# Frequency Shuffling and Enhancement for Open Set Recognition

**Lijun Liu[1,2], Rui Wang[1,2*], Yuan Wang[3],**
**Lihua Jing[1,2], Chuan Wang[1]**
[1]Institute of Information Engineering, CAS, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Department of Electronic Engineering,Tsinghua University
{liulijun, wangrui, jinglihua, wangchuan}@iie.ac.cn, wy23@mails.tsinghua.edu.cn

## Abstract

Open-Set Recognition (OSR) aims to accurately identify known classes while effectively rejecting unknown classes to guarantee reliability. Most existing OSR methods focus on learning in the spatial domain, where subtle texture and global structure are potentially intertwined. Empirical studies have shown that DNNs trained in the original spatial domain are inclined to over-perceive subtle texture. The biased semantic perception could lead to catastrophic over-confidence when predicting both known and unknown classes. To this end, we propose an innovative approach by decomposing the spatial domain to the frequency domain to separately consider global (low-frequency) and subtle (high-frequency) information, named Frequency Shuffling and Enhancement (FreSH). To alleviate the overfitting of subtle texture, we introduce the High-Frequency Shuffling (HFS) strategy that generates diverse high-frequency information and promotes the capture of low-frequency invariance. Moreover, to enhance the perception of global structure, we propose the Low-Frequency Residual (LFR) learning procedure that constructs a composite feature space, integrating low-frequency and original spatial features. Experiments on various benchmarks demonstrate that the proposed FreSH consistently trumps the state-of-the-arts by a considerable margin.

## Introduction

Deep Neural Networks (DNNs) have dominated various visual recognition tasks and yielded outstanding performance (He et al. 2016; Krizhevsky, Sutskever, and Hinton 2017). Traditional DNNs show strong capabilities in Closed-Set Recognition (CSR), where test samples are limited to known classes that appeared in training phase (*i.e., closed-set*). In the more realistic and challenging open-set setting, the model may face unknown classes (*i.e., open-set*) during inference, and vanilla DNNs tend to predict them incorrectly as known classes with high confidence (Yang et al. 2020).

Recently, the Open-Set Recognition (OSR) task (Scheirer et al. 2012) has drawn considerable attention, which not only requires distinguishing between the training categories but also indicates whether an image belongs to a category that has never encountered (Yang et al. 2020). Existing efforts for OSR are inspired by two aspects: generative-based and discriminative-based methods. The generative-based methods generate unknown classes (Ge et al. 2017; Chen et al. 2021b) or reconstruct known classes (Oza and Patel 2019; Huang et al. 2022). The discriminative-based methods focus on designing various distance loss functions to constrain embedding distributions  (Chen et al. 2020; Lu et al. 2022). However, the above methods train with the original spatial-domain images (marked as 'Original' in Figure 1), where the global structure and local texture of the object are potentially coupled together. In fact, these semantic elements play distinct roles during representation learning. Inspired by the frequency domain theory, the spatial-domain image can be disassembled through frequency transformation (Wang et al. 2020; Yao et al. 2022), where the global structure is encompassed within low-frequency bands (marked as 'Low' in Figure 1), and the local texture is contained in high-frequency bands (marked as 'High').

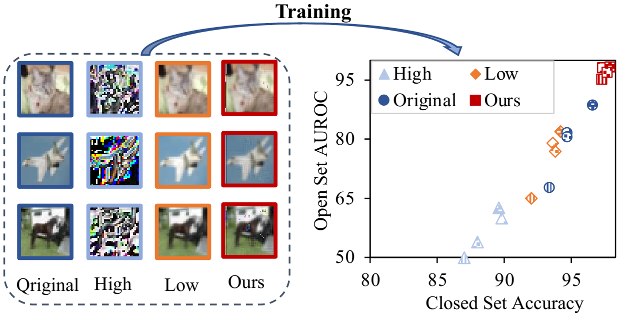

Figure 1: A comparison of spatial-domain models trained with the high-frequency band, low-frequency band, original spatial images, and the FreSH on four datasets. CNNs trained with the original images exhibit similar performance trends to those trained with low-frequency bands.

