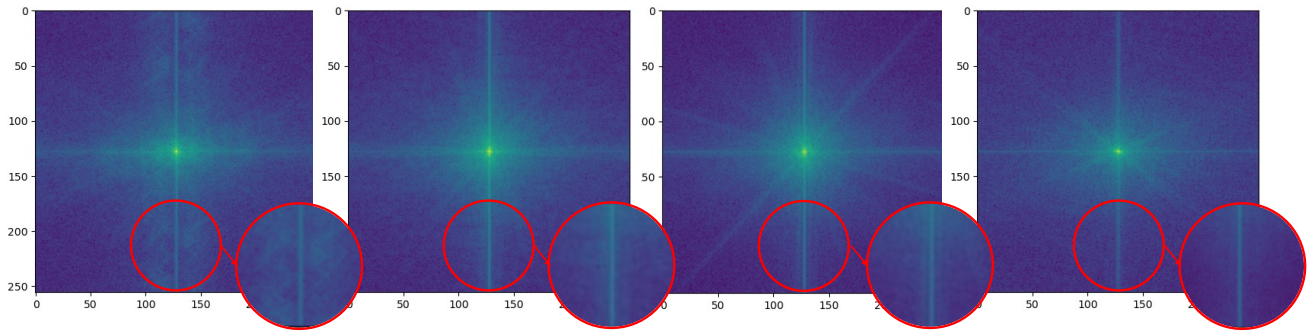
Domain-Generalized Object Anti-Spoofing: Bridging Gaps and Patch Selection for Robust Detection across Domains

– Supplementary Material –



(a) Largest power spectrum (b) Smallest power spectrum (c) Random power spectrum (d) Real image power spectrum

Figure A. Power spectrum analysis. (a), (b), and (c) depict the power spectrum of spoof images. (d) presents the power spectrum of a real image. In (a), patches with larger power spectrum values indicate the presence of more significant noise compared to (b), (c), and (d). In this approach, we select patches with a large power spectrum to detect high-frequency anomalies, such as moiré patterns and atypical illuminations, commonly encountered in spoofing images.

A. Analyses on Frequency Domain

Figure A represents the power spectrum analysis of the spoof and real images. Among the spoof images (a-c), it is evident that patches featuring the largest power spectra conspicuously demonstrate a significantly heightened presence of noise in comparison to patches with the smallest power spectra and random spectra. Moreover, when compared to the authentic images in (d), (a) exhibits more pronounced noise. The results show that patches with higher power spectrum values encompass a significant presence of high-frequency components, such as moiré patterns and atypical illuminations, as shown in the original paper.

To capture these localized spoofing cues present within images, we propose a frequency domain analysis-based patch selection method. Our experimental results in the original manuscript demonstrate that extracting patches based on frequency domain analysis yields better results compared to random patch extraction.

Figure B in the main experimental section of the paper presents an enlarged view of the magnitude of Figure 3. The image at the top of the grid represents the average magnitude of patches with the larger power spectrum we proposed. In contrast, the image at the bottom depicts the results obtained by sampling a smaller power spectrum, in contrast to our proposed approach. As evident from the depicted images, the magnitudes created using our proposed larger power spectrum exhibit distinct differences, particularly in high frequencies, compared to the other three magnitudes.

B. Domain Generalization Results

Table A presents the results of training the backbone network trained on OULU-NPU with the MToFNet dataset. It shows the significant efficiency of the proposed method in terms of training parameters. Training the entire parameters of the ResNet-18 network is approximately 83 times more than training through the LoRA module. Despite having fewer training parameters, the proposed method achieves the best average HTER and EER on the DoFNet and GOSet datasets. While the MToFNet and DoFNet datasets are distinct, the MToFNet dataset includes some data that is similar to the DoFNet dataset. Consequently, GOLab, ResNet-18, and ResNet-18 with our patch selection exhibit significant overfitting to the training data from MToFNet, resulting in lower error rates on DoFNet. However, all of these methods surpass an error rate of 40% on the GOSet dataset. Table B represents the results of training the backbone network trained on OULU-NPU with the DoFNet dataset. The proposed method achieves the best average HTER and AUC on the MToFNet and GOSet datasets at 15.3% and 87.9%, respectively.

