Self-distillation Enhanced Vertical Wavelet Spatial Attention for Person Re-identification

Yuxuan Zhang[1], Huibin Tan[1], Long Lan[1], Xiao Teng[1], Jing Ren[1(✉)], and Yongjun Zhang[2]

[1] Institute for Quantum Information & State Key Laboratory of High Performance Computing, College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China
{yuxuanzhang21,tanhb,long.lan,tengxiao14,renjing}@nudt.edu.cn
[2] Artificial Intelligence Research Center (AIRC), National Innovation Institute of Defense Technology (NIIDT), Beijing 100071, China
jyzhang@nudt.edu.cn

Abstract. Person re-identification is a challenging problem in computer vision, aiming to accurately match and recognize the same individual across different viewpoints and cameras. Due to significant variations in appearance under different scenes, person re-identification requires highly discriminative features. Wavelet features contain richer phase and amplitude information as well as rotational invariance, demonstrating good performance in various visual tasks. However, through our observations and validations, we have found that the vertical component within wavelet features exhibits stronger adaptability and discriminability in person re-identification. It better captures the body contour and detailed information of pedestrians, which is particularly helpful in distinguishing differences among individuals. Based on this observation, we propose a vertical wavelet spatial attention only with the vertical component in the high frequency specifically designed for feature extraction and matching in person re-identification. To enhance spatial semantic consistency and facilitate the transfer of knowledge between different layers of wavelet attention in the neural network, we introduce a self-distillation enhancement method to constrain shallow and deep spatial attention. Experimental results on Market-1501 and DukeMTMC-reID datasets validate the effectiveness of our model.

Keywords: wavelet spatial attention · person Re-ID · self-distillation

1 Introduction

Person re-identification, an important task within computer vision, aims to identify a specific individual in a video sequence or set of images captured by non-overlapping cameras. It plays a crucial role in scenarios such as tracking fugitives, locating missing persons, and tracing abducted individuals, particularly women and children.

Y. Zhang and H. Tan—Co-first authors of the article.

S. Rudinac et al. (Eds.): MMM 2024, LNCS 14555, pp. 1–13, 2024.
https://doi.org/10.1007/978-3-031-53308-2_1

Typically, person re-identification models utilize various features, including appearance, body shape, and clothing, for recognizing individuals. However, real-world scenarios introduce challenges such as blurred images, variations in shooting distance and angle, occlusions, and changes in pedestrian posture. Moreover, the inherent similarity of features among different individuals further complicates the task, making it more challenging than traditional target recognition. Consequently, achieving high discriminability in features becomes imperative for effective person re-identification.

The accurate extraction of valid pedestrian features becomes a key issue in improving the precision of person re-identification. Current mainstream research approaches have explored pedestrian features from several perspectives, including pixel features [1], local and global feature [2], and cross-domain invariant features [3], which have achieved good results. In a recent study, frequency features of images were introduced to the vision tasks [4]. The authors demonstrated that global average pooling is a special case of frequency domain compression and introduced discrete cosine transform (DCT) to extend signal compression towards richer multi-spectrum information, shifting our attention to the spatial frequencies of images. Meanwhile, [5] shown that low-frequency information, which represents semantic and labeling information, is a primary feature commonly used in computer vision tasks. On the other hand, high-frequency information contains valuable image texture and boundary details but also introduces noise that hampers model training. Traditional neural network models primarily utilize the low-frequency semantic information and are trained without explicitly considering the high-frequency information. However, as model accuracy improves, overfitting tendencies due to the dominance of high-frequency noise emerge. In summary, the frequency domain signal contains more information, but needs to be analyzed and selected.

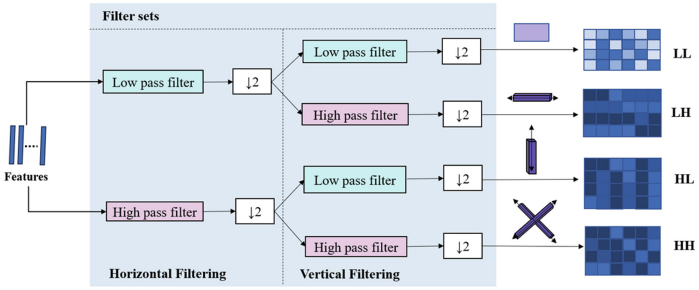

Fig. 1. Decomposition of images by 2D DWT. Low pass filter represents scale filter, high pass filter represents wavelet filter, and "↓2" means downsampling. They form the discrete wavelet transform.

