VAC-Net performance under different λ values. (Left is Market-1501 dataset and right is DukeMTMC-reID dataset)

effective approach in dealing with viewpoint and scale variations and thus can improve performance.

4.3.2 Quantitative Analysis. In Table 1, the model achieves obvious performance improvements on three public datasets whether we apply the B.+F., B.+S or B.+S.+F. (i.e., ESC) method, which demonstrates that it is effective to guide the network to learn a consistent embedding feature space among different transforms from the same images. Specifically, compared with the baseline model, our proposed ESC architecture obtains 0.6%, 1.2% and 4.0% gains in Rank-1 and 1.5%, 2.3% and 5.0% gains in mAP on the Market-1501, DukeMTMC-reID and MSMT17 datasets, respectively. Similarly, the ESC method also surpasses the B.+F. and B.+S., which illustrates that the better consistency of the embedding feature space among on three transforms is more effective than that with them separately. Besides that, we further enhance the consistent embedding feature space with two additional visual supervised losses. In particular, when only the IW-loss is added, the ESC method is improved by 0.7%, 1.8%, and 4.5% gains in the Rank-1 accuracy and 2.3%, 3.9%, and 6.2% gains in the mAP accuracy on the Market-1501, DukeMTMC-reID and MSMT17 datasets, respectively. And it illustrates that visual supervised loss is indicative of dealing with the viewpoint changes and scale variations. It can be noted that the results with LW-loss is superior to that with IW-loss especially on the MSMT17 dataset. The ESC architecture with LW-loss achieves 0.8%/1.7% improvements in Rank-1/mAP compared with IW-loss, which demonstrates that supervising visual attention consistency among different stages has more contribution to improve the model performance. Finally, the best performance is obtained when appending ESC+IW-loss+LW-loss (VAC-Net), which shows the superiority of our proposed method.

4.4 Parameter Analysis

As the shown in the Figure. 7, It can be seen that the model achieves the best performance when λ is set to 0.01. We also observe that the model accuracy increases with λ becoming larger, while it begins

Table 2: Comparison with the existing methods on Market-1501, DukeMTMC-reID, and MSMT17 dataset. The Rank (R1) and mAP accuracies are reported. It is clear that VAC-Net achieves excellent performance on all datasets, surpassing most methods by a clear margin. (Best and second best results in red and blue respectively. "–" indicates that the results are not available)

Method	Publication	Backbone	Market-1501		DukeMTMC-reID		MSMT17	
			R1	mAP	R1	mAP	R1	mAP
HA-CNN [17]	CVPR'18	HA-CNN	91.2	75.7	80.5	63.8	-	-
LGMANet [29]	ICIP'19	ResNet50	94.0	82.7	87.2	73.9	-	-
MHN [2]	ICCV'19	PCB	95.1	85.0	89.1	77.2	-	-
CE-SAN [47]	SPL'20	ResNet50	94.1	84.1	84.8	74.2	77.3	55.0
DPFL [3]	ICCV'17	Inception	88.9	73.1	79.2	60.6	-	-
MLFN [1]	CVPR'18	ResNeXt	90.0	74.3	81.0	62.8	-	-
CASN [41]	CVPR'19	ResNet50	94.4	82.8	87.7	73.7	-	-
RIN [35]	ICASSP'19	ResNet50	86.10	67.60	77.20	56.9	-	-
OSNet [46]	ICCV'19	OSNet	94.8	84.9	88.6	73.5	78.7	52.9
PCB+RPP [31]	ECCV'18	ResNet50	93.8	81.6	83.3	69.2	68.2	40.4
DGNet [42]	CVPR'19	ResNet50	94.8	86.0	86.6	74.8	77.2	52.3
IANet [11]	CVPR'19	ResNet50	94.4	83.1	87.1	73.4	75.5	46.8
RE [44]	AAAI'20	ResNet50	87.08	71.31	79.31	62.44	-	-
FFLN [39]	TMM'20	ResNet50	93.8	81.8	83.3	68.2	74.3	43.6
FA-Net [19]	TIP'21	ResNet50	95.0	84.6	88.7	77.0	76.8	51.0
CDNet [15]	CVPR'21	CDNet	95.1	86.0	88.6	76.8	78.9	54.7
VACNet	This work	ResNet50	95.1	86.1	89.5	77.1	81.3	56.3

to decrease when λ is greater than 0.01. Therefore, λ is a reasonable choice when it is 0.01.

4.5 Comparison with State-of-the-Art Methods

In this section, our method will be compared with state-of-the-art approaches including attention-based methods [2, 17, 29, 47], viewpoint invariance-based methods [35, 41], multi-scale based methods [1, 3, 46], and other methods [11, 15, 19, 31, 39, 42, 44]. As shown in Table 2, we achieve better performance than attention-based methods. Specifically, we obtain a higher mAP than the MHN [2] method by 1.1% on the Market-1501 dataset, and 1%/2%, 4.7%/2.9%, and 4%/1.3% improvement in R1/mAP than CE-SAN [47] on the three datasets. This indicates that VAC-Net also can learn effective attention region under the supervision of the IW-loss and LW-loss. Next, compared with viewpoint invariance-based [35, 41], our method achieves better results. Unlike [35] which learns consistent attention regions of all images from the same identity, we argue that the model should have consistent attention region for different transforms of the same image and a little difference between pairs of images from the same identity. Next, in comparison to the multi-scale feature extraction works [1, 3, 46], we obtain a clearly overwhelming performance increase. Specifically, compared with the OSNet, the R1/mAP are increaseed by 0.3%/1.2% on the Market-1501 dataset, 0.9%/3.6% on the DukeMTMC-reID dataset and 2.6%/3.4% on the MSMT17 dataset. We postulate the reason is that only the richness of multi-scale features is considered in OSNet and the consistency of cross-scale features. is not enough. Overall, we achieve the best performance on almost all datasets than other methods, which demonstrates the superiority of our approach.