As illustrated in the left of Figure 1, it is more straightforward for humans to identify low-frequency images compared to unintuitive high-frequency images. And this observation has also inspired other computer vision researchers (Van den Branden Lambrecht and Kunt 1998; Luo et al. 2022). However, empirical studies on neural networks (Yin et al. 2019; Wang et al. 2020) reveal that CNNs trained in the original spatial domain are more inclined to

[*]Corresponding author

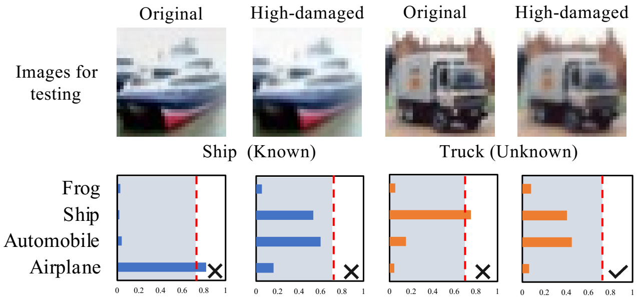

Figure 2: Unexpected prediction reversal when using high-frequency damaged images. The red line is the confidence threshold, below which is identified as unknown.

fit high-frequency visual representations, as shown in Figure 2. When the high-frequency bands are damaged, humans can easily maintain predictions the same as the original spatial domain, but the predictions of the CNNs are unexpectedly reversed (The closed-set "Ship" mislabeled as "Unknown" and a mislabeled "Truck" turns to "Unknown"). These inversions reveal that the spatial-domain CNNs are excessively sensitive to the changes in high-frequency information, thereby posing challenges in obtaining satisfactory confidence scores for OSR task. That is the over-confidence phenomenon (on known and unknown classes) could be attributed to the inductive bias learned in the spatial domain.

To correct the inductive bias and alleviate the over-confidence for OSR, we propose a novel **Fre**quency-based **S**huffling and En**h**ancement (FreSH) framework, which changes the network's preference for frequency bands. The FreSH framework consists of High-Frequency Shuffling (HFS) strategy and Low-Frequency Residual (LFR) learning. Firstly, the HFS strategy is designed to alleviate the overfitting of high-frequency texture during the training phase. It constructs disparate high-frequency variants and mixes them with the original low-frequency bands, which sparks the CNNs' potential for learning robust high-frequency features and focusing on the invariance of low-frequency information. Secondly, to emphasize the superiority of low-frequency information in recognition, we skip-connect the low-frequency features from shallow layers to deeper layers, which is called Low-Frequency Residual (LFR) learning. In this way, the global structure of objects is enhanced in the composite feature space that integrates frequency and spatial features. Overall, our FreSH framework captures more discriminative features by calibrating the fitting bias of traditional spatial-domain CNNs, thereby achieving satisfactory accuracy and robustness.

Our main contributions can be summarized as follows: *i)* We propose Frequency Shuffling and Enhancement (FreSH) framework for OSR, which relieves the induction bias of the original spatial domain and alleviates the over-confidence for unknown detection and known classification. *ii)* We propose the High-Frequency Shuffling (HFS) strategy to encourage the network to learn robust high-frequency bands and the Low-Frequency Residual (LFR) learning to enhance the global structure of objects. *iii)* Extensive experiments on multiple benchmarks demonstrate that the proposed method significantly improves the performance of OSR and remarkably surpasses existing methods.