C. Training and Inference

In both training and inference, we consistently use the power spectrum to extract the top k patches from each image. This method ensures that we focus on the most informative regions during both phases. The reason this approach is feasible is that, during training with the FAS dataset, we

Table A. Experimental results on the domain-generalization with face dataset and object dataset. The blue checks represent training through the LoRA module on the object dataset. The red fonts indicate the addition of MToFNet training into the same backbone that was pre-trained with OULU-NPU. The term #Param. indicates the number of train parameters when the backbone network is trained on object dataset. ResNet-18 undergoes fine-tuning across all parameters. In contrast, the proposed method is trained using notably fewer parameters through the utilization of the LoRA module.

Method	Train set		#Param.	DoFNet			GOSet			AVG.		
	Face	Object		HTER	EER	AUC	HTER	EER	AUC	HTER	EER	AUC
Atoum et al.	✓	✓	3.5M	19.0	18.3	91.8	39.6	39.7	64.9	29.3	29.0	78.3
GOLab	✓	✓	3.0M	**1.5**	**1.5**	**99.9**	41.0	41.1	65.3	21.2	21.3	82.6
ResNet-18	✓	✓	11.17M	8.3	7.6	97.4	40.2	40.1	63.3	24.2	23.8	80.3
ResNet-18 w Patch	✓	✓	11.22M	7.6	7.6	97.2	49.8	49.7	48.1	28.7	28.6	72.6
Ours	✓	✓	0.136M	6.6	7.8	97.1	**27.3**	**27.3**	**80.6**	**17.5**	**16.9**	**88.8**

Table B. Experimental results on the domain-generalization with face dataset and object dataset. The blue checks represent training through the LoRA module on the object dataset. The red fonts indicate the addition of DoFNet training into the same backbone that was pre-trained with OULU-NPU. The term #Param. indicates the number of train parameters when the backbone network is trained on object dataset. ResNet-18 undergoes fine-tuning across all parameters. In contrast, the proposed method is trained using notably fewer parameters through the utilization of the LoRA module.

Method	Train set		#Param.	MToFNet			GOSet			AVG.		
	Face	Object		HTER	EER	AUC	HTER	EER	AUC	HTER	EER	AUC
Atoum et al.	✓	✓	3.5M	46.1	46.1	55.9	44.1	44.0	59.4	45.1	45.0	57.6
GOLab	✓	✓	3.0M	50.0	49.3	47.3	58.5	58.6	37.4	54.2	53.9	42.3
ResNet-18	✓	✓	11.17M	20.5	20.4	87.8	33.2	33.3	69.0	26.8	26.8	78.4
ResNet-18 w Patch	✓	✓	11.22M	9.6	9.6	95.3	34.9	34.8	71.8	22.2	22.2	83.5
Ours	✓	✓	0.136M	**4.0**	**4.0**	**99.0**	**26.6**	**26.6**	**76.8**	**15.3**	**15.3**	**87.9**

excluded any images with partial attacks, ensuring that only fully spoofed or real images were used. Additionally, in the OAS dataset, partial attacks are not present, further justifying the use of this patch selection method without concern for missing important spoof cues in localized areas. This consistency enhances the model's ability to generalize across both FAS and OAS datasets.

D. Description of LoRA for Convolution Layer

We applied LoRA to the convolutional layers of a ResNet network, enhancing domain generalization performance compared to fine-tuning all model parameters. The original convolutional layer equation is as follows:

$$y = W * x + b \tag{1}$$

where x represents the input, W denotes the filter (weights) of the convolutional layers, and b is the bias term. The symbol $*$ signifies the convolution operation. The equation for a convolutional layer with LoRA applied is as follows:

$$y = (W + \alpha \cdot (B \times A)) * x + b \tag{2}$$

where α is the scaling factor, and A and B are the LoRA matrices.