5 CONCLUSION

In this paper, we design a simple yet effective Visual Attention Consistent Network (VAC-Net) to learn a consistent embedding feature space for person ReID to address viewpoint and scale variations. Firstly, the embedding space consistency (ESC) architecture is proposed to enable the network to learn consistent embedding feature space, thus accurately distinguish persons with different viewpoint changes and scale variants. What's more, we design two loss (i.e., IW-loss and LW-loss) to supervise the consistency of attention regions among the three branches and the consistency of semantic consistency among different stages in each branch, which further improve the embedding space consistency. VAC-Net illustrates its excellent performance through extensive experiments.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (No. 61973066, 61471110), Foundation Project of National Key Laboratory (6142002301), and Fundamental Research Funds for the Central Universities (N172608005, N182608004).

REFERENCES

[1] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. 2018. Multi-level factorisation net for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2109–2118.

[2] Binghui Chen, Weihong Deng, and Jiani Hu. 2019. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 371–381.

[3] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. 2017. Person re-identification by deep learning multi-scale representations. In *Proceedings of the IEEE international conference on computer vision workshops.* 2590–2600.

[4] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. 2016. Exploiting cyclic symmetry in convolutional neural networks. In *International conference on machine learning.* PMLR, 1889–1898.

[5] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. 2015. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition* 48, 10 (2015), 2993–3003.

[6] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. 2009. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence* 32, 9 (2009), 1627–1645.

[7] Yang Fu, Yunchao Wei, Yuqian Zhou, Honghui Shi, Gao Huang, Xinchao Wang, Zhiqiang Yao, and Thomas Huang. 2019. Horizontal pyramid matching for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 8295–8302.

[8] Douglas Gray, Shane Brennan, and Hai Tao. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*, Vol. 3. Citeseer, 1–7.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[10] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).

[11] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. 2019. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 9317–9326.

[12] Srikrishna Karanam, Yang Li, and Richard J Radke. 2015. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *Proceedings of the IEEE international conference on computer vision.* 4516–4524.

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Xu Lan, Xiatian Zhu, and Shaogang Gong. 2018. Person search by multi-scale matching. In *Proceedings of the European conference on computer vision (ECCV).* 536–552.

[15] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. 2021. Combined Depth Space based Architecture Search For Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 6729–6738.

[16] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 152–159.

[17] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2285–2294.

[18] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2197–2206.

[19] Yiheng Liu, Wengang Zhou, Jianzhuang Liu, Guo-Jun Qi, Qi Tian, and Houqiang Li. 2021. An end-to-end foreground-aware network for person re-identification. *IEEE Transactions on Image Processing* 30 (2021), 2060–2071.

[20] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.* 0–0.

[21] Diego Marcos, Michele Volpi, Nikos Komodakis, and Devis Tuia. 2017. Rotation equivariant vector field networks. In *Proceedings of the IEEE International Conference on Computer Vision.* 5048–5057.

[22] Hyunjong Park and Bumsub Ham. 2020. Relation network for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11839–11847.

[23] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Tao Xiang, and Xiangyang Xue. 2017. Multi-scale deep learning architectures for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision.* 5399–5408.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).

[25] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision.* Springer, 17–35.

[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 815–823.

[27] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision.* 618–626.

[28] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV).* 402–419.

[29] Lingchuan Sun, Jianlei Liu, Yingxin Zhu, and Zhuqing Jiang. 2019. Local to Global with Multi-Scale Attention Network for Person Re-Identification. In *2019 IEEE International Conference on Image Processing (ICIP).* IEEE, 2254–2258.

[30] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 6398–6407.

[31] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV).* 480–496.

[32] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 79–88.

[33] Fei Xiong, Mengran Gou, Octavia Camps, and Mario Sznaier. 2014. Person re-identification using kernel-based metric learning methods. In *European conference on computer vision.* Springer, 1–16.

[34] Mingfu Xiong, Jun Chen, Zhongyuan Wang, Chao Liang, Bohan Lei, and Ruimin Hu. 2017. A multi-scale triplet deep convolutional neural network for person re-identification. In *Pacific-Rim Symposium on Image and Video Technology.* Springer, 30–41.

[35] Dongshu Xu, Jun Chen, Chao Liang, Zheng Wang, and Ruimin Hu. 2019. Cross-view identical part area alignment for person re-identification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2462–2466.

[36] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. 2019. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1389–1398.

[37] Xun Yang, Meng Wang, and Dacheng Tao. 2017. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing* 27, 2 (2017), 791–805.

[38] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Deep metric learning for person re-identification. In *2014 22nd International Conference on Pattern Recognition.* IEEE, 34–39.

[39] Cairong Zhao, Xinbi Lv, Zhang Zhang, Wangmeng Zuo, Jun Wu, and Duoqian Miao. 2020. Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification. *IEEE Transactions on Multimedia* 22, 12 (2020), 3180–3195.

[40] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision.* 1116–1124.

[41] Meng Zheng, Srikrishna Karanam, Ziyan Wu, and Richard J Radke. 2019. Re-identification with consistent attentive siamese networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5735–5744.

[42] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. 2019. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2138–2147.

[43] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision.* 3754–3762.

[44] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13001–13008.

[45] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2921–2929.

[46] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2019. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 3702–3712.

[47] Xiaoguang Zhu, Jiuchao Qian, Haoyu Wang, and Peilin Liu. 2020. Curriculum enhanced supervised attention network for person re-identification. *IEEE Signal Processing Letters* 27 (2020), 1665–1669.