## Related Work

**Open-Set Recognition.** The OSR task is outlined by (Scheirer et al. 2012) from the perspective of open space risk. The current methods ameliorate the OSR performance with the powerful representation capabilities inherent in DNNs, which can be broadly categorized into generative-based methods and discriminative-based methods. The first category is dedicated to generating known or unknown classes. G-openmax (Ge et al. 2017) and OSRCI (Neal et al. 2018) employ the formidable power of GANs (Goodfellow et al. 2014) to generate unknown samples. AEs (Kingma and Welling 2013) are employed by (Yoshihashi et al. 2019; Sun et al. 2020) to reconstruct known classes. C2AE (Oza and Patel 2019), GFROSR (Perera et al. 2020) and CapsNet (Guo et al. 2021) use Conditional Variational Auto-Encoder (CVAE) to minimize reconstruction errors. While the generative methods rely on auxiliary networks, which inevitably incur extra computational costs. The second category is discriminative-based approaches, which optimize the classifier or feature extractor. OpenMax (Bendale and Boult 2016) replaces the softmax operator with OpenMax and CPN (Yang et al. 2020) introduce convolutional prototype network. PROSER (Zhou, Ye, and Zhan 2021) proposes classifier placeholders for unknown classes. Hybrid (Zhang et al. 2020) adds a flow density estimator to reject unknown samples. Other methods optimize the feature embedding with prototype learning (Yang et al. 2018; Lu et al. 2022), supervised contrastive learning (Kodama et al. 2021) and spatial attention mechanism (Liu et al. 2022). However, most of them focus on optimization within the spatial domain. Contrary to them, we innovatively address the OSR task from the perspective of frequency domain.

**Frequency Domain Learning.** Frequency domain analysis is a powerful tool to expose the semantic elements of images. Recently, frequency domain theory has shown overwhelming performance on various DNNs tasks, such as image super-resolution (Li, You, and Robles-Kelly 2018; Fritsche, Gu, and Timofte 2019), image rescaling (Xiao et al. 2020), forgery detection (Li et al. 2021; Wang et al. 2023) and adversarial attacks(Sharma, Ding, and Brubaker 2019; Yin et al. 2019). Yin *et al.*(Yin et al. 2019) find that trained deep models are susceptive to high-frequency perturbations in adversarial settings. Guo *et al.*(Guo, Frank, and Weinberger 2018; Sharma, Ding, and Brubaker 2019) propose adversarial attacks targeting low-frequency images, which demonstrates the pivotal role of low-frequency bands in the model prediction. Some works exploit frequency domain theory to explain the generalization of CNNs. Wang *et al.*(Wang et al. 2020) notice that CNNs could capture high-frequency features that are usually untraceable to humans. Chen *et al.*(Chen et al. 2021a) qualitatively study the impact of magnitude spectrum and phase spectrum on the generalization behavior of CNNs.

