Advancing Person Re-identification with Frequency Domain Augmentation Using DHWT

Hongbo Jiang 31520241154483[1], Jiaqi Song 36920241153246[2], Zexu Wang 23220231151806[3]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University

[2]Institute of Artificial Intelligence, Xiamen University

[3]National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University

Abstract

Although many interesting and noteworthy works have emerged in pedestrian re-identification (ReID) tasks, most of them primarily impose loss constraints directly on RGB images and leverage large-scale pre-trained models for training. However, there has been very little research focusing on the evolution of models and recognition performance from the perspective of frequency dimensions. We believe that considering how to constrain high- and low-frequency information during the model evolution process is a worthwhile and intriguing problem. To validate our idea, in this paper, we utilize the Discrete Haar Wavelet Transform (DHWT) to decompose RGB image inputs into high-frequency and low-frequency components. We then observe the changes in these two types of information during the model evolution process. First, we propose a DHWT-based Frequency-Driven Augmentation (FDA) structure, which can be easily integrated into the model training pipeline. Second, to support frequency-adaptive enhancement, we introduce a Low-High Frequency Similarity Loss (LHFS loss) that constrains high- and low-frequency information, enabling the model to effectively distinguish between these two components.Experiments conducted on ReID benchmark datasets validate the effectiveness of our approach.

Introduction

Pedestrian re-identification (ReID) aims to find the most matching image from a dataset given a query image (Ye et al. 2021; Zhang et al. 2020a,b; Zhuang et al. 2020). Among these, some works have adopted ResNet as their backbone (Wang et al. 2020, 2018; Zhu et al. 2020), while TransReID has effectively improved the performance of transformers on ReID tasks (He et al. 2021).

Subsequent studies have not only utilized transformers but also fully explored their properties, leading to many fascinating developments (Tan et al. 2022b; Zhu et al. 2022a,b). However, it is worth noting that these efforts primarily focus on enhancing recognition performance by considering the model's evolution from the perspective of spatial information.

In this paper, we propose a novel approach by investigating the network evolution process from the frequency domain. As shown in Figure **1**, we compared the testing results of TransReID and ResNet101 on the MSMT17 and Market1501 datasets. Additionally, we used three different data types: the original images, the low-frequency components obtained by applying DHWT to the original images, and the high-frequency components.

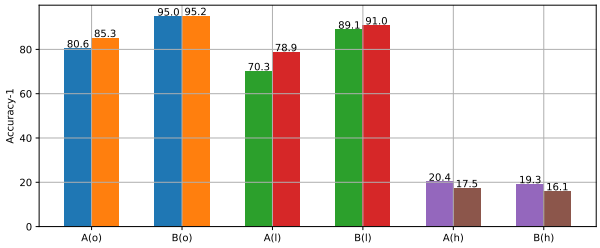

Figure 1: Comparison of ResNet101 and TransReID on Different Datasets. In this analysis, A represents ResNet101, and B represents TransReID. The notations (o), (l), and (h) correspond to the original images, the low-frequency components obtained by applying DHWT to the original images, and the high-frequency components, respectively. Accuracy-1 refers to the probability that the top-ranked result in the recognition process is the correct match (Rank-1 accuracy).

Models perform better with original images than with low- and high-frequency components. As shown in Figure 1, for the same model, the recognition performance on original images is consistently better than on low- and high-frequency components. Specifically, for ResNet101, the Rank-1 Accuracy achieved using original images is 10.3% higher than that of the low-frequency components and 60.2% higher than that of the high-frequency components. The primary reason for this, we believe, is that low-frequency components only capture coarse visual information, while the missing fine-grained details are closely associated with high-frequency information.

