Deep Frequency Filtering for Domain Generalization

Shiqi Lin[1]* Zhizheng Zhang[2] Zhipeng Huang[1]* Yan Lu[2] Cuiling Lan[2] Peng Chu[2]
Quanzeng You[2] Jiang Wang[2] Zicheng Liu[2] Amey Parulkar[2] Viraj Navkal[2] Zhibo Chen[1]
[1]University of Science and Technology of China [2]Microsoft
{linsq047,hzp1104}@mail.ustc.edu.cn chenzhibo@ustc.edu.cn
{zhizzhang,yanlu,culan,pengchu,quyou,jiangwang,zliu,amey.parulkar,vnavkal}@microsoft.com

Abstract

Improving the generalization ability of Deep Neural Networks (DNNs) is critical for their practical uses, which has been a longstanding challenge. Some theoretical studies have uncovered that DNNs have preferences for some frequency components in the learning process and indicated that this may affect the robustness of learned features. In this paper, we propose Deep Frequency Filtering (DFF) for learning domain-generalizable features, which is the first endeavour to explicitly modulate the frequency components of different transfer difficulties across domains in the latent space during training. To achieve this, we perform Fast Fourier Transform (FFT) for the feature maps at different layers, then adopt a light-weight module to learn attention masks from the frequency representations after FFT to enhance transferable components while suppressing the components not conducive to generalization. Further, we empirically compare the effectiveness of adopting different types of attention designs for implementing DFF. Extensive experiments demonstrate the effectiveness of our proposed DFF and show that applying our DFF on a plain baseline outperforms the state-of-the-art methods on different domain generalization tasks, including close-set classification and open-set retrieval.

1. Introduction

Domain Generalization (DG) seeks to break through the *i.i.d.* assumption that training and testing data are identically and independently distributed. This assumption does not always hold in reality since domain gaps are commonly seen between the training and testing data. However, collecting enough training data from all possible domains is costly and even impossible in some practical environments. Thus, learning generalizable feature representations is of high practical value for both industry and academia.

Recently, a series of research works [78] analyze deep learning from the frequency perspective. These works, represented by the F-Principle [75], uncover that there are different preference degrees of DNNs for the information of different frequencies in their learning processes. Specifically, DNNs optimized with stochastic gradient-based methods tend to capture low-frequency components of the training data with a higher priority [74] while exploiting high-frequency components to trade the robustness (on unseen domains) for the accuracy (on seen domains) [66]. This observation indicates that different frequency components are of different transferability across domains.

In this work, we seek to learn generalizable features from a frequency perspective. To achieve this, we conceptualize Deep Frequency Filtering (DFF), which is a new technique capable of enhancing the transferable frequency components and suppressing the ones not conducive to generalization in the latent space. With DFF, the frequency components of different cross-domain transferability are dynamically modulated in an end-to-end manner during training. This is conceptually simple, easy to implement, yet remarkably effective. In particular, for a given intermediate feature, we apply Fast Fourier Transform (FFT) along its spatial dimensions to obtain the corresponding frequency representations where different spatial locations correspond to different frequency components. In such a frequency domain, we are allowed to learn a spatial attention map and multiply it with the frequency representations to filter out the components adverse to the generalization across domains.

The attention map above is learned in an end-to-end manner using a lightweight module, which is instance-adaptive. As indicated in [66, 74], low-frequency components are relatively easier to be generalized than high-frequency ones while high-frequency components are commonly exploited to trade robustness for accuracy. Although this phenomenon can be observed consistently over different instances, it does not mean that high-frequency com-

*This work was done when Shiqi Lin and Zhipeng Huang were interns at Microsoft Research Asia.

ponents have the same proportion in different samples or have the same degree of effects on the generalization ability. Thus, we experimentally compare the effectiveness of task-wise filtering with that of instance-adaptive filtering. Here, the task-wise filtering uses a shared mask over all instances while the instance-adaptive filtering uses unshared masks. We find the former one also works but is inferior to our proposed design by a clear margin. As analyzed in [10], the spectral transform theory [32] shows that updating a single value in the frequency domain globally affects all original data before FFT, rendering frequency representation as a global feature complementary to the local features learned through regular convolutions. Thus, a two-branch architecture named Fast Fourier Convolution (FFC) is introduced in [32] to exploit the complementarity of features in the frequency and original domains with an efficient ensemble. To evaluate the effectiveness of our proposed DFF, we choose this two-branch architecture as a base architecture and apply our proposed frequency filtering mechanism to its spectral transform branch. Note that FFC provides an effective implementation for frequency-space convolution while we introduce a novel frequency-space attention mechanism. We evaluate and demonstrate our effectiveness on top of it.

Our contributions can be summarized in the following:

- We discover that the cross-domain generalization ability of DNNs can be significantly enhanced by a simple learnable filtering operation in the frequency domain.
- We propose an effective Deep Frequency Filtering (DFF) module where we learn an instance-adaptive spatial mask to dynamically modulate different frequency components during training for learning generalizable features.
- We conduct an empirical study for the comparison of different design choices on implementing DFF, and find that the instance-level adaptability is required when learning frequency-space filtering for domain generalization.

2. Related Work

2.1. Domain Generalization

Domain Generalization (DG) aims to improve the generalization ability of DNNs from source domains to unseen domains, which is widely needed in different application scenarios. The challenges of DG have been addressed from *data*, *model*, and *optimization* perspectives. From the *data* perspective, augmentation [26, 64, 79] and generation [61,65,87] technologies are devised to increase the diversity of training samples so as to facilitate generalization. From the *model* perspective, some efforts are made to enhance the generalization ability by carefully devising the normalization operations in DNNs [44, 53, 60] or adopting an ensemble of multiple expert models [47, 88]. From the *optimization* perspective, there are many works designing different training strategies to learn generalizable features. which is a dominant line in this field. To name a few, some works learn domain-invariant feature representations through explicit feature alignment [21, 31, 51], adversarial learning [20, 22, 40], gradient-based methods [3, 33, 41], causality-based methods [39] or meta-learning based method [69], *etc*. In this work, we showcase a conceptually simple operation, *i.e*., learnable filtration in the frequency domain, can significantly strengthen the generalization performance on unseen domains, verified on both the close-set classification and open-set retrieval tasks.