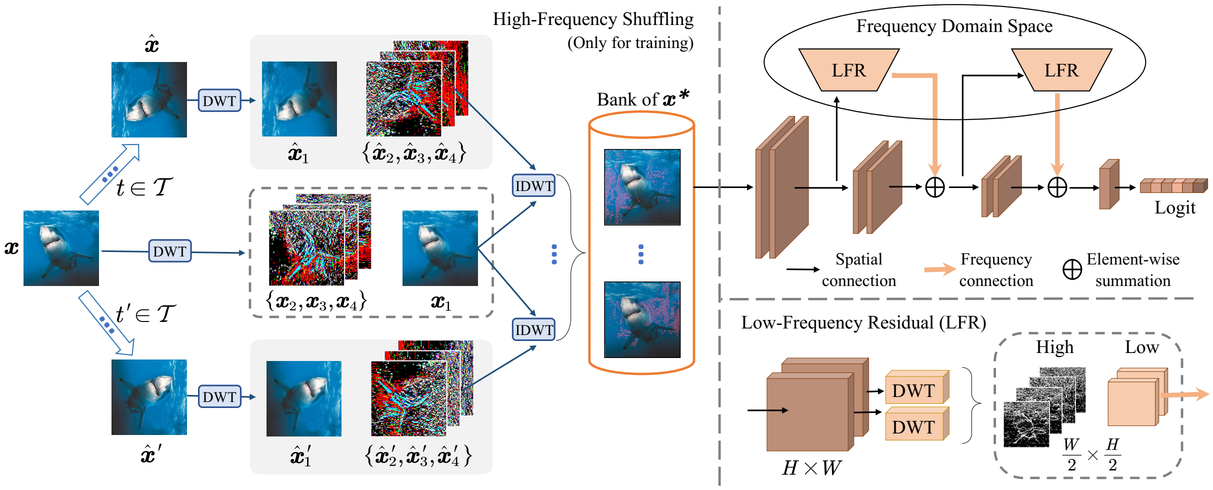

Figure 3: Overview of the proposed Frequency Shuffling and Enhancement (FreSH) framework. The High-Frequency Shuffling (HFS) strategy is denoted on the left. HFS randomly generates different high-frequency materials for every instance, and we draw it twice for simplicity. The low-frequency residual connection is performed between two convolutional stages. Without loss of generality, we show a backbone consisting of four convolutional stages.

## Method

In this section, we first review the formulation of a typical frequency transform method — Discrete Wavelet Transform (DWT), which establishes a foundation for our study. Then, based on the aforementioned observations, we set our goals as weakening the role of high-frequency bands while enhancing low-frequency bands, which are achieved by the proposed two main components: HFS and LFR, respectively. As shown in Figure 3, our framework takes several augmented views from an image $\boldsymbol{x}$ as input. These views are generated by performing HFS with different transformations. Subsequently, we construct a composite feature space by feeding low-frequency bands directly into the embedding space of deep layers, called LFR. More details will be described in the following subsections.

### Discrete Wavelet Transformation

In contrast to other frequency analysis tools such as Fourier transform, Discrete Wavelet Transform (DWT) (Mallat 1989) can capture frequency features combined with high-precision spatial information, thus making it a more effective way for vision tasks. Furthermore, Bae *et al.* (Bae, Yoo, and Chul Ye 2017) show that wavelet transformations can provide topologically simpler data flow patterns, thus facilitating more efficient pattern recognition. Therefore, we employ DWT with the efficient Haar wavelet filter (Haar 1911) by default to obtain different frequency bands.

Figure 4 illustrates 2D-DWT with Haar kernels. Suppose $\boldsymbol{f}_{\mathrm{L}}$ and $\boldsymbol{f}_{\mathrm{H}}$ are the low-pass and high-pass filters of a standard 1D wavelet decomposition. For a 2D-image $\boldsymbol{x} \in \mathbb{R}^{H \times W}$, row-wise and column-wise 1D-DWT are conducted, which are defined as 2D transform. The corresponding 2D filters are denoted as $\{\boldsymbol{f}_{\mathrm{LL}}, \boldsymbol{f}_{\mathrm{LH}}, \boldsymbol{f}_{\mathrm{HL}}, \boldsymbol{f}_{\mathrm{HH}}\}$. Taking Haar wavelet as an example, the 2D low pass filter $\boldsymbol{f}_{\mathrm{LL}}$ is formulated as:

$$\boldsymbol{f}_{\mathrm{LL}} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}. \tag{1}$$

The high-pass filters $\boldsymbol{f}_{\mathrm{LH}}$, $\boldsymbol{f}_{\mathrm{HL}}$, and $\boldsymbol{f}_{\mathrm{HH}}$ are defined as:

$$\begin{aligned} \boldsymbol{f}_{\mathrm{LH}} &= \frac{1}{2}\begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}, \boldsymbol{f}_{\mathrm{HL}} = \frac{1}{2}\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \\ \boldsymbol{f}_{\mathrm{HH}} &= \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \end{aligned} \tag{2}$$