TransReID performs worse than ResNet101 in utilizing high-frequency information. As also shown in Figure **1**, for original images and low-frequency components, TransReID consistently outperforms ResNet101 in recognition performance. However, the opposite is true for high-frequency components. This observation, coupled with the findings in the first point, motivated our work to explore how the relationship between high- and low-frequency information evolves during the training of models with Transformer architectures. Our research reveals that the Transformer struc-

ture significantly reduces the distinction between high- and low-frequency information. In other words, Transformers lose the ability to differentiate between high- and low-frequency information during model evolution. This will be elaborated further in Section **Frequency Similarity**.

First, inspired by the observed changes in this relationship, we propose a DHWT-based Frequency Domain Augmentation (FDA) structure. This structure can be seamlessly integrated into the training pipeline and removed during the inference phase, ensuring that it does not affect the model's inference speed. Thus, FDA can be considered a plug-and-play auxiliary training component.

Second, to enable FDA to effectively constrain the relationship between high- and low-frequency information, we introduce a Low-High Frequency Similarity Loss (LHFS loss).

The contributions of our FDA can be summarized as follows:

- Our work uncovers the differences in how various models process original images, low-frequency information, and high-frequency information. Furthermore, it demonstrates that the relationship between high- and low-frequency information undergoes significant changes during network evolution.
- Based on the observed evolution patterns of high- and low-frequency information relationships, we propose a plug-and-play FDA structure and introduce the LHFS loss to constrain the relationship between these two types of information.
- Experiments conducted on the MSMT17 and Market-1501 datasets reveal that our FDA method can improve the model's recognition capability under the same conditions and achieve competitive recognition results.

Related work

CNN-based Person Re-identification

Recent advancements in person ReID have leveraged Convolutional Neural Networks (CNNs) to extract robust and discriminative features for identifying individuals across different camera views. Many CNN-based methods aim to enhance feature representations by combining global and local information.

For instance, Multiple Granularity Network (MGN) (Wang et al. 2018) utilizes a multi-branch architecture to capture both global and local features at different granularities, allowing for improved robustness in varying visual conditions. However, CNN-based methods often rely on predefined body part regions or supervised domain-specific knowledge, which can limit their generalization in real-world scenarios. To address these issues, Smoothing Adversarial Domain Attack (SADA) and p-Memory Reconsolidation (pMR) (Wang et al. 2020) propose techniques for cross-domain knowledge transfer by aligning feature distributions between labeled source and unlabeled target domains. Additionally, methods like Identity-guided Semantic Parsing (ISP) (Zhu et al. 2020) employ identity-guided learning for pixel-level alignment, enabling the model to effectively localize both human body parts and personal belongings, which are critical for ReID tasks.

Transformer-based Person Re-identification

Transformer-based models have gained significant attention in person ReID due to their ability to capture global context and long-range dependencies, overcoming limitations of CNNs, which focus on local receptive fields. These models process images as sequences of patches, enabling them to learn robust and holistic feature representations.

A key approach, TransReID (He et al. 2021), introduces a pure Transformer-based ReID framework. It encodes images into patch sequences and improves feature discrimination through two innovations: the Jigsaw Patch Module (JPM), which enhances feature diversity, and the Side Information Embeddings (SIE), which helps mitigate camera/view biases. This model achieves state-of-the-art results on several ReID benchmarks, demonstrating the power of Transformers in feature learning.

For fine-grained recognition, the Dual Cross-Attention Learning (DCAL) method (Zhu et al. 2022a) extends self-attention by introducing Global-Local Cross-Attention (GLCA) and Pair-Wise Cross-Attention (PWCA). GLCA enhances the interaction between global and local features, while PWCA regularizes attention learning to focus on relevant parts. This reduces misleading attention and improves the model's ability to differentiate subtle features in ReID tasks.

Transformer models also address occlusion challenges in ReID. The Dynamic Prototype Mask (DPM) (Tan et al. 2022b) introduces a Hierarchical Mask Generator to align occluded features and a Head Enrich Module to enhance feature aggregation, allowing for better handling of occluded or partial information.

In conclusion, Transformer-based methods provide significant advantages for person re-identification, including improved feature extraction, fine-grained recognition, and occlusion handling, achieving superior performance compared to traditional CNN-based methods.