The wavelet transform is a method of transforming a signal between the time and frequency domains. Wavelet functions of different scales and frequencies can be used to capture changes in the different frequency components of a signal. Large wavelet coefficients correspond to the more significant frequency components of the signal, while smaller wavelet coefficients correspond to the less significant frequency components. By selectively processing these wavelet coefficients, we can extract the key frequency

domain features in the signal and perform signal analysis, processing and feature extraction. Figure 1 shows the Haar wavelet as an example of how the decomposition of image spatial frequencies in different directions can be achieved using the two-dimensional discrete wavelet transform. Wavelet features have shown good performance in various vision tasks such as image classification [6], target detection [7] and image segmentation [8], and have also been used for tasks such as blur detection [9], denoising [10] and improving the mathematical interpretability of neural networks [11].

In person re-identification, we observe that pedestrians usually stand vertically, while roads and occlusions in the background are generally oriented horizontally. Therefore, the spatial features of pedestrians should be continuous in the vertical direction, and the segmentation line with the background is also essentially vertical. Based on the above assumptions, we extracted the vertical component of the wavelet features on the dataset. Experiments verified that the component can better capture the body contour and detail information of pedestrians, and has stronger adaptability and is more discriminative in the person re-identification task. Due to the low resolution of the person re-identification data, we select a pedestrian image from an open-source library for better displaying the effect of the wavelet transform. As shown in Fig. 2(a), the upper image is the horizontal high-frequency component contains in the original image, and the bottom image shows the vertical high-frequency component. The red box is enlarged to clearly show that the pedestrian profile information is disambiguated in the horizontal direction, while it is displayed more clearly in the vertical component.

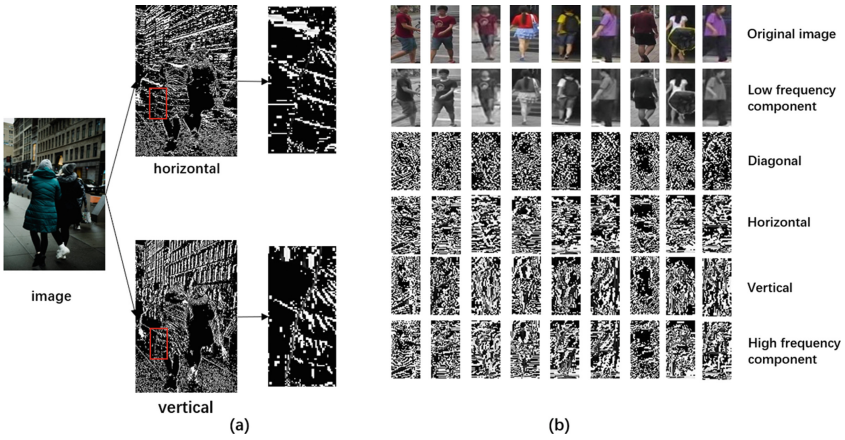

Fig. 2. Spatial structure preservation effect of images under wavelet transform.

Based on the aforementioned observations, this paper presents a vertical wavelet spatial attention module for feature extraction and matching in person re-identification, which uses the Haar wavelet transform to extract the vertical components of an image and constructs a vertical wavelet attention to obtain a useful high frequency. Experiments show that the model is able to better capture body contour and detail information

of pedestrians, which helps to distinguish differences between different pedestrians. Furthermore, to enhance the consistency of spatial semantics and facilitate knowledge transfer across different layers of the neural network, this paper introduces a self-distillation enhancement mechanism. This mechanism constrains the spatial attention in both shallow and deep layers by inserting a vertical wavelet attention module into each layer. These two modules generate separate spatial attention maps after extracting different network features, which are then interconnected through distillation loss functions. This enables bidirectional knowledge transfer within the model, facilitating improved prediction performance through self-learning.