2.2. Frequency Domain Learning

Frequency analysis has been widely used in conventional digital image processing for decades [4, 55]. Recently, frequency-based operations, *e.g*., Fourier transform, set forth to be incorporated into deep learning methods for different purposes in four aspects: 1) accelerating the training or facilitating the optimization of DNNs [12, 36, 49, 52, 56, 57]; 2) achieving effective data augmentation [28, 45, 73, 76]; 3) learning informative representations of non-local receptive fields [10,48,59,63,77]; 4) helping analyze and understand some behaviors of DNNs [66,74,75,78] as a tool. As introduced before, prior theoretical studies from the frequency perspective uncover that different frequency components are endowed with different priorities during training and contribute differently to the feature robustness. This inspires us to enhance the generalization ability of DNNs through modulating different frequency components.

In [10], for intermediate features, a 1×1 convolution in the frequency domain after FFT to learn global representations. However, such global representations capture global characteristics while losing local ones, thus have been demonstrated complementary with the features learned in the original latent space. To address this, a two-branch architecture is proposed in [10] to fuse these two kinds of features. This problem also exists in our work but is not our focus. Thereby, we adopt our proposed frequency filtering operation in the spectral transform branch of the two-branch architecture proposed in [10] for effectiveness evaluation. Besides, in [59], a learnable filter layer is adopted to self-attention (*i.e*., transformer) to mix tokens representing different spatial locations, which may seem similar with ours at its first glance but is actually not. The learnable filter in [59] is implemented by network parameters while that of ours is the network output thus instance-adaptive. We theoretically analyze and experimentally compare them in the following sections. Besides, with a different purpose from token mixing, we are devoted to improve the generalization ability of DNNs.

2.3. Attention Mechanisms

Attention has achieved great success in many visual tasks. It can be roughly categorized into selective attention [5, 27, 30, 58, 67, 71, 81, 82, 89] and self-attention [1, 6, 7, 17, 19, 30, 54, 68] upon their working mechanisms. The former one explicitly learns a mask to enhance task-beneficial features and suppress task-unrelated features. In contrast, self-attention methods commonly take affinities of tokens as the attentions weights to refine the token representations via message passing, wherein the attention weights can be understood to model the importance of other tokens for the query token. Our proposed frequency filtering is implemented with a simple selective attention applied in the frequency domain for the intermediate features of DNNs. There have been a few primary attempts [58, 89] exploiting frequency representations to learn more effective attention. In these works [58, 89], channel attention weights are proposed to learned from multiple frequency components of 2D DCT, where they are still used to modulate channels in the original feature space. In our work, we further investigate the frequency filtering where the the learning and using of attention weights are both in the frequency domain. We make the first endeavour to showcase the effectiveness of such a conceptually simple mechanism for the DG field, and would leave more delicate designs of attention model architectures for the frequency domain in our future work.

3. Deep Frequency Filtering

3.1. Problem Definition and Core Idea

In this paper, we aim to reveal that the generalization ability of DNNs to unseen domains can be significantly enhanced through an extremely simple mechanism, *i.e.*, an explicit frequency modulation in the latent space, named Deep Frequency Filtering (DFF). To shed light on this core idea, we first introduce the problem definition of Domain Generalization (DG) as preliminaries. Given K source domains $\mathcal{D}_s = \{D_s^1, D_s^2, \cdots, D_s^K\}$, where $D_s^k = (\mathbf{x}_i^k, y_i^k)_{i=1}^{N_k}$ denotes the k-th domain consisting of N_k samples $\mathbf{x}_i^k$ with their corresponding labels y_i^k, the goal of DG is to enable the model trained on source domains $\mathcal{D}_s$ perform as well as possible on unseen target domains $\mathcal{D}_t$, without additional model updating using the data in $\mathcal{D}_t$. When different domains share the same label space, it corresponds to a closed-set DG problem, otherwise an open-set problem.

As introduced before, the studies for the behaviors of DNNs from the frequency perspective [66, 74, 75] have uncovered that the DNNs have different preferences for different frequency components of the learned intermediate features. The frequency characteristics affect the trade-off between robustness and accuracy [66]. This inspires us to improve the generalization ability of DNNs through modulating different frequency components of different transfer difficulties across domains during training. Achieved by a simple filtering operation, transferable frequency components are enhanced while the components prejudice to cross-domain generalization are suppressed.

3.2. Latent Frequency Representations

Different from previous frequency-based methods [45, 73, 76] applied in the pixel space (*i.e.*, the side of inputs), we adopt our proposed filtering operation in the latent space. In this section, we briefly recall a conventional signal processing tool Fast Fourier Transform (FFT). We adopt it for obtaining the feature representations in the frequency domain, then discuss the characteristics of such representations.

Given the intermediate features $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, we perform a 2D fast Fourier transform (*i.e.*, an accelerated version [14] of 2D discrete Fourier transform) for each channel independently to get the corresponding frequency representations $\mathbf{X}_F \in \mathbb{R}^{2C\times H\times(\lfloor \frac{W}{2}\rfloor+1)}$. We formulate this transform $\mathbf{X}_F$=$FFT(\mathbf{X})$ as below (where the channel dimension is omitted for brevity):

$$\mathbf{X}_F(x,y) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} \mathbf{X}(h,w)e^{-j2\pi\left(x\frac{h}{H}+y\frac{w}{W}\right)}. \tag{1}$$

The frequency representation $\mathbf{X}_F$ can be transferred to the original feature space via an inverse FFT, succinctly expressed as $\mathbf{X}$=$iFFT(\mathbf{X}_F)$, which can be formulated as:

$$\mathbf{X}(h,w) = \frac{1}{H\cdot W}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} \mathbf{X}_F(x,y)e^{j2\pi\left(x\frac{h}{H}+y\frac{w}{W}\right)}. \tag{2}$$

The $\mathbf{X}_F \in \mathbb{R}^{2C\times H\times(\lfloor \frac{W}{2}\rfloor+1)}$ above denotes the frequency representation of $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, which concatenates the real and imaginary parts after FFT (each one has C channels). Besides, thanks to the conjugate symmetric property of FFT, $\mathbf{X}_F$ only needs retain the half of spatial dimensions thus has a spatial resolution of $H \times (\lfloor \frac{W}{2}\rfloor + 1)$. For the frequency representation $\mathbf{X}_F$, there are two utilizable properties: 1) Different frequency components of the original feature $\mathbf{X}$ are decomposed into elements at different spatial locations of $\mathbf{X}_F$, which could be viewed as a frequency-based disentanglement and re-arrangement for $\mathbf{X}$. This property makes the learning in frequency domain efficient in practice, and more importantly, allows us to achieve frequency filtering with a simple devised spatial attention module. 2) $\mathbf{X}_F$ is a naturally global feature representation, as discussed in [10], which can facilitate the suppression of globally distributed domain-specific information, such as illumination, imaging noises, *etc*. Next, we shed light on the specific filtering operation on $\mathbf{X}_F$.