Application of Frequency Information in Vision

Recent advancements in deep learning have explored the utility of frequency information, particularly through Fourier and wavelet transforms, to improve the performance of vision models. One prominent direction involves leveraging the spectral components of images to enhance model generalization and reduce the impact of domain shifts. For instance, some approaches have focused on using Fourier phase information, which is less susceptible to changes in domain distributions. A notable example is the Fourier-based domain generalization method, which incorporates a data augmentation strategy, known as amplitude mix, to improve generalization across different domains. By interpolating between the amplitude spectrums of source and target images, the method encourages the model to focus on invariant phase information, leading to state-of-the-art results in domain generalization tasks without requiring complex adversarial training techniques (Xu et al. 2021).

Wavelet transforms have also gained attention for their ability to capture both low-frequency and high-frequency components of images. The Deep Wavelet Super-Resolution (DWSR) method (Guo et al. 2017) exploits this by predicting missing high-frequency details in wavelet space, which significantly enhances the image resolution while maintaining computational efficiency. This approach not only simplifies the training process by reducing the need to learn low-frequency components but also demonstrates competitive performance compared to traditional super-resolution techniques. Similarly, the Wavelet Vision Transformer (Wave-ViT) introduces a novel method of down-sampling using wavelet transforms, which is both efficient and invertible. This enables a better trade-off between accuracy and computational cost, particularly by preserving high-frequency details, such as texture information, that are typically lost in traditional down-sampling methods used in Vision Transformers (Yao et al. 2022).

In addition, frequency information has been applied to unsupervised domain adaptation, where a simple Fourier Transform is used to align the low-frequency spectra of source and target domains, mitigating the domain shift without complex training procedures. This method has shown promising results in semantic segmentation tasks, particularly in scenarios where high-quality annotations are scarce in the target domain (Yang and Soatto 2020). Furthermore, the concept of feature space deep residual learning has been explored to enhance image restoration tasks, particularly in situations where CNNs struggle with images containing intricate patterns and structures. By applying an analytic mapping to a feature space, the method improves the learning process by focusing on spectral characteristics, demonstrating superior performance in tasks such as denoising and single-image super-resolution (Bae, Yoo, and Ye 2017).

These studies collectively highlight the importance of frequency-based techniques in improving both the robustness and efficiency of vision models, paving the way for more effective and scalable solutions in various vision tasks.

Proposed Solution

We first analyze the changes in the relationship between high- and low-frequency information during network evolution in Section **Frequency Similarity**. Then, in Section **Frequency Domain Augment**, we introduce the Frequency Domain Augment approach. Finally, in Section **Low-High Frequency Similarity Loss**, we use the LHFS loss to constrain the relationship between high- and low-frequency information during the model evolution process.

Frequency Similarity

Based on the preliminary experiments presented in the Introduction section, it is evident that both low-frequency and high-frequency information are equally important for the pedestrian re-identification task. To explore the role of these components in model evolution, we apply the Discrete Haar Wavelet Transform (DHWT) to process the input images or feature vectors, extracting both low-frequency and high-frequency information. For the input RGB image, img , we have:

$$low, high = ||DHWT(img)||_2 \tag{1}$$

where low and $high$ represent the low-frequency and high-frequency components, respectively.

After obtaining the high-frequency and low-frequency components, we calculate the cosine similarity between the high-frequency and low-frequency information.

$$cosine\ similarity = \frac{low \cdot high}{||low||_2||high||_2} \tag{2}$$

After obtaining the similarity between the two types of information, we compare the changes in the cosine similarity of high-frequency and low-frequency information before and after processing the input images by ResNet and Transformer network architectures on the MSMT17 dataset.