Main contributions of this paper:

1. A vertical wavelet spatial attention module based on the Haar wavelet transform is proposed to selectively extract the vertical component from the high-frequency information of the image, so as to better capture the body contour and detail information of pedestrians and improve the discriminability of features.
2. A self-distillation enhancement mechanism is introduced to enhance the coherence of spatial attention in the shallow and deep layers and to induce the transfer of spatial attention knowledge in the neural network between different layers.
3. A self-distillation enhanced vertical wavelet spatial attention is applied to the person re-identification task and the effectiveness of the model is verified.

2 Related Work

2.1 Person Re-identification

A series of early works have been conducted around reducing noise interference and obtaining favorable features. Song *et al.* [12] designed a mask-guided contrastive attention model to filter the background directly using RGB image. Yu *et al.* [13] divides the features into a series of sub-features, with contextual information retaining. To considering global features, Jiao *et al.* [14] designs two additional attention: Hard region-level attention and Soft pixel-level attention, which can automatically locate the most active parts and eliminate background noise. Some more recent work has provided new insights. Li *et al.* [15] consider not only global and local features, but also the intrinsic relation of local features to retrieve lost information due to feature segmentation. Zhang *et al.* [16] considers that low-resolution images contain information that is not available in high and super-resolution images, and generates pedestrian features from a multi-resolution perspective. However, these studies are all carried out in the color-block dimension and have not been translated to the frequency domain.

2.2 Wavelet Transforms in CNN

The wavelet transform is an advancement of the Fourier transform and is widely used in the field of image processing. After the rise of neural networks, researchers found that large-scale models went under better than using wavelet algorithms alone, and began trying to combine wavelets and neural networks [17]. In general, wavelet transforms are more often used in the low-level tasks such as deraining [18] and deblurring [19].

Wave-CNet [7] replaces the maxpool with a wavelet transform and retains only the low-frequency part when downsampling to achieve Noise-Robust. Wave-kernel [8] replaces the first convolution layer directly with the wavelet transform. DWAN [20] points out that global average pooling has the disadvantage of insufficient channel information, so it switches to wavelet transform and introduces the Haar wavelet [21],which has the advantages of simple formulas, low computational effort, and easy implementation.

2.3 Knowledge Distillation Mechanism

Knowledge distillation is a method of model compression proposed by Hinton [22], there are teacher network and student network in the training stage. And the self-distillation is a deformation of knowledge distillation to train without teacher networks. Zhang *et al.* [23] add bottleneck and FC layers after each network block as a separate classifier, and calculate loss with the final layer of classification results. Xu *et al.* [24] pass data separately into the feature extraction network, and then use the output to calculate KL scatter, making the original and distorted images guide each other. Li *et al.* [25] split the batch into two networks, and get the target network by sharing the weights and averaging. Depending on the task, the exact structure of the self-distillation varies.

3 Method

In this section, we propose a novel approach for person re-identification called Self-distillation Enhanced Vertical Wavelet Spatial Attention. Our approach enhances the ResNet50 network by adding a vertical wavelet spatial attention module to transform the feature representation and filter noise in high-frequency components, which strengthens pedestrian features. Additionally, we employ a self-distillation enhancement mechanism to improve the network's performance. We present the method's architecture in detail, organized in a bottom-up order of network construction, which includes the framework in Sect. 3.1, vertical wavelet spatial attention module in Sect. 3.2, and the self-distillation enhancement mechanism in Sect. 3.3.

3.1 Overall Model Framework

Baseline Model Settings for Person Re-identification. In this study, we select article [26] as the baseline for our experiments and largely follow the proposed tricks in our replication experiments. We decide to use triplet loss instead of center loss since it is a more classical method, and based on the reported results in the article, center loss results in only 0.2 mAP increase in the model's performance.