For DWT decomposition, four filters are used to convolve with input image $\boldsymbol{x}$ to obtain four subbands $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, and $\boldsymbol{x}_4$. Specifically, the low-frequency subbands $\boldsymbol{x}_1$ is defined as $(\boldsymbol{f}_{LL} \otimes \boldsymbol{x}) \downarrow_2$. The $(i, j)$-th value of $\boldsymbol{x}_1$ after the 2D Haar transform can be calculated as:

$$\begin{aligned} \boldsymbol{x}_1(i, j) = &\frac{1}{2}\boldsymbol{x}(2i-1, 2j-1) + \frac{1}{2}\boldsymbol{x}(2i-1, 2j) \\ &+ \frac{1}{2}\boldsymbol{x}(2i, 2j-1) + \frac{1}{2}\boldsymbol{x}(2i, 2j). \end{aligned} \tag{3}$$

The remaining three high-frequency subbands $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, and $\boldsymbol{x}_4$ can also be defined in an analogous manner. The wavelet decomposition can be regarded as a special downsampling operation with four decoupled channels. Different from irreversible pooling operations, the original image $\boldsymbol{x}$ can be accurately reconstructed by the Inverse DWT (IDWT):

$$\boldsymbol{x} = \mathrm{IDWT}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4). \tag{4}$$

This means that the four frequency bands of DWT could cover both the global structure and details of the image with negligible information loss. As shown in Figure 4, the low-frequency subbands $\boldsymbol{x}_1$ contains global structure information, which directly determines the intuitive discrimination of objects. The recognition process of humans also depended primarily on these low-frequency features, as demonstrated in Figure 1. The high-frequency subbands $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, and $\boldsymbol{x}_4$ contain horizontal, vertical and diagonal texture details, respectively, which are essential for fine-grained recognition.

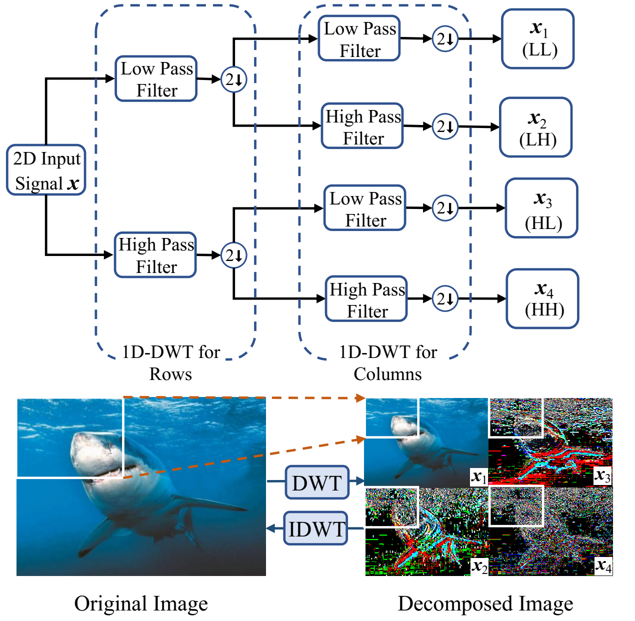

Figure 4: Illustration of Discrete Wavelet Transform (DWT) with Haar filter. The image is downsampled into four frequency subbands via wavelet decomposition.

## High-Frequency Shuffling

As stated above, the robustness of the human visual system primarily benefits from global structure information rather than the unintuitive high-frequency edges (Luo et al. 2022; Van den Branden Lambrecht and Kunt 1998). In contrast, CNNs tend to focus excessively on high-frequency information (Wang et al. 2020; Yin et al. 2019), which is considered the reason for their overconfident predictions. Inspired by the powerful generalization ability of humans, we argue that reducing excessive dependence on high-frequency bands and enhancing the ability to capture low-frequency bands are potential routes for both closed-set classification and open-set recognition tasks.