3.3. Latent-space Frequency Filtering

Our goal is to adaptively modulate different frequency components over different network depths during training.

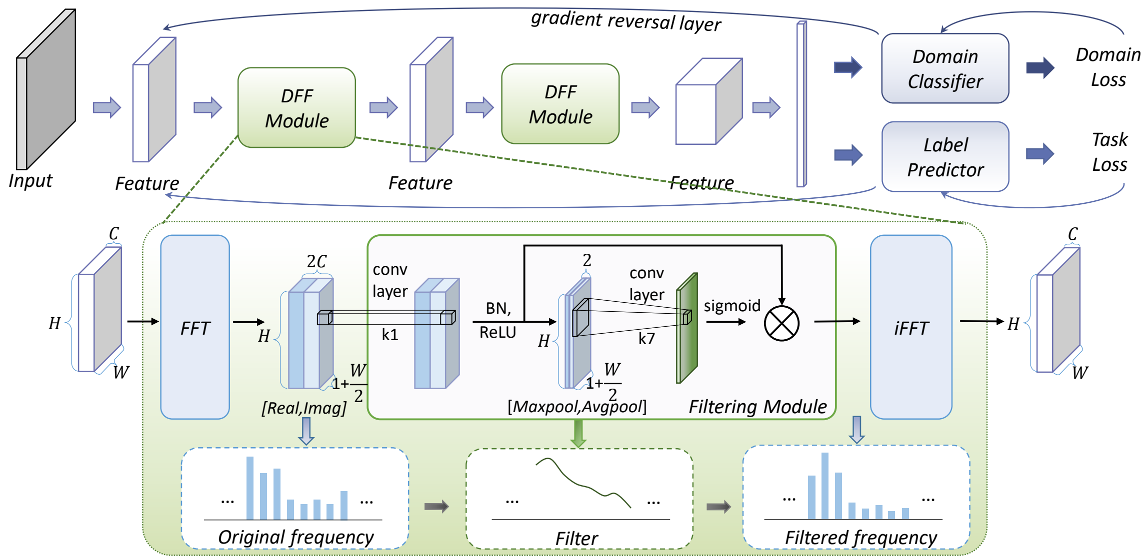

Figure 1. Illustration of our proposed Deep Frequency Filtering (DFF) module. DFF learns an instance-adaptive spatial mask to dynamically modulate different frequency components during training for learning generalizable features, which consists of three key operations: a 2D fast Fourier transform (FFT) to convert the input spatial features to the frequency domain, an filtering module to enhance the transferable components while suppressing the generalization-detrimental ones, and a 2D inverse FFT (iFFT) to map the features back to the orginal latent space.

We thus propose to apply a frequency filtering operation on $\mathbf{X}_F$ to enhance the transferable components while suppressing the generalization-detrimental ones. Thanks to the first hallmark of $\mathbf{X}_F$ discussed in Sec. 3.2, the frequency filtering operation is allowed to be implemented with a spatial attention on $\mathbf{X}_F$. Given a frequency representation $\mathbf{X}_F \in \mathbb{R}^{2C\times H\times(\lfloor\frac{W}{2}\rfloor+1)}$, the proposed frequency filtering mechanism is formulated as follows:

$$\mathbf{X}'_F = \mathbf{X}_F \otimes \mathbf{M}_S(\mathbf{X}_F), \tag{3}$$

where $\otimes$ denotes element-wise multiplication. $\mathbf{M}_S(\cdot)$ refers to the attention module to learn a spatial mask with a resolution of $H \times (\lfloor\frac{W}{2}\rfloor + 1)$. This mask is copied along the channel dimension of $\mathbf{X}_F$ accordingly for the element-wise multiplication, filtering out the components adverse to the generalization in $\mathbf{X}_F$. The frequency feature after filtering is denoted by $\mathbf{X}'_F$. Our contributions lie in revealing such a frequency filtering operation in the latent space can bring impressive improvements for DG, although using a lightweight attention architecture designed for the features in original latent space [71] to implement $\mathbf{M}_S(\cdot)$. This provides another alternative to the field of DG, which is conceptually simple, significantly effective, but previously unexplored. Besides, we further conduct an empirical study to investigate the effectiveness of different attention types in implementing our conceptualized deep frequency filtering in the experiment section. The specific architecture design for the attention module is not our current focus, but is worth being explored in the future.

Here, we introduce an extremely simple instantiation for Eq. (3). We use this for verifying our proposed concept *deep frequency filtering* in this paper. For $\mathbf{X}_F \in \mathbb{R}^{2C\times H\times(\lfloor\frac{W}{2}\rfloor+1)}$ that consists of real and imaginary parts after FFT, inspired by the attention architecture design in [71], we first adopt a 1×1 convolutional layer followed by Batch Normalization (BN) and ReLU activation to project $\mathbf{X}_F$ to an embedding space for the subsequent filtration. After embedding, as shown in Fig. 1, we follow the spatial attention architecture design in [71] to aggregate the information of $\mathbf{X}_F$ over channels using both average-pooling and max-pooling operations along the channel axis, generating two frequency descriptors denoted by $\mathbf{X}_F^{avg}$ and $\mathbf{X}_F^{max}$, respectively. These two descriptors can be viewed as two compact representations of $\mathbf{X}_F$ in which the information of each frequency component is compressed separately by the pooling operations while the spatial discriminability is still preserved. We then concatenate $\mathbf{X}_F^{avg}$ with $\mathbf{X}_F^{max}$ and use a large-kernel 7×7 convolution layer followed by a sigmoid function to learn the spatial mask. Mathematically, this instantiation can be formulated as:

$$\mathbf{X}'_F = \mathbf{X}_F \otimes \sigma(\mathrm{Conv}_{7\times 7}([\mathrm{AvgPool}(\mathbf{X}_F), \mathrm{MaxPool}(\mathbf{X}_F)])), \tag{4}$$

where σ denotes the sigmoid function. The $[\cdot,\cdot]$ is a concatenation operation. $\mathrm{AvgPool}(\cdot)$ and $\mathrm{MaxPool}(\cdot)$ denote the average and max pooling operations, respectively. $\mathrm{Conv}_{7\times 7}(\cdot)$ is a convolution layer with the kernel size of 7. Albeit using a large-size kernel, the feature $[\mathrm{AvgPool}(\mathbf{X}_F), \mathrm{MaxPool}(\mathbf{X}_F)]$ has only two channels through the information squeeze by pooling operations such that this step is still very computationally efficient in practice. We omit the embedding of $\mathbf{X}_F$ in this formulation for brevity. We believe using more complex attention architectures, such as [16, 50, 82], is of the potentials to achieve higher improvements, and we expect more effective instantiations of our conceptualized Deep Frequency Filtering.

Discussion. The proposed Deep Frequency Filtering is conceptually new design to achieve instance-adaptive frequency modulation in the latent space of DNNs. It also corresponds to a novel neural operation albeit using an off-the-shelf architecture design as an exampled instantiation. Compared to prior frequency-domain works [10, 59], we make the first endeavour to introduce an explicit instance-adaptive frequency selection mechanism into the optimization of DNNs. From the perspective of attention, conventional attention designs [27, 68, 71, 82] learn masks from deep features in the original latent space, and adopt the learned masks to these features themselves to achieve feature modulation. FcaNet [58] strives to a further step by learning channel attention weights from the results of frequency transform. But the learned attention weights are still used for the original features. In this aspect, we are the first to learn attention weights from frequency representations and also use the learned masks in the frequency domain to achieve our conceptualized frequency filtering.

3.4. Post-filtering Feature Restitution

The features captured in the frequency domain have been demonstrated to be global and complementary to the local ones captured in the original latent space in [10]. Thus, a simple two-branch is designed to exploit this complementarity to achieve an ensemble of both local and global features in [10]. This architecture is naturally applicable to the restitution of complementary local features as a post-filtering refinement in the context of our proposed concept. We thus evaluate the effectiveness of our proposed method on top of the two-branch architecture in [10]. Specifically, similar to [10], we split the given intermediate feature $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ along its channel dimension into $\mathbf{X}^g \in \mathbb{R}^{rC \times H \times W}$ and $\mathbf{X}^l \in \mathbb{R}^{(1-r)C \times H \times W}$. The two-branch architecture can be formulated as:

$$Y^l = f_l(\mathbf{X}^l) + f_{g \to l}(\mathbf{X}^g), \quad Y^g = f_g(\mathbf{X}^g) + f_{l \to g}(\mathbf{X}^l), \tag{5}$$

where $f_l(\cdot)$, $f_g(\cdot)$, $f_{l \to g}(\cdot)$ and $f_{g \to l}(\cdot)$ denote four different transformation functions. Among them, $f_l(\cdot)$, $f_{l \to g}(\cdot)$ and $f_{g \to l}(\cdot)$ are three regular convolution layers. In [10], the $f_g(\cdot)$ corresponds to the spectral transform implemented by their proposed convolution operation in the frequency domain. On top of it, we evaluate the effectiveness of our proposed DFF by adding this operation into the spectral transform branch of this architecture to achieve an explicit filtering operation in the frequency domain. The contribution on this two-branch architecture design belongs to [10].

3.5. Model Training

In addition to commonly used task-related loss functions (for classification or retrieval), we train a domain classifier with a domain classification loss and adopt a gradient reversal layer [20]. These are commonly used in DG research works for explicitly encouraging the learning of domain-invariant features and the suppression for features conducive to domain generalization. The feature extractor is optimized for minimizing the task losses while maximizing the domain classification loss simultaneously.

4. Experiments

4.1. Datasets and Settings

We evaluate the effectiveness of our proposed Deep Frequency Filtering (DFF) for Domain Generalization (DG) on **Task-1**: the close-set classification task and **Task-2**: the open-set retrieval task, *i.e.*, person re-identification (ReID).

For Task-1, Office-Home dataset [34] is a commonly used domain generalization (DG) benchmark on the task of classification. It consists of 4 domains (*i.e.*, Art (Ar), Clip Art (Cl), Product (Pr), Real-World (Rw)). Among them, three domains are used for training and the remaining one is considered as the unknown target domain for testing.

For Task-2, person ReID is a representative open-set retrieval task, where different domains do not share their label space. i) following [44, 83], we take four large-scale datasets (CUHK-SYSU (CS) [72], MSMT17 (MS) [70], CUHK03 (C3) [38] and Market-1501 (MA) [85]). For evaluation, a model is trained on three domains and tested on the remaining one. ii) several large-scale ReID datasets *e.g.*, CUHK02 [37], CUHK03, Market-1501 and CUHK-SYSU, are viewed as multiple source domains. Each small-scale ReID dataset including VIPeR [23], PRID [25], GRID [46] and iLIDS [86] is used as an unseen target domain. To comply with the General Ethical Conduct, we exclude DukeMTMC from the source domains.

We adopt ResNet-18 and ResNet-50 [24] as our backbone for Task-1 and Task-2, respectively. All reported results are obtained by the averages of five runs. We provide more implementation details in the supplementary material.

4.2. Ablation Study

4.2.1 The effectiveness of DFF

To investigate the effectiveness of our proposed Deep Frequency Filtering (DFF), we compare it with the ResNet baselines (*Base*) and the ResNet-based FFC [10] models

Table 1. Performance comparisons of our proposed Deep Frequency Filtering (DFF) with the baselines and the models with deep filtering in the original feature space. "*Base*" refers to the vanilla ResNet baseline. In "*SBase*", we use ResNet-based FFC in [10], serving as a strong baseline. "*Ori-F*" refers to a filtering operation in the original feature space, adopted in the local branch f_l and global branch f_g, respectively. When adopted in f_g, the FFT/iFFT operations are discarded from f_g so that the filtering is in the original feature space instead of the frequency space. "*Fre-F*" represents our proposed Deep Frequency Filtering.