Our model employs ResNet50 as the backbone network. The final downsampling parameter is changed to 1. This increases the spatial resolution of the output feature maps, helping to capture more detail of the pedestrian images. To further improve the performance of the model, we randomly erase parts of the input images. This encourages it to learn better representations of pedestrian images with occlusions and increases the robustness. We use a learning rate warm-up strategy during training to ensure that the learning rate is always within an appropriate range. When calculating the loss, we use the

sum of the triplet loss (Eq. (1)) and the cross-entropy loss of label smoothing (Eq. (2)) to prevent overfitting, where a and p are from the same category, a and n are from the different category. And d is calculating distance. The ε is a constant. The y indicates the true label and p_i is prediction logits. As the cross-entropy loss mainly optimizes the cosine distance, while the triplet loss concentrates on the Euclidean distance, we introduce a BN layer after the backbone network to explore the feature space more exhaustively.

$$Loss1 = \max(d(a,p) - d(a,n) + \varepsilon, 0) \tag{1}$$

$$Loss2 = \begin{cases} \sum_{i=1}^{N} -\left(1 - \frac{N-1}{N}\varepsilon\right)\log p_i, y = i \\ \sum_{i=1}^{N} -\frac{\varepsilon}{N}\log p_i, y \neq i \end{cases} \tag{2}$$

The Architecture of Self-distillation Enhanced Vertical Wavelet Spatial Attention for Person Re-identification. We present a novel approach for person re-identification using a 4-layers pipelined backbone network, as depicted in Fig. 3. Our proposed framework comprises of the original resnet50 network represented by the blue cube, a feedforward network containing the yellow and grey cubes, the vertical wavelet spatial attention module represented by the green parts and the self-distillation mechanism represented by the faint yellow part. Features from different stages are extracted to calculate different losses.

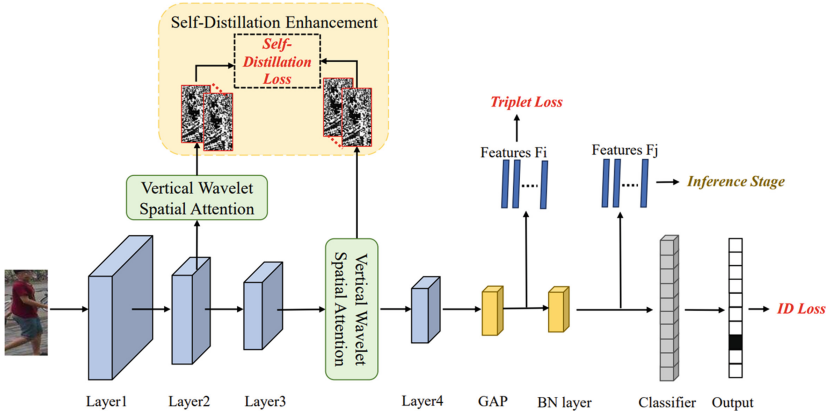

Fig. 3. The overall architecture of the method (Color figure online).

Our vertical wavelet spatial attention module can be encoded into any layer of the network based on the specific task requirements and experimental data. In this work, we put it between the third and fourth layers. By doing so, we can significantly enhance the model's performance in the person re-identification task.

To avoid increasing the parameter space of the backbone network, we employ an independent auxiliary network for self-distillation. Since the Haar wavelets are fully reconstructed, our wavelet spatial attention module does not change the feature size before and after decomposition and reconstruction, thus being plug and play and requiring only minor code changes for the module to shift between different layers.

3.2 Vertical Wavelet Spatial Attention Module

Inspired by wavelet spatial attention (WSA) used in image processing tasks, we introduce vertical wavelet spatial attention (VWSA) into the person re-identification model using the Haar wavelet transform.

Wavelet Spatial Attention in Person Re-identification. The WSA block comprises three main components, namely the wavelet decomposition, aggregation, and attention generation modules. Since Haar wavelet transform can be seen as the permutation and combination of the low-pass filter and high-pass filter, the two-dimensional discrete wavelet transform (2D DWT) is able to decompose features extracted from the network into four components: low-frequency approximation features (LL), horizontal features (LH), vertical features (HL), and diagonal features (HH) as illustrated in Fig. 1. Through summing up the high-frequency components and concatenating them with the low-frequency components, WSA produces aggregated features Fw. These aggregated features are then passed through the attention transformation network to determine the attention coefficients, forming the output wavelet attention, ATTw. Multiplying the attention and original feature results in the final output. The process is depicted in Fig. 4(a).