Specifically, we introduce the High-Frequency Shuffle (HFS) strategy. For a training sample $(\boldsymbol{x}, y)$, the main principle of HFS is to shuffle the high-frequency subbands within a reasonable range while leaving both the low-frequency subbands $\boldsymbol{x}_1$ and the label $y$ unchanged. In order to generate rich high-frequency material from $\boldsymbol{x}$, we establish a transformation set $\mathcal{T} = \{t_1, t_2, \dots t_K\}$, consisting of $K$ image transformation strategies. For each epoch, an element $t_i$ is randomly selected from $\mathcal{T}$ to obtain the transformed image $\hat{\boldsymbol{x}} = t_i(\boldsymbol{x})$, which is used to replace the high-frequency subbands of $\boldsymbol{x}$. Further, the two images are decomposed by DWT, and we get two sets of frequency bands $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4\}$ and $\{\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \hat{\boldsymbol{x}}_3, \hat{\boldsymbol{x}}_4\}$, respectively. With IDWT, HFS performs reconstruction from the wavelet domain to the spatial domain, denoted as:

$$\boldsymbol{x}^* = \mathrm{IDWT}(\boldsymbol{x}_1, \hat{\boldsymbol{x}}_2, \hat{\boldsymbol{x}}_3, \hat{\boldsymbol{x}}_4). \tag{5}$$

Since $\boldsymbol{x}^*$ is an augmented counterpart of $\boldsymbol{x}$, the label of $\boldsymbol{x}^*$ is naturally the same as that of the original $\boldsymbol{x}$. Then we train the network using the cross-entropy loss of samples $(\boldsymbol{x}^*, y)$ so that it learns robust high-frequency information. The HFS process is shown on the left side of Figure 3 with two transformations for simplicity. In the inference stage, we use the original test images directly without HFS.

By performing high-frequency perturbations, we encourage the network to learn low-frequency invariance, which extends the discriminative capacity for known and unknown classes. It is worth noting that HFS generates a bank of diverse samples based solely on the original samples without introducing additional hyperparameters to be tuned or parameters to be trained. In addition, HFS can be used as a novel data augmentation strategy in other recognition tasks to reduce the inductive bias of existing DNNs.

## Low-Frequency Residual Learning

To formulate the low-frequency residual learning throughout the network, we first define a typical deep convolutional neural network $\mathcal{F}$. Without losing generality, we presume that $\mathcal{F}$ is a visual recognition network consisting of $N$ convolutional stages, a fully connected layer and a softmax layer. Each of the convolutional stage consists of several convolutional layers, a non-linear activation layer, and a batch normalization layer (Ioffe and Szegedy 2015). Formally, the function $\mathcal{F}$ is denoted as:

$$\mathcal{F} = S \circ L \circ B_N \circ \cdots \circ B_1, \tag{6}$$

where '$\circ$' denotes the function composition operation. $S$ represents the softmax function, $L$ is the linear function, and $B_i$ denotes the function of the $i$-th convolutional stage, after which the feature maps are downsampled.

The proposed low-frequency residual learning network consists of the function $\mathcal{F}$ mentioned above and wavelet transform modules corresponding to each convolutional stage, as illustrated in Figure 3. Taking the $i$-th wavelet transform module as an example, it decomposes the input feature map of $B_i$ and obtains the low-frequency subbands denoted as $\mathrm{DWT}_{\mathrm{low}}(\boldsymbol{F}_i)$, where $\boldsymbol{F}_i$ is the input feature of $B_i$. Then we design a skip-connection protocol to feed low-frequency subbands from shallow layers to deeper ones. Formally, the skip connection is denoted as:

$$\boldsymbol{F}_{i+1} = B_i(\boldsymbol{F}_i; \boldsymbol{W}_i) + \mathrm{DWT}_{\mathrm{low}}(\boldsymbol{F}_i), \tag{7}$$