Method	Source→Target					
	MS+CS+C3→MA		MS+MA+CS→C3		MA+CS+C3→MS	
	mAP	R1	mAP	R1	mAP	R1
Base	59.4	83.1	30.3	29.1	18.0	41.9
SBase (FFC)	66.2	84.7	35.8	35.4	19.4	44.8
Ori-F in f_l	66.9	85.0	36.2	35.9	19.8	45.1
Ori-F in f_g	61.9	83.5	32.7	31.9	18.4	42.8
Fre-F (Ours)	**71.1**	**87.1**	**41.3**	**41.1**	**25.1**	**50.5**

Table 2. Performance comparisons of different implementations of the frequency filtering operation. "*Task.*" refers to the filtering operation using a task-level attention mask where the mask is implemented with network parameters and is shared over different instances. "*Ins.*" denotes the filtering operation using learned instance-adaptive masks. "*C*" and "*S*" represents the filtering performed along the channel and spatial dimensions, respectively.

Method	Source→Target					
	MS+CS+C3→MA		MS+MA+CS→C3		MA+CS+C3→MS	
	mAP	R1	mAP	R1	mAP	R1
Base	59.4	83.1	30.3	29.1	18.0	41.9
Task.(C)	62.7	80.0	32.1	31.4	19.5	44.9
Task.(S)	68.6	85.8	37.0	36.3	20.8	45.4
Ins.(C)	69.8	86.2	36.4	35.9	21.0	45.7
Ins.(S) (Ours)	**71.1**	**87.1**	**41.3**	**41.1**	**25.1**	**50.5**

(*SBase*) that serve as the strong baselines, respectively. The experiment results are in Table 1. There are two crucial observations: 1) Our *Fre-F (DFF)* consistently outperforms *Base* by a clear margin. This demonstrates the effectiveness of our proposed method. With a simple instantiation, our proposed Deep Frequency Filtering operation can significantly improve the generalization ability of DNNs. 2) Our *Fre-F (DFF)* brings more improvements than *SBase (FFC)*. This indicates that frequency-domain convolutions are inadequate for model generalization. Instead, DFF is an effective solution for classification generalization task.

4.2.2 Frequency-domain v.s. Original-domain filtering

Feature filtering operations can be implemented in either the original feature domain or the frequency domain. We conduct experiments to study the impact of different implementation in both domains. Table 1 indicates that the proposed frequency-domain filtering (*Fre-F*) outperforms the original-domain feature filtering (*Ori-F*). This demonstrates the importance of modulating different frequency components in the latent space for domain generalization.

Table 3. The ablation results on the influence of the two-branch architecture. "*FFC*" refers to the frequency-domain convolution work [10] without our proposed filtering operation. "*only f_g*" denotes the setting in which we adopt the *FFC* or our proposed *DFF* over the complete feature without the splitting along the channel dimension, corresponding to using a single branch architecture. "f_l + f_g" represents the setting in which we split the feature along the channel dimension and adopt frequency-domain operations (*FFC* or our *DFF*) on one half of them.

Method	Source→Target					
	MS+CS+C3→MA		MS+MA+CS→C3		MA+CS+C3→MS	
	mAP	R1	mAP	R1	mAP	R1
Base	59.4	83.1	30.3	29.1	18.0	41.9
FFC(only f_g)	60.1	80.4	26.1	24.8	17.3	40.2
FFC($f_l + f_g$)	66.2	84.7	35.8	35.4	19.4	44.8
Ours(only f_g)	64.2	83.4	29.3	28.1	17.6	40.4
Ours($f_l + f_g$)	**71.1**	**87.1**	**41.3**	**41.1**	**25.1**	**50.5**

4.2.3 The importance of instance-adaptive attention

Our proposed DFF is allowed to be implemented using a spatial attention on the frequency representations of features. In Table 2, we compare the performance of adopting task-level and instance-level attention mask as well as channel and spatial filtering, respectively. We can observe that *Ins. (S)* clearly and consistently outperforms *Task. (S)* over all settings on both tasks. The results suggest that using instance-adaptive attention mask is essential for implementing our proposed DFF. We observe that different instances correspond to diversified frequency components in the feature space. Thus, a task-level attention mask is weak for all instances, which may explain the performance gaps in Table 2. We need to perform the modulation of feature frequency components with instance-adaptive weights.

4.2.4 Spatial v.s. Channel

DFF can be implemented with a spatial attention module, since different spatial positions of the frequency representations correspond to different frequencies. Admittedly, we can also adopt the channel attention to the frequency representations, which can be viewed as a representation refinement within each frequency component rather than perform Deep Frequency Filtering. The results in Table 2 show that spatial attention consistently outperforms channel attention in both task-level and instance-level settings. This indicates that frequency filtering (or selection) is more important than the refinement of frequency-domain representations for domain generalization.

4.2.5 The Influence of the Two-branch Architecture

As mentioned above, we adopt the two-branch architecture proposed by FFC [11] as a base architecture and apply our frequency filtering mechanism to the spectral transform branch. As shown in Table 3, the performances of

Table 4. Performance (classification accuracy %) comparison with the state-of-the-art methods on close-set classification task. We use ResNet-18 as backbone. Best in bold.

Method	Source→Target				Avg
	Cl,Pr,Rw→Ar	Ar,Pr,Rw→Cl	Ar,Cl,Rw→P	Ar,Cl,Pr→Rw	
Baseline	52.2	45.9	70.9	73.2	60.5
CCSA [51]	59.9	49.9	74.1	75.7	64.9
D-SAM [18]	58.0	44.4	69.2	71.5	60.8
MMD-AAE [35]	56.5	47.3	72.1	74.8	62.7
CrossGrad [61]	58.4	49.4	73.9	75.8	64.4
JiGen [8]	53.0	47.5	71.5	72.8	61.2
RSC [29]	58.4	47.9	71.6	74.5	63.1
MixStyle [88]	58.7	53.4	74.2	75.9	65.5
Ours	**65.4**	**53.7**	**74.9**	**76.5**	**67.6**

Table 5. Performance (%) comparison with the state-of-the-art methods on the open-set person ReID task. "Source" refers to the multiple training datasets, i.e., Market-1501 (MA), DukeMTMC-reID (D), CUHK-SYSU (CS), CUHK03 (C3) and CUHK02 (C2). "All" represents using MA+D+CS+C3+C2 as source domains. We do not include DukeMTMC-reID (D) in the training domains since this dataset has been discredited by the creators, denoted as "All w/o D".