The WSA block functions as a structure-independent add-on module integrated into existing network architectures due to its modular design. It can be added to each layer of the network and ensure constant processing of the feature size. The effects of this module vary due to the differences in the sizes of the feature space in each layer.

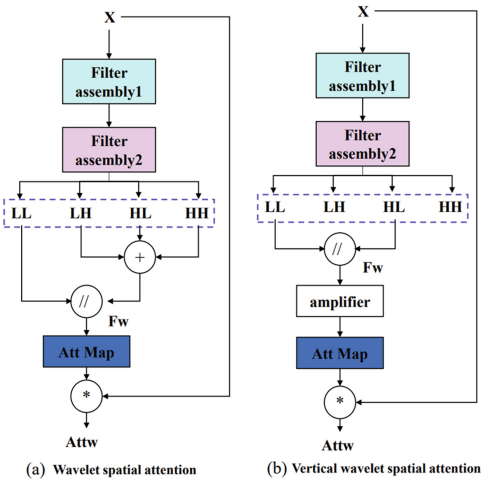

Fig. 4. Vertical Wavelet Spatial Attention Module. Fig(a) is the original wavelet spatial attention module, and Fig(b) is a vertical wavelet attention module. In this figure, "+" means add, "//" means splice and "*" means multiply.

Vertical High Frequency Extraction and Noise Filtering. For person re-identification tasks, we analyze the characteristics of the task. We believe that there are more continuous

contours in the vertical direction than the horizontal direction in pedestrian, and propose a method to extract contours by filtering high-frequency information from the vertical direction.

We utilize the wavelet transform to segment low frequency and different components of high-frequency information. We introduce the vertical wavelet spatial attention (VWSA) module, which integrates component features without affecting the feature size by discarding some high-frequency information. The preservation of the low-frequency component is crucial due to the significant amount of image approximation information it retains. Moreover, we directly splice the low-frequency component with the vertical component to maximize the retention of feature information, which enhances the pedestrian profile.

As shown in Fig. 4(b), the VWSA module extracts smaller wavelet feature values after discarding some components. Hence, we add signal amplifiers to obtain a more distinguishable attention map to avoid entering the inert region of the activation function.

3.3 Self-distillation Enhancement

We propose a self-distillation enhancement mechanism to enhance consistency in the ResNet50 network. Specifically, we aim to add the VWSA module after layer 2 and layer 3, and enable them to learn from each other. To achieve this goal, we extract the wavelet spatial attention after layer 2 and layer 3, and employ them to calculate an additional distillation loss function L3, such that the features converge and both can better focus on pedestrian features (Fig. 5).

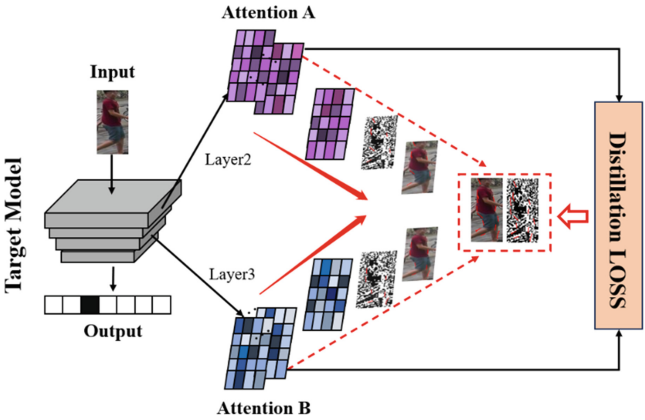

Fig. 5. Self-distillation enhancement mechanism. The Distillation Loss push two features from middle layers of the target network to achieve complementarity. The red arrow parts indicate that the attention A and attention B may focus on boundary at different locations and gradually complement each other under the constraint of the loss function.