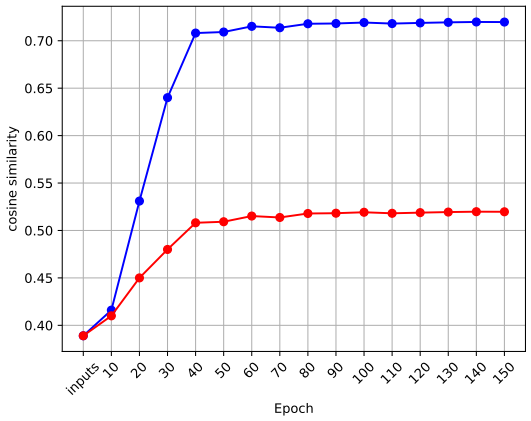

Figure 2: The curve illustrates the changes in high and low frequency information as they evolve through TransReID and ResNet101. The vertical axis represents the similarity between high and low frequency information after applying DHWT to the input images or feature vectors. The horizontal axis represents the iteration index. The red curve represents ResNet101, while the blue curve represents TransReID.

As shown in Figure **2**, we can observe that after the input image is processed by the transformer, the similarity between high-frequency and low-frequency information significantly increases. At the same time, we observe a similar trend in ResNet101. However, it is noteworthy that ResNet101 does not achieve a very high similarity between high and low-frequency information upon convergence. Therefore, we hypothesize that this phenomenon occurs because convolution has a better ability to distinguish between high and low-frequency information compared to the self-attention mechanism.

Frequency Domain Augment

We hope that the pedestrian re-identification model can effectively distinguish both high-frequency and low-frequency information in pedestrian images, thereby fully leveraging these two types of information to extract fine-grained details from the images.

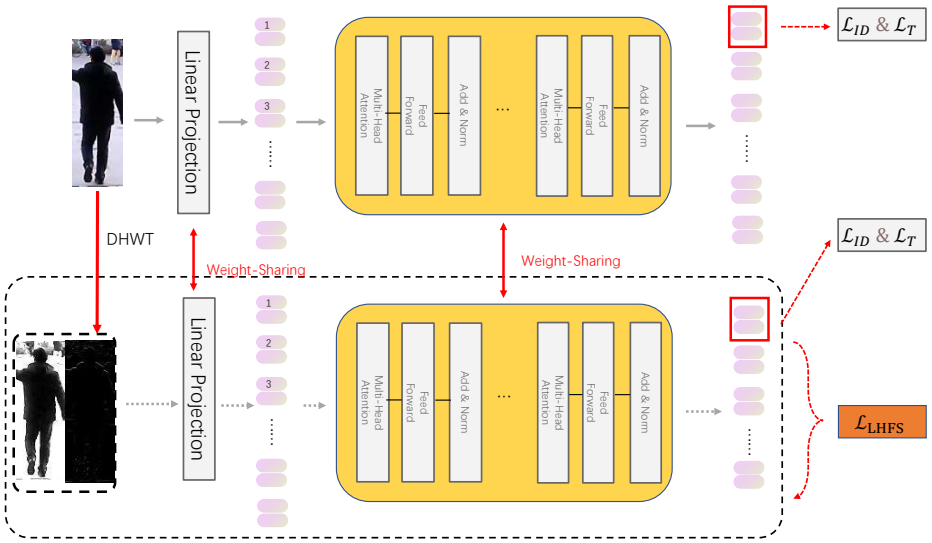

Figure 3: Overall Diagram of FDA. The diagram includes the plug-and-play FDA Module (FDAM) and the LHFS loss function. L_{ID} and L_T represent ID loss and Triplet loss, respectively. The part highlighted in red corresponds to the global embedding output by TransReID. The dashed box indicates the FDAM, which, along with the LHFS loss function, is only used during model training. FDA is a plug-and-play method designed to regulate the similarity of high and low-frequency information during the model's evolution, without introducing additional computational overhead during model inference.

To enable the model to effectively differentiate between high- and low-frequency information, we design a Frequency Domain Augmentation (FDA) approach. Figure **3** shows the overall framework during training. This augmentation method uses a plug-and-play module that allows the model to learn the ability to distinguish between high- and low-frequency information extracted via DHWT during training, rather than confusing the two types of information.