where Haar wavelet transformation is adopted as the default filter. The dimensions of $B_i(\boldsymbol{F}_i; \boldsymbol{W}_i)$ and $\mathrm{DWT}_{\mathrm{low}}(\boldsymbol{F}_i)$ are equal, so element-wise addition is conducted channel by channel. In this way, the function $B_i(\boldsymbol{F}_i; \boldsymbol{W}_i)$ is restructured to learn the low-frequency residual representations instead of fitting the original complex mapping. Based on Eq. 7, we perform residual connections for all convolutional stages (from $B_2$ to $B_{N-1}$) in the network without introducing additional parameters. Theoretically, if a convolutional stage can fit the low-frequency features, it can asymptotically fit the residuals of the low-frequency subbands. In contrast, training the latter will be easier because the residual connection facilitates identity mapping (He et al. 2016), which results in more sparse representations.

| Methods | MNIST | SVHN | CIFAR10 | CIFAR+10 | CIFAR+50 | Tiny-ImageNet |
|---|---|---|---|---|---|---|
| Softmax | 97.8 | 88.6 | 67.7 | 81.6 | 80.5 | 57.7 |
| OpenHybrid (Zhang et al. 2020) | 99.5 | 94.7 | 95.0 | 96.2 | 95.5 | 79.3 |
| PROSER (Zhou, Ye, and Zhan 2021) | - | 94.3 | 89.1 | 96.0 | 95.3 | 69.3 |
| ARPL (Chen et al. 2021b) | 99.6 | 96.3 | 90.1 | 96.5 | 94.3 | 76.2 |
| ARPL+CS (Chen et al. 2021b) | **99.7** | 96.7 | 91.0 | 97.1 | 95.1 | 78.2 |
| DIAS (Moon et al. 2022) | 99.2 | 94.3 | 85.0 | 92.0 | 91.6 | 73.1 |
| BCR (Cho and Choo 2022) | - | 95.6 | 94.8 | 96.1 | 95.7 | 78.5 |
| ODL (Liu et al. 2022) | 99.5 | 94.3 | 85.7 | 89.1 | 88.3 | 76.4 |
| KPF (Xia et al. 2023) | 99.6 | 96.3 | 89.9 | 96.6 | 94.3 | 76.0 |
| PMAL (Lu et al. 2022) | 99.5 | 96.3 | 94.6 | 96.0 | 94.3 | 81.8 |
| FreSH | 99.6 | **98.1** | **95.2** | **98.3** | **96.9** | **83.8** |

Table 1: The AUROC results of detecting known and unknown samples.

# Experiments

## Experimental Setup

**Datasets.** Due to the variety of real-world open-set scenarios, we utilize openness (Scheirer et al. 2012) to measure the complexity of open-set tasks, which is defined as:

$$\text{Openness } = 1 - \sqrt{\frac{2 \times |C_{\mathrm{TR}}|}{|C_{\mathrm{TR}}| + |C_{\mathrm{TE}}|}}, \qquad (8)$$

where $|C_{\mathrm{TR}}|$ represents the number of known categories in the training phase, and $|C_{\mathrm{TE}}|$ is the total number of known and unknown categories during testing. Following standard OSR protocols (Guo et al. 2021; Neal et al. 2018), we evaluate the proposed method on MNIST (LeCun et al. 2010), SVHN (Netzer et al. 2011), CIFAR10 (Krizhevsky, Hinton et al. 2009), CIFAR+10, CIFAR+50, Tiny-ImageNet (Le and Yang 2015). Their openness are from 13.39% to 62.86%.

**Implementation Details.** The establishment of open-set recognition scenarios relies on a split protocol for known and unknown classes. Since different splits often give rise to unfair comparisons, we follow the commonly used splits (Neal et al. 2018). For the backbone network, we follow the benchmark protocol (Neal et al. 2018) using VGG32. We use the Adam optimizer with a batch size of 128 for 600 epochs. The learning rate starts at 0.1 and decays by a factor of 0.1 every 120 epochs. All experiments are conducted with NVIDIA RTX 3090 GPU support.