By applying LoRA, we decompose the adjustment to the convolutional weights into two low-rank matrices A and B, which are learned during fine-tuning. This approach allows us to efficiently adapt the convolutional filters with fewer parameters compared to fine-tuning the entire weight matrix W. This results in a more parameter-efficient adaptation process, improving domain generalization without the need for extensive retraining of the entire network. Additionally, LoRA's low-rank adaptation helps in reducing overfitting by limiting the number of trainable parameters, making the model more robust to new, unseen domains.

E. Ablation Study

In this section, we report the ablation studies regarding LoRA positional configuration and patch configuration(the number and size of the patch) of the proposed method.

E.1. LoRA Configuration

As demonstrated in Figure A, capturing subtle spoofing cues necessitates an understanding of low-level information. Through the exploration of architectural variations, we conclude that fusing the first convolution block with the stage 1 layer leads to optimal performance outcomes, as depicted in Table C. This aligns with the hypothesis that com-

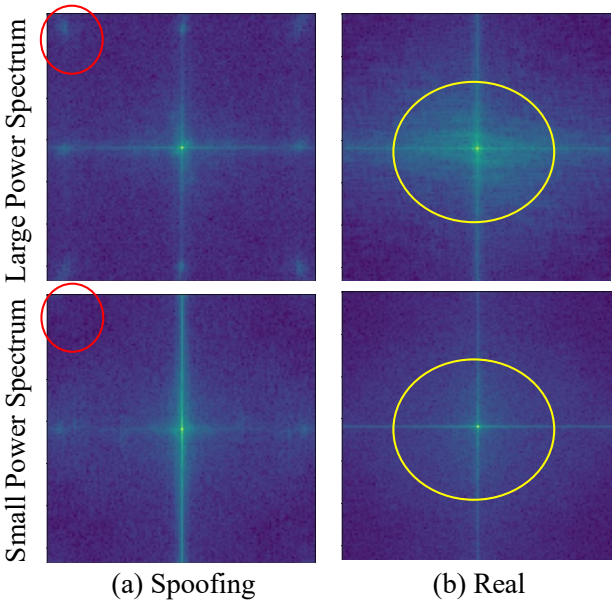


Figure B. This serves as an extra depiction for Figure 3 in the primary experimental section. (a) illustrates the magnitude of spoofing patches, while (b) showcases the magnitude of patches extracted from real images. The upper row reflects sampling predicated on higher spectrum values within the patches, whereas the lower row involves sampling based on lower spectrum values. For spoofing images, the differences are evident at the ends of the large power spectrum (i.e., high frequency components), whereas for real images, the differences appear in the middle (i.e., low frequency components).

Table C. Ablation study on LoRA configuration. We evaluate the experiments on the DoFNet dataset. Each Configuration denotes the position ResNet-18 network.

LoRA Configuration	HTER	EER	AUC
Conv1	21.8	22.4	88.2
Stage1	29.9	29.5	80.0
Stage2	26.9	26.5	78.4
Stage3	53.8	53.0	39.6
Stage4	19.7	19.8	**91.3**
FC	22.5	22.4	86.8
Ours (Conv1 + Stage1)	**17.2**	**17.8**	87.3

prehending low-level information is crucial for capturing subtle cues, as discussed earlier in the original manuscript.

E.2. Patch Size

Table E presents the results of the ablation study conducted on patch sizes. Based on these evaluations, we opted for an image patch size of 64. The patch images of sizes 128 and 256 are more likely to include unnecessary information. An image with a patch size of 32 is too small to capture diverse information effectively.

E.3. Patch Number

Table F represents the results of the ablation study conducted on the selected number of patches. Based on these findings, we chose this specific patch number. Similar to the results of the patch size ablation study, when the number of patches is high, it can lead to the inclusion of unnecessary information. Similar to the findings of the patch size ablation study, a higher number of patches can lead to the inclusion of unnecessary information. Conversely, having only 1 or 2 patches is too few to effectively capture information.