Method	Source	Target:VIPeR(V)		Target:PRID(P)		Target:GRID(G)		Target:iLIDS(I)		Mean of V,P,G,I	
		R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
CDEL [43]	All	38.5	-	57.6	-	33.0	-	62.3	-	47.9	-
DIMN [62]	All	51.2	60.1	39.2	52.0	29.3	41.1	70.2	78.4	47.5	57.9
DDAN [9]	All	56.5	60.8	62.9	67.5	46.2	50.9	78.0	81.2	60.9	65.1
RaMoE [15]	All	56.6	64.6	57.7	67.3	46.8	54.2	85.0	90.2	61.5	69.1
SNR [31]	All	49.2	58.0	47.3	60.4	39.4	49.0	77.3	84.0	53.3	62.9
CBN [90]	All	49.0	59.2	61.3	65.7	43.3	47.8	75.3	79.4	57.2	63.0
Person30K [2]	All	53.9	60.4	60.6	68.4	50.9	56.6	79.3	83.9	61.1	67.3
DIR-ReID [80]	All	58.3	62.9	71.1	75.6	47.8	52.1	74.4	78.6	62.9	67.3
MetaBIN [13]	All	56.2	66.0	72.5	79.8	49.7	58.1	79.7	85.5	64.5	72.4
QAConv$_{50}$ [42]	All w/o D	57.0	66.3	52.3	62.2	48.6	57.4	75.0	81.9	58.2	67.0
M^{3}L [84]	All w/o D	60.8	68.2	55.0	65.3	40.0	50.5	65.0	74.3	55.2	64.6
MetaBIN [13]	All w/o D	55.9	64.3	61.2	70.8	50.2	57.9	74.7	82.7	60.5	68.9
Ours	All w/o D	**65.7**	**74.2**	**71.8**	**78.6**	**56.4**	**65.5**	**83.6**	**88.3**	**69.4**	**76.7**

FFC(only f_g) and *FFC($f_l + f_g$)* are inferior to *Ours(only f_g)* and *Ours($f_l + f_g$)* respectively, which indicates that simply filtering frequency components of features can bring striking improvement of the generalization ability. And the performance gap between *Ours($f_l + f_g$)* and Ours(only f_g) demonstrates that the two-branch structure for Post-filtering Feature Restitution can restore the globally filtered features. Furthermore, the complementary local filtering helps the learning of generalizable feature representation.

4.3. Comparison with the State-of-the-arts

4.3.1 Performances on close-set classification (Task-1)

In Table 4, we show the comparisons with the state-of-the-art approaches for Task-1 on Office-Home dataset. All our reported results are averaged over five runs. We observe that our proposed DFF outperforms most existing DG methods and achieves **67.6%** classification accuracy on average using ResNet-18, and outperforms the second-best method MixStyle [88] by **2.2%** average classification accuracy.

4.3.2 Performances on open-set retrieval (Task-2)

As shown in Table 5 and Table 6, our DFF model achieves significant performance improvements on all settings. Specifically, (see the Table 5), the mean performance of our method exceeds the second-best by **8.9%** in R1 accuracy and **7.8%** in mAP. As shown in Table 6, our DFF outperforms the second-best by **3.4%, 6.7%, 5.2%** in R1 accuracy and **8.0%, 7.1%, 7.3%** in mAP on Market-1501, CUHK03, MSMT17, respectively. When under the setting (*i.e.*, the testing set of the seen domains are also included for training model), our DFF performs better than the second-best by **7.3%, 8.1%, 5.2%** in R1 accuracy and **13.8%, 8.5%, 6.5%** in mAP. The results demonstrate that our DFF can significantly improve the generalization ability of learned features even with a simple learned filtering operation in the frequency domain.

4.4. Complexity Analysis

In Table 7, we compare the complexities of our DFF and vanilla ResNet models. The GFLOPs are calculated with input size of 224×224. FFT and inverse FFT are parameter-free and our filtering design uses average-pooling and max-pooling operations to reduce computational cost. Thus, our DFF module only brings limited extra GFLOPs and parameters. Our experiment results demonstrate significant performance gain over vanilla ResNet variants.

4.5. Visualization of Learned Masks

We visualize the learned masks at different depths used for Deep Frequency Filtering in our proposed scheme. The visualization results are in Fig. 2. We draw two observations: 1) The models equipped with DFF tend to enhance

Table 6. Performance (%) comparison with the state-of-the-art methods on the open-set person ReID task. Evalution on four large-scale person ReID benchmarks including Market-1501 (MA), Cuhk-SYSC (CS), CUHK03 (C3) and MSMT17 (MS). 'Com-" refers to combining the train and test sets of source domains for training. The $\mathrm{M}^3\mathrm{L}$ with a superscript * denote the model adopting IBN-Net50 as backbone. Without this superscript, ResNet-50 is taken as the backbone.

Method	Source	Market-1501		Source	CUHK03		Source	MSMT17	
		mAP	R1		mAP	R1		mAP	R1
SNR [31]	MS+CS+C3	34.6	62.7	MS+CS+MA	8.9	8.9	CS+MA+C3	6.8	19.9
$\mathrm{M}^3\mathrm{L}$ [84]		58.4	79.9		20.9	31.9		15.9	36.9
$\mathrm{M}^3\mathrm{L}^*$ [84]		61.5	82.3		34.2	34.4		16.7	37.5
QAConv_{50} [42]		63.1	83.7		25.4	24.8		16.4	45.3
MetaBIN [13]		57.9	80.0		28.8	28.1		17.8	40.2
Ours		**71.1**	**87.1**		**41.3**	**41.1**		**25.1**	**50.5**
SNR [31]	Com-(MS+CS+C3)	52.4	77.8	Com-(MS+CS+MA)	17.5	17.1	Com-(CS+MA+C3)	7.7	22.0
$\mathrm{M}^3\mathrm{L}$ [84]		61.2	81.2		32.3	33.8		16.2	36.9
$\mathrm{M}^3\mathrm{L}^*$ [84]		62.4	82.7		35.7	36.5		17.4	38.6
QAConv_{50} [42]		66.5	85.0		32.9	33.3		17.6	46.6
MetaBIN [13]		67.2	84.5		43.0	43.1		18.8	41.2
Ours		**81.0**	**92.3**		**51.5**	**51.2**		**25.3**	**51.8**

Table 7. Comparison of Parameters (#Para), GFLOPs and the Top-1 classification accuracy (Acc.) on ImageNet-1K between the models equipped with DFF and their corresponding base models. "*DFF-R18/R50*" denote the ResNet-18/-50 models equipped with our DFF.