## Benchmark Comparisons

**Unknown Detection.** As shown in Table 1, we use AUROC to measure the ability to detect unknown classes that are invisible during training. Following (Neal et al. 2018), we averaged the results over five randomized trials. We do not list some previous works because they report lower results than ARPL (Chen et al. 2021b). We achieve state-of-the-art performance on five datasets except 0.1% lower than ARPL on MNIST. However, the parameter number of ARPL expands tenfold more than that of ours due to additional auxiliary networks. Moreover, remarkable improvements are achieved on large-scale Tiny-ImageNet over both generative-based and discriminative-based methods. These methods do not pay attention to the inductive bias in the original spatial domain. Instead, we learn discriminative features through joint optimization in both frequency and spatial domains, which improves the reliability of OSR models.

| Method | SVHN | CIF10 | CIF+10 | CIF+50 | TINY |
|---|---|---|---|---|---|
| Softmax | 96.6 | 93.4 | 94.7 | 94.7 | 73.1 |
| CPN † | 96.7 | 92.9 | 94.8 | 95 | 81.4 |
| RPL † | 95.3 | 94.3 | 94.6 | 94.7 | 81.3 |
| ARPL † | 94.3 | 87.9 | 94.7 | 92.9 | 65.9 |
| PROSER | 96.5 | 92.6 | - | - | 52.1 |
| DIAS | 97.0 | 94.7 | 96.4 | 96.4 | 70.0 |
| ODL | 96.5 | 92.8 | 94.7 | 94.7 | 73.1 |
| PMAL † | 96.5 | 96.3 | 96.4 | 96.9 | 84.4 |
| FreSH | **97.4** | **97.3** | **97.9** | **97.7** | **87.6** |

Table 2: A comparison of the closed-set accuracy. The results of methods marked by † are from PMAL (Lu et al. 2022). Other results are from the original paper.

**Closed-Set Accuracy.** Table 2 shows the closed-set accuracy comparison between our method and existing methods on five datasets. According to Table 2, the FreSH does not sacrifice the accuracy of known classes when detecting unknown classes. Moreover, it significantly improves the closed-set accuracy, providing state-of-the-art results on all datasets, especially with a gain of 3.1% on TinyImageNet. We attribute it to the fact that our frequency enhancement strategy helps the network pay attention to global structure, leading to more compact and discriminative representations for known classes. The consistent improvement in AUROC and Acc demonstrates that unknown detection and closed-set recognition are highly correlated, which is also in line with the observation in (Vaze et al. 2021).

**Open-Set Recognition.** To further measure the trade-off between unknown detection and known recognition at various confidence thresholds, we report the Open-Set Classification Rate (OSCR) (Dhamija, Günther, and Boult 2018) score for a comprehensive evaluation of OSR. Our experimental settings and major results shown in Table 3 are derived from (Chen et al. 2021b). The proposed FreSH consistently improves OSCR performance by a substantial margin,

| Method | MNIST | SVHN | CIFAR10 | CIFAR+10 | CIFAR+50 | Tiny-ImageNet |
|---|---|---|---|---|---|---|
| Softmax | 99.2 ± 0.1 | 92.8 ± 0.4 | 83.8 ± 1.5 | 90.9 ± 1.3 | 88.5 ± 0.7 | 60.8 ± 5.1 |
| GCPL (Yang et al. 2018) | 99.1 ± 0.2 | 93.4 ± 0.6 | 84.3 ± 1.7 | 91.0 ± 1.7 | 88.3 ± 1.1 | 59.3 ± 5.3 |
| RPL (Chen et al. 2020) | 99.4 ± 0.1 | 93.6 ± 0.5 | 85.2 ± 1.4 | 91.8 ± 1.2 | 89.6 ± 0.9 | 53.2 ± 4.6 |
| ARPL (Chen