Low-High Frequency Similarity Loss

To prevent high-frequency and low-frequency information from being confused by the model during network evolution, we propose an innovative Low-High Frequency Similarity Loss (LHFS loss). This loss function is a similarity-based loss that helps ensure the model properly distinguishes between high-frequency and low-frequency components.

$$L_{LHFS} = \frac{\sum_{i=1}^{n} \text{low}_i \cdot \text{high}_i}{\sqrt{\sum_{i=1}^{n} \text{low}_i^2} \cdot \sqrt{\sum_{i=1}^{n} \text{high}_i^2} + \alpha} - 1 \quad (3)$$

where $^{\text{low}}{}_i$ and $^{\text{high}}{}_i$ represent the low-frequency and high-frequency components of the i-th image, respectively. The parameter α is a hyperparameter that controls the scale of the loss function. As can be seen, when the LHFS loss is low, the similarity between the high-frequency and low-frequency information in the output feature vectors is also low, which aligns with our training goal.

Experiments

Datasets and Evaluation Metrics

We conduct extensive experiments on three standard person ReID benchmarks: Market-1501 (Zheng et al. 2015), MSMT17 (Wei et al. 2018), CUHK03-NP (Li et al. 2014). Tabel **1** shows details of above datasets. Following conventions in the ReID community (He et al. 2021, 2020; Yan et al. 2020), we adopt Cumulative Matching Characteristic (CMC) curves and the mean Average Precision (mAP) to evaluate the quality of different methods.

Dataset	IDs	Images	Cameras
Market-1501	1,501	32,668	6
MSMT17	4,101	126,441	15
CUHK03-NP	1,467	13,164	2

Table 1: Details of the datasets used in our experiments.

Implementation Details

Following TransReID (He et al. 2021), all input images are resized to 256×128, and the training images are augmented with random horizontal flipping, padding, random cropping, and random erasing. The batch size is set to 64 with 4 images per ID, and the SGD optimizer is employed with a momentum of 0.9 and a weight decay of 0.0001. The learning rate is initialized as 0.008 with cosine learning rate decay. The parameter α in Eq. (**3**) is set to 0.01. All experiments are performed with one Nvidia A100 GPU with FP16 training.

Method	Market1501		MSMT17		CUHK03-NP labeled	
	R1(%)	mAP(%)	R1(%)	mAP(%)	R1(%)	mAP(%)
CNN-based methods						
CBDB-Net (TCSVT 21) (Tan et al. 2022a)	94.4	85.0	-	-	77.8	76.6
CDNet (CVPR 21) (Li, Wu, and Zheng 2021)	95.1	86.0	78.9	54.7	-	-
C2F (CVPR 21) (Zhang et al. 2021)	94.8	87.7	-	-	80.6	79.3
ViT-based methods						
TransReID (ICCV 21) (He et al. 2021)	95.2	89.9	85.3	67.4	81.7	79.6
PFD (AAAI 22) (Wang et al. 2021)	95.5	89.7	83.8	67.4	-	-
ABDNNet+NFormer (CVPR 22) (Wang et al. 2022)	95.7	**93.0**	80.8	62.2	80.6	79.1
DCAL (CVPR 22) (Zhu et al. 2022a)	94.7	87.5	83.1	64.0	-	-
FDA (ours)	**95.9**	90.2	**85.6**	**68.0**	**82.1**	**79.9**

Table 2: Comparison of Different Methods on Three Datasets.

Comparison with State-of-the-Art

We compare our proposed method with several state-of-the-art ReID methods, the results of which are shown in Table **2**. Our method achieves competitive performance on both the Market-1501, MSMT17 and CUHK03-NP labeled datasets, demonstrating the effectiveness of our approach. Particularly, with the TransReID baseline, our method achieves 95.9%/90.2%, 85.6%/68.0%, 82.1%/79.9% Rank-1/mAP on Market1501, MSMT17, CUHK03-NP labeled datasets, respectively.