Model	#Para	GFLOPs	Acc.	Model	#Para	GFLOPs	Acc.
ResNet18	11.7M	1.8	69.8	ResNet50	25.6M	4.1	76.3
DFF-R18	12.2M	2.0	72.3	DFF-R50	27.7M	4.5	77.9

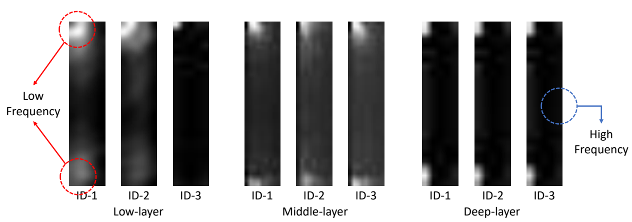

Figure 2. Visualization of the learned spatial masks of DFF for the filtering in the frequency domain. The grayscale denotes the mask value, while "white" and "black" correspond to "1" and "0", respectively. For each mask, the left top and bottom corners correspond to the low-frequency components while the right middle corresponds to the high-frequency components. The masks at different depths are resized to the same resolution.

relatively low frequency components while suppressing relatively high ones in the latent frequency domain, which is consistently observed over different instances (IDs). This is in line with the results of theoretical study on the relationship between frequency and robustness [66, 74, 75]. 2) The learned masks are instance-adaptive, demonstrating our proposed scheme can achieve instance-adaptive frequency modulation/filtering in the latent space.

4.6. Visualization of Learned Feature Maps

We compare the feature maps extracted by the model equipped with our proposed DFF and the one without DFF (see Fig. 3). We observe that the features learned by the model equipped with DFF have higher responses for human-body regions than those learned by the baselines model without DFF. This indicates that DFF enables neural networks to focus more precisely on target regions and suppress unrelated feature components (*e.g.*, backgrounds).

Figure 3. We compare the feature maps extracted by the model without (left) and the model with DFF (right). The *lighter* the color is, the *larger* the feature value is.

Moreover, we also perform t-SNE visualization for the ReID feature vectors learned by the baseline and the model with DFF. The results and analysis are in the supplementary.

5. Conclusion

In this paper, we first conceptualize Deep Frequency Filtering (DFF) and point out that such a simple mechanism can significantly enhance the generalization ability of deep neural networks across domains. This provides a novel alternative for this field. Furthermore, we discuss the implementations of DFF and showcase the implementation with a simple spatial attention in the frequency domain can bring stunning performance improvements for DG. Extensive experiments and ablation studies demonstrate the effectiveness of our proposed method. We leave the exploration on more effective instantiations of our conceptualized DFF in the future work, and encourage more combinations and interplay between conventional signal processing and deep learning technologies.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6836–6846, 2021. 3

[2] Yan Bai, Jile Jiao, Wang Ce, Jun Liu, Yihang Lou, Xuetao Feng, and Ling-Yu Duan. Person30k: A dual-meta generalization network for person re-identification. In *CVPR*, pages 2123–2132, 2021. 7

[3] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. *NIPS*, 31, 2018. 2

[4] Gregory A Baxes. *Digital image processing: principles and applications*. John Wiley & Sons, Inc., 1994. 2

[5] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *ICCV*, pages 3286–3295, 2019. 3

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, pages 813–824, 2021. 3

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020. 3

[8] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, pages 2229–2238, 2019. 7

[9] Peixian Chen, Pingyang Dai, Jianzhuang Liu, Feng Zheng, Qi Tian, and Rongrong Ji. Dual distribution alignment network for generalizable person re-identification. In *AAAI*, volume 6, 2021. 7

[10] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In *NeurIPS*, pages 4479–4488, 2020. 2, 3, 5, 6

[11] Lu Chi, Guiyu Tian, Yadong Mu, Lingxi Xie, and Qi Tian. Fast non-local neural networks with spectral residual learning. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2142–2151, 2019. 6

[12] Kamran Chitsaz, Mohsen Hajabdollahi, Nader Karimi, Shadrokh Samavi, and Shahram Shirani. Acceleration of convolutional neural network using fft-based split convolutions. *arXiv preprint arXiv:2003.12621*, 2020. 2

[13] Seokeon Choi, Taekyung Kim, Minki Jeong, Hyoungseob Park, and Changick Kim. Meta batch-instance normalization for generalizable person re-identification. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2021. 7, 8

[14] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965. 3

[15] Yongxing Dai, Xiaotong Li, Jun Liu, Zekun Tong, and Ling-Yu Duan. Generalizable person re-identification with relevance-aware mixture of experts. In *CVPR*, pages 16145–16154, 2021. 7

[16] Zihang Dai, Hanxiao Liu, Quoc Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. In *NeurIPS*, volume 34, 2021. 5

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[18] Antonio D'Innocente and Barbara Caputo. Domain generalization with domain-specific aggregation modules. In *GCPR*, pages 187–198. Springer, 2018. 7

[19] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, pages 6824–6835, 2021. 3

[20] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 5

[21] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *ICCV*, pages 2551–2559, 2015. 2

[22] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. Dlow: Domain flow for adaptation and generalization. In *CVPR*, pages 2477–2486, 2019. 2

[23] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, pages 262–275, 2008. 5

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[25] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *SCIA*, pages 91–102, 2011. 5

[26] Narges Honarvar Nazari and Adriana Kovashka. Domain generalization using shape representation. In *ECCV*, pages 666–670, 2020. 2

[27] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 3, 5

[28] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *CVPR*, pages 6891–6902, 2021. 2

[29] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. Self-challenging improves cross-domain generalization. In *ECCV*, pages 124–140. Springer, 2020. 7

[30] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *NeurIPS*, 28, 2015. 3

[31] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *CVPR*, pages 3143–3152, 2020. 2, 7, 8