E.4. Learning Rate of LoRA Module

Table G presents the results of the ablation study conducted on learning rate of LoRA module. Based on these experiments, we use a learning rate of 1e-4, as it resulted in the best HTER and EER performance.

E.5. ResNet Backbone Scale

We compare the performance in the same LoRA settings as the backbone model size increases, as shown in Table D. As the results indicate, enlarging the model size does not lead to performance improvements. For a fair comparison in most anti-spoofing tasks, we use ResNet-18. Therefore, we have fine-tuned our method for the ResNet-18 setting, and it is observed that using ResNet-50 or 101 with the same settings results in decreased performance.

F. Dataset Description

The Object Anti-spoofing dataset employed in this study encompasses the following three components:

F.1. GOSet

GOSet [3] constitutes a spoofing dataset comprising 2,849 video samples, organized into 24 spoof subjects and 7 spoof mediums. The GOSet data, captured in video format, possess the characteristic of blurred images, with simple backgrounds and a zoomed-in focus on the objects.

F.2. DoFNet

DoFNet [1] consists of 2,757 images. It incorporates 6 subjects and includes 3 attack mediums. Unlike other datasets, DoFNet offers two images with differing focal points, one focusing on the foreground and the other on the background simultaneously. Due to its image-based nature, DoFNet provides a very limited number of images, making domain generalization without utilizing Depth of Field highly restricted.

Table D. Ablation study on ResNet backbone scale. We used the GOSet dataset to train the models. We evaluate the experiments on the DoFNet dataset and MToFNet dataset.

Method	Train set		DoFNet			MToFNet			AVG.		
	Face	Object	HTER	EER	AUC	HTER	EER	AUC	HTER	EER	AUC
ResNet-50	✓	✓	31.4	31.1	77.0	26.2	26.4	80.1	28.8	28.7	78.5
ResNet-101	✓	✓	42.6	42.8	59.6	33.9	33.6	72.3	38.2	38.2	65.9
Ours (ResNet-18)	✓	✓	**17.2**	**17.8**	**87.3**	**8.0**	**7.7**	**97.2**	**12.6**	**12.7**	**92.2**

Table E. Ablation study varying the size of the input image patches. The evaluation is conducted on the DoFNET dataset.

Patch Size	HTER	EER	AUC
32	32.4	31.6	73.0
128	37.8	37.7	59.3
256	38.0	38.2	68.5
Ours (64)	**17.2**	**17.8**	**87.3**

Table F. Ablation study varying the number of the input image patches with 64×64 size. The entire evaluation is conducted on the DoFNET dataset.

Patch Num	HTER	EER	AUC
1	22.8	22.9	86.8
2	22.8	22.9	80.7
3	26.9	26.5	82.0
4	26.1	26.5	76.3
10	29.1	28.0	75.7
Ours (5)	**17.2**	**17.8**	**87.3**

Table G. Ablation study on the learning rate of LoRA module. The evaluation is conducted on the DoFNET dataset.

Learning Rate	HTER	EER	AUC
5e-5	18.2	18.3	**90.1**
1e-3	20.8	20.4	85.8
1e-2	36.5	36.7	67.9
Ours (1e-4)	**17.2**	**17.8**	87.3

F.3. MToFNet

MToFNet [2] comprises 12,529 images, encompassing 27 spoof subjects and 16 spoof mediums. It also simultaneously provides depth maps collected using a Time-of-Flight sensor. Notably, MToFNet includes images from DoFNet.

References

[1] Yonghyun Jeong, Jongwon Choi, Doyeon Kim, Sehyeon Park, Minki Hong, Changhyun Park, Seungjai Min, and Youngjune Gwon. Dofnet: Depth of field difference learning for detecting image forgery. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3

[2] Yonghyun Jeong, Doyeon Kim, Jaehyeon Lee, Minki Hong, Solbi Hwang, and Jongwon Choi. mtofnet: Object anti-spoofing with mobile time-of-flight data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 38–47, 2022. 4

[3] Joel Stehouwer, Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Noise modeling, synthesis and classification for generic object anti-spoofing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7294–7303, 2020. 3