To avoid increasing the size of the backbone network while improving its performance, a VWSA block is added as the additional auxiliary network after layer 2. The

added network structure is illustrated in the faint yellow area of Fig. 3. We measure the distillation loss using MSE Loss as shown in Eq. (3). The overall loss function is shown in Eq. (4), where att_i represents the spatial attention of different layers, and the α, β, and γ are hyperparameters.

$$Loss3 = \frac{1}{N}\sum_{k=1}^{N}(att_k^i - att_k^j)^2 \tag{3}$$

$$Loss = \alpha * Loss1 + \beta * Loss2 + \gamma * Loss3 \tag{4}$$

4 Experience

In this section, we present the experimental setup, datasets, and results of our proposed approach. We conducted experiments on the Market1501 and DukeMTMC-reID datasets, and evaluated the performance of our model using the Rank-1 and mAP metrics. Further, we analyzed the effectiveness of each modification by evaluating the performance of different stages of model adjustment and ablation study.

4.1 Datasets and Settings

Market1501 [27] and DukeMTMC-reID [28] are most widely used datasets for person re-identification. These two datasets have the similar organization structure of train, query and gallery sets. In contrast to the Market1501 dataset, the DukeMTMC-reID dataset has pedestrians that only appear in one camera as interference terms.

To ensure the preciseness of the experiment, we basically followed the same experimental setup as in the baseline [26]. For the loss function, we proposed Eq. (4). We keep the same part unchanged as the baseline, i.e. $\alpha = 1$ and $\beta = 1$.

4.2 Experience Results

To observe the effect of the discrete wavelet transform based on the Haar wavelet on the person re-identification dataset more intuitively, we conducted experiments on some images. As shown in Fig. 2(b), we selected several photos with different resolutions for the discrete wavelet transform, and the results show that the wavelet transform is able to better maintain the spatial structure in the images. The low-frequency component maintains the general information of the image well, but blurs textures and contours. The high frequency contains information in different directions and is comprehensive but cluttered. The information in the diagonal direction gives little indication of pedestrian features, and the horizontal frequency contains information in which the pedestrian contours are cut off and the pedestrian features are connected to the horizontal background. Pedestrian contours and details, such as leg information, can be better retained in the information contained in the vertical direction.

We make incremental improvements to the baseline to verify the effect of each step on the improvement of the model. The experiments' results are shown in Table 1. It shows

that the combination of the wavelet transform and the person re-identification task is able to achieve good improvements in both Rank-1 and mAP metrics. Based on it, the high-frequency component obtained by vertically high-pass filtering the wavelet attention is also able to obtain an improvement, which is more evident on mAP. The further addition of the self-distillation module is able to continue to push the mAP even higher while maintaining the Rank-1 level. The performance of the model firstly demonstrates that CNN can learn high-frequency information that is not visible to people. Secondly, it demonstrates that the person re-identification model based on wavelet attention can achieve better performance by filtering out noisy information from the high-frequency component. Finally, it also demonstrates the effectiveness of self-distillation architecture using attention between layers. As can be seen, our model is competitive among other advanced methods.

Table 1. Results of the model experiment on Market1501 and DukeMTMC-reID datasets.

Method		Datasets			
		Market1501		DukeMTMC-reID	
		Rank-1 (%)	mAP (%)	Rank-1 (%)	mAP (%)
PCB [2]	ECCV18	93.8	81.6	83.3	69.2
CAMA [29]	CVPR19	94.7	84.5	85.8	72.9
SNR [30]	CVPR20	94.4	84.7	85.9	73.0
GASM [31]	ECCV20	95.3	84.7	88.3	74.4
RANGEv2 [32]	PR22	94.7	86.8	87.0	78.2
Baseline [26]		94.1	85.7	86.2	75.9
+wavelet		94.4	86.4	86.8	76.9
+vertical wavelet		94.6	87.0	87.6	77.2
+vertical wavelet & self-distillation		95.2	87.4	87.9	77.4

4.3 Ablation Study

We propose a vertical wavelet spatial attention with a frequency selection mechanism, and in order to investigate the effectiveness of frequencies in each direction for the person re-identification, we conduct corresponding ablation experiments on Market1501. In addition, the wavelet attention module can be inserted after different layers of the model, and we also present the results of the ablation experiments for where the module is inserted here.