Comparison to ViT-based Methods. Some typical works (e.g., TransReID (He et al. 2021), PFD (Wang et al. 2021) and DCAL (Zhu et al. 2022a)), extract discriminative part features for accurate alignment. Rather than aligning fine-grained parts, our FDA method benefits the ViT to preserve pivotal high-frequency components of images, to extract discriminative person representations. Compared to NFormer (Wang et al. 2022) which aggregates hierarchical features from CNN with Transformer blocks, our FDA method does not modify the model architecture. It is only necessary during training and can be discarded during inference, without bringing extra computation costs.

Comparison to CNN-based Methods. Compared with the competing method C2F (Zhang et al. 2021), our FDA outperforms it by 1.1%/2.5% and 1.5%/0.6% Rank-1/mAP on Market1501 and CUHK03-NP labeled datasets when taking the TransReID as the baseline. By virtue of our FDA, the ViT could not only build long-distance dependencies of low-frequency components but also capture key high-frequency components of person images. This benefits the ViT to extract discriminative person representations.

Ablation Study

To evaluate the effectiveness of our proposed FDA method, we conduct an ablation study on the MSMT17 dataset. As shown in Table **3**, we compare the performance of the ViT baseline with different combinations of FDA and LHFS loss. The results demonstrate that both FDA and LHFS loss contribute to the improvement in Rank-1 accuracy and mAP. Specifically, the combination of FDA and LHFS loss achieves the best performance, indicating that our proposed method effectively enhances the ViT's ability to extract discriminative person representations.

Index	FDA	L_{LHFS}	R1 (%)	mAP (%)
1			80.4	59.3
2	✓		82.3	67.4
3	✓	✓	85.6	68.0

Table 3: Ablation study over MSMT17 dataset.

Effectiveness of FDA. By comparing the results of Index 1 and Index 2, we observe that the FDA method improves the ViT's performance by 1.9% in Rank-1 accuracy and 8.1% in mAP. This demonstrates that FDA effectively enhances the ViT's ability to extract discriminative person representations by preserving high-frequency components of person images.

Effectiveness of LHFS Loss. By comparing the results of Index 2 and Index 3, we observe that the LHFS loss improves the ViT's performance by 3.3% in Rank-1 accuracy and 0.6% in mAP. This indicates that the LHFS loss effectively constrains the relationship between high- and low-frequency components, enabling the ViT to extract more discriminative person representations.

Conclusion

In this paper, we propose a novel approach to enhance person re-identification by considering the evolution of models from the perspective of frequency dimensions. We introduce a Frequency Domain Augmentation (FDA) method that leverages the Discrete Haar Wavelet Transform (DHWT) to decompose images into high-frequency and low-frequency components. We observe the changes in the relationship between high- and low-frequency information during network evolution and propose a Low-High Frequency Similarity Loss (LHFS loss) to constrain this relationship. Experi-

mental results on the MSMT17, Market-1501 and CUHK03-NP labeled datasets demonstrate the effectiveness of our approach, achieving competitive performance compared to state-of-the-art methods. Our work provides valuable insights into the role of frequency information in person re-identification and opens up new directions for future research in this area.

References

Bae, W.; Yoo, J.; and Ye, J. C. 2017. Beyond Deep Residual Learning for Image Restoration: Persistent Homology-Guided Manifold Simplification. arXiv:1611.06345.

Guo, T.; Mousavi, H. S.; Vu, T. H.; and Monga, V. 2017. Deep Wavelet Prediction for Image Super-Resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1100–1109.

He, L.; Liao, X.; Liu, W.; Liu, X.; Cheng, P.; and Mei, T. 2020. FastReID: A Pytorch Toolbox for General Instance Re-identification. arXiv:2006.02631.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. TransReID: Transformer-based Object Re-Identification. arXiv:2102.04378.