[32] Yitzhak Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004. 2

[33] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, volume 32, 2018. 2

[34] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, pages 5542–5550, 2017. 5

[35] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409, 2018. 7

[36] Shaohua Li, Kaiping Xue, Bin Zhu, Chenkai Ding, Xindi Gao, David Wei, and Tao Wan. Falcon: A fourier transform based approach for fast and secure convolutional neural network predictions. In *CVPR*, pages 8705–8714, 2020. 2

[37] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *CVPR*, pages 3594–3601, 2013. 5

[38] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 5

[39] Xin Li, Zhizheng Zhang, Guoqiang Wei, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Confounder identification-free causal visual feature learning. *CoRR*, abs/2111.13420, 2021. 2

[40] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, pages 624–639, 2018. 2

[41] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *ICML*, pages 3915–3924. PMLR, 2019. 2

[42] Shengcai Liao and Ling Shao. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In *ECCV*, pages 456–474. Springer, 2020. 7, 8

[43] Ci-Siang Lin, Yuan-Chia Cheng, and Yu-Chiang Frank Wang. Domain generalized person re-identification via cross-domain episodic learning. In *ICPR*, pages 6758–6763. IEEE, 2021. 7

[44] Jiawei Liu, Zhipeng Huang, Liang Li, Kecheng Zheng, and Zheng-Jun Zha. Debiased batch normalization via gaussian process for generalizable person re-identification. In *AAAI*, 2022. 2, 5

[45] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *CVPR*, pages 1013–1023, 2021. 2, 3

[46] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1988–1995. IEEE, 2009. 5

[47] Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. Best sources forward: domain generalization through source-specific nets. In *ICIP*, pages 1353–1357, 2018. 2

[48] Xintian Mao, Yiming Liu, Wei Shen, Qingli Li, and Yan Wang. Deep residual fourier transformation for single image deblurring. *arXiv preprint arXiv:2111.11745*, 2021. 2

[49] Michael Mathieu, Mikael Henaff, and Yann LeCun. Fast training of convolutional networks through ffts. *arXiv preprint arXiv:1312.5851*, 2013. 2

[50] Diganta Misra, Trikay Nalamada, Ajay Uppili Arasanipalai, and Qibin Hou. Rotate to attend: Convolutional triplet attention module. In *WACV*, pages 3139–3148, 2021. 5

[51] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *ICCV*, pages 5715–5725, 2017. 2, 7

[52] Varsha Nair, Moitrayee Chatterjee, Neda Tavakoli, Akbar Siami Namin, and Craig Snoeyink. Fast fourier transformation for optimizing convolutional neural networks in object recognition. *arXiv preprint arXiv:2010.04257*, 2020. 2

[53] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018. 2

[54] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *NeurIPS*, 2021. 3

[55] Ioannis Pitas. *Digital image processing algorithms and applications*. John Wiley & Sons, 2000. 2

[56] Anish Prabhu, Ali Farhadi, Mohammad Rastegari, et al. Butterfly transform: An efficient fft based neural architecture design. In *CVPR*, pages 12024–12033, 2020. 2

[57] Harry Pratt, Bryan Williams, Frans Coenen, and Yalin Zheng. Fcnn: Fourier convolutional neural networks. In *ECML PKDD*, pages 786–798, 2017. 2

[58] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. Fcanet: Frequency channel attention networks. In *ICCV*, pages 783–792, 2021. 3, 5

[59] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *NeurIPS*, volume 34, 2021. 2, 5

[60] Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *ECCV*, pages 68–83, 2020. 2

[61] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. Generalizing across domains via cross-gradient training. *ICLR*, 2018. 2, 7

[62] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *CVPR*, pages 719–728, 2019. 7

[63] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, pages 2149–2159, 2022. 2

[64] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *CVPR workshops*, pages 969–977, 2018. 2

[65] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. *NeurIPS*, 31, 2018. 2

[66] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *CVPR*, 2020. 1, 2, 3, 8

[67] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *CVPR*, pages 11531–11539, 2020. 3

[68] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 3, 5

[69] Guoqiang Wei, Cuiling Lan, Wenjun Zeng, Zhizheng Zhang, and Zhibo Chen. Toalign: task-oriented alignment for unsupervised domain adaptation. *NeurIPS*, 34:13834–13846, 2021. 2

[70] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018. 5

[71] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 3, 4, 5

[72] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *CVPR*, pages 3415–3424, 2017. 5

[73] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *CVPR*, pages 14383–14392, 2021. 2, 3

[74] Zhiqin John Xu. Understanding training and generalization in deep learning by fourier analysis. *arXiv preprint arXiv:1808.04295*, 2018. 1, 2, 3, 8

[75] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *ICONIP*, 2019. 1, 2, 3, 8

[76] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *CVPR*, pages 4085–4095, 2020. 2, 3

[77] Qiaosi Yi, Jinhao Liu, Le Hu, Faming Fang, and Guixu Zhang. Contrastive learning for local and global learning mri reconstruction. *arXiv preprint arXiv:2111.15200*, 2021. 2

[78] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *NeurIPS*, 2019. 1, 2

[79] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*, pages 2100–2110, 2019. 2

[80] Yi-Fan Zhang, Hanlin Zhang, Zhang Zhang, Da Li, Zhen Jia, Liang Wang, and Tieniu Tan. Learning domain invariant representations for generalizable person re-identification. *arXiv preprint arXiv:2103.15890*, 2021. 7

[81] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *CVPR*, pages 10407–10416, 2020. 3

[82] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *CVPR*, pages 3186–3195, 2020. 3, 5

[83] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*, 2021. 5

[84] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Nicu Sebe. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*, pages 6277–6286, 2021. 7, 8

[85] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 5

[86] Wei-Shi Zheng, Shaogang Gong, and Tao Xiang. Associating groups of people. In *BMVC*, volume 2, pages 1–11, 2009. 5

[87] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation. In *AAAI*, pages 13025–13032, 2020. 2

[88] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *TIP*, 2021. 2, 7

[89] Ying Zhu, Runwei Ding, Weibo Huang, Peng Wei, Ge Yang, and Yong Wang. Hmfca-net: Hierarchical multi-frequency based channel attention net for mobile phone surface defect detection. *Pattern Recognit Lett*, 153:118–125, 2022. 3

[90] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *ECCV*, pages 140–157. Springer, 2020. 7