The Effect of Choosing Different Directions in High Frequency. Table 2 shows the effect of choosing to retain different frequencies for the person re-identification task. LL represents the low-frequency component, LH represents the horizontal high frequency,

Table 2. The effect of choosing different directions in high frequency.

Feature selection	Rank-1 (%)	mAP (%)
LL + LH	94.3	86.3
LL + HL	94.6	87.0
LL + HH	94.6	86.5
LL + LH + HL + HH	94.4	86.4

HL represents the vertical high frequency, and HH represents the diagonal high frequency. It indicates that the horizontal component cuts off the pedestrian profile feature and makes it discrete, while the diagonal direction contains compromising information and the vertical direction can best maintain the continuity of the pedestrian profile feature. This also verifies our hypothesis.

The Effect of Choosing Different Insertion Positions. We tried to encode the wavelet spatial attention module after different layers of the model and compare the final effect. The VW represents the wavelet attention module that completes the vertical frequency selection. The results are shown in Table 3. The results demonstrate that the wavelet attention module is movable and the effect varies with the size of the feature space size. This experiment result becomes the basis for constructing a self-distillation enhancement mechanism.

Table 3. The effect of choosing different insertion positions.

Insertion after	Rank-1 (%)	mAP (%)
Layer1	94.4	85.9
Layer2	94.5	85.9
Layer3	94.4	86.4
Layer4	93.9	86.0
Layer1(VW)	94.6	86.8
Layer2(VW)	94.9	86.9
Layer3(VW)	94.6	87.0
Layer4(VW)	94.3	86.2

5 Conclusion

Summarizing our work, in order to break through the limitations of the pixel level for the person re-identification, we introduce Haar wavelets into the CNN to build Haar wavelet attention in spatial; to make better use of the effective high-frequency components contained in the pictures for accurate knowledge, we select vertical components

that better match the upright pedestrian from the high frequency; to enhance learning and transfer knowledge in the CNN, we add a self-distillation enhancement mechanism to enable vertical wavelet spatial attention between different layers to learn from each other. Our proposed method, so far, offers an effective way to refine the attention maps of the ResNet50 network and improve its performance.

Acknowledgment. This work was supported by: The Self networks (CNNs). Furthermore, the Transformer has also-directed Project of State Key Laboratory of High Performance Computing: 202101-18.

References

1. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical Gaussian descriptor for person re-identification. In: 2016 IEEE CVPR, pp. 1363–1372 (2016)
2. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline) (2018). http://arxiv.org/abs/1711.09349
3. Liu, J., Zha, Z.-J., Chen, D., Hong, R., Wang, M.: Adaptive transfer network for cross-domain person re-identification. In: 2019 IEEE/CVF CVPR, pp. 7195–7204 (2019)
4. Qin, Z., Zhang, P., Wu, F., Li, X.: FcaNet: frequency channel attention networks. In: 2021 IEEE/CVF International Conference on Computer Vision, pp. 763–772 (2021)
5. Wang, H., Wu, X., Huang, Z., Xing, E.P.: High frequency component helps explain the generalization of convolutional neural networks, http://arxiv.org/abs/1905.13545 (2020)
6. Li, Q., Shen, L., Guo, S., Lai, Z.: Wavelet integrated CNNs for noise-robust image classification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7243–7252 (2020)
7. Alaba, S.Y., Ball, J.E.: WCNN3D: wavelet convolutional neural network-based 3D object detection for autonomous driving. Sensors **22**, 7010 (2022)
8. Li, Q., Shen, L.: WaveSNet: wavelet integrated deep networks for image segmentation. http://arxiv.org/abs/2005.14461 (2020)
9. Tong, H., Li, M., Zhang, H., Zhang, C.: Blur detection for digital images using wavelet transform. In: 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), pp. 17–20. IEEE, Taipei, Taiwan (2004)
10. Luo, X., Zhang, J., Hong, M., Qu, Y., Xie, Y., Li, C.: Deep wavelet network with domain adaptation for single image demoireing. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1687–1694 (2020)
11. Leterme, H., Polisano, K., Perrier, V., Alahari, K.: On the shift invariance of max pooling feature maps in convolutional neural networks. http://arxiv.org/abs/2209.11740 (2022)
12. Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1179–1188. IEEE, Salt Lake City, UT (2018)
13. Yu, R., Zhou, Z., Bai, S., Bai, X.: Divide and fuse: a re-ranking approach for person re-identification. In: Proceedings of the British Machine Vision Conference 2017, p. 135. British Machine Vision Association, London, UK (2017). https://doi.org/10.5244/C.31.135
14. Jiao, S., Wang, J., Hu, G., Pan, Z., Du, L., Zhang, J.: Joint attention mechanism for person re-identification. IEEE Access **7**, 90497–90506 (2019)
15. Li, W., et al.: Collaborative attention network for person re-identification. J. Phys. Conf. Ser. **1848**, 012074 (2021). https://doi.org/10.1088/1742-6596/1848/1/012074