Li, H.; Wu, G.; and Zheng, W.-S. 2021. Combined Depth Space based Architecture Search For Person Re-identification. arXiv:2104.04163.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 152–159.

Tan, H.; Liu, X.; Bian, Y.; Wang, H.; and Yin, B. 2022a. Incomplete Descriptor Mining With Elastic Loss for Person Re-Identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1): 160–171.

Tan, L.; Dai, P.; Ji, R.; and Wu, Y. 2022b. Dynamic Prototype Mask for Occluded Person Re-Identification. arXiv:2207.09046.

Wang, G.; Lai, J.-H.; Liang, W.; and Wang, G. 2020. Smoothing Adversarial Domain Attack and P-Memory Reconsolidation for Cross-Domain Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning Discriminative Features with Multiple Granularities for Person Re-Identification. In *Proceedings of the 26th ACM international conference on Multimedia*, 274–282. ACM.

Wang, H.; Shen, J.; Liu, Y.; Gao, Y.; and Gavves, E. 2022. NFormer: Robust Person Re-identification with Neighbor Transformer. arXiv:2204.09331.

Wang, T.; Liu, H.; Song, P.; Guo, T.; and Shi, W. 2021. Pose-guided Feature Disentangling for Occluded Person Re-identification Based on Transformer. arXiv:2112.02466.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person Transfer GAN to Bridge Domain Gap for Person Re-identification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 79–88.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A Fourier-based Framework for Domain Generalization. arXiv:2105.11120.

Yan, C.; Pang, G.; Bai, X.; Zhou, J.; and Gu, L. 2020. Beyond Triplet Loss: Person Re-identification with Fine-grained Difference-aware Pairwise Loss. arXiv:2009.10295.

Yang, Y.; and Soatto, S. 2020. FDA: Fourier Domain Adaptation for Semantic Segmentation. arXiv:2004.05498.

Yao, T.; Pan, Y.; Li, Y.; Ngo, C.-W.; and Mei, T. 2022. Wave-ViT: Unifying Wavelet and Transformers for Visual Representation Learning. arXiv:2207.04978.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. H. 2021. Deep Learning for Person Re-identification: A Survey and Outlook. arXiv:2001.04193.

Zhang, A.; Gao, Y.; Niu, Y.; Liu, W.; and Zhou, Y. 2021. Coarse-to-Fine Person Re-Identification with Auxiliary-Domain Classification and Second-Order Information Bottleneck. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 598–608.

Zhang, T.; Xie, L.; Wei, L.; Zhang, Y.; Li, B.; and Tian, Q. 2020a. Single Camera Training for Person Re-Identification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07): 12878–12885.

Zhang, T.; Xie, L.; Wei, L.; Zhuang, Z.; Zhang, Y.; Li, B.; and Tian, Q. 2020b. UnrealPerson: An Adaptive Pipeline towards Costless Person Re-identification. arXiv:2012.04268.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable Person Re-identification: A Benchmark. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 1116–1124.

Zhu, H.; Ke, W.; Li, D.; Liu, J.; Tian, L.; and Shan, Y. 2022a. Dual Cross-Attention Learning for Fine-Grained Visual Categorization and Object Re-Identification. arXiv:2205.02151.

Zhu, K.; Guo, H.; Liu, Z.; Tang, M.; and Wang, J. 2020. Identity-Guided Human Semantic Parsing for Person Re-Identification. arXiv:2007.13467.

Zhu, K.; Guo, H.; Yan, T.; Zhu, Y.; Wang, J.; and Tang, M. 2022b. PASS: Part-Aware Self-Supervised Pre-Training for Person Re-Identification. arXiv:2203.03931.

Zhuang, Z.; Wei, L.; Xie, L.; Zhang, T.; Zhang, H.; Wu, H.; Ai, H.; and Tian, Q. 2020. Rethinking the Distribution Gap of Person Re-identification with Camera-based Batch Normalization. arXiv:2001.08680.