16. Zhang, G., Chen, Y., Lin, W., Chandran, A., Jing, X.: Low resolution information also matters: learning multi-resolution representations for person re-identification. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, pp. 1295–1301, (2021)
17. Fujieda, S., Takayama, K., Hachisuka, T.: Wavelet convolutional neural networks, http://arxiv.org/abs/1805.08620 (2018)
18. Huang, H., Yu, A., Chai, Z., He, R., Tan, T.: Selective wavelet attention learning for single image deraining. Int. J. Comput. Vis. **129**, 1282–1300 (2021)
19. Zou, W., Jiang, M., Zhang, Y., Chen, L., Lu, Z., Wu, Y.: SDWNet: a straight dilated network with wavelet transformation for image deblurring. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), pp. 1895–1904 (2021)
20. Yang, Y., et al.: Dual wavelet attention networks for image classification. IEEE Trans. Circuits Syst. Video Technol. 1 (2022)
21. Mallat, S.G.: A Wavelet Tour of Signal Processing: The Sparse Way. Elsevier/Academic Press, Amsterdam, Boston (2009)
22. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network, http://arxiv.org/abs/1503.02531 (2015)
23. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: improve the performance of convolutional neural networks via self distillation, http://arxiv.org/abs/1905.08094 (2019)
24. Xu, T.-B., Liu, C.-L.: Data-distortion guided self-distillation for deep neural networks. AAAI **33**, 5565–5572 (2019). https://doi.org/10.1609/aaai.v33i01.33015565
25. Li, G., Togo, R., Ogawa, T., Haseyama, M.: Self-knowledge distillation based self-supervised learning for Covid-19 detection from chest X-ray images. In: ICASSP 2022, pp. 1371–1375 (2022)
26. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W.: Bag of tricks and a strong baseline for deep person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1487–1495. IEEE, Long Beach, CA, USA (2019)
27. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable Person re-identification: a benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 1116–1124. IEEE, Santiago, Chile (2015)
28. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by GAN improve the person re-identification baseline in vitro, http://arxiv.org/abs/1701.07717 (2017)
29. Yang, W., Huang, H., Zhang, Z., Chen, X., Huang, K., Zhang, S.: Towards rich feature discovery with class activation maps augmentation for person re-identification. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1389–1398. IEEE, Long Beach, CA, USA (2019)
30. Jin, X., Lan, C., Zeng, W., Chen, Z., Zhang, L.: Style normalization and restitution for generalizable person re-identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3140–3149. IEEE, Seattle, WA, USA (2020). https://doi.org/10.1109/CVPR42600.2020.00321
31. He, L., Liu, W.: Guided saliency feature learning for person re-identification in crowded scenes. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12373, pp. 357–373. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58604-1_22
32. Wu, G., Zhu, X., Gong, S.: Learning hybrid ranking representation for person re-identification. Pattern Recogn. **121**, 108239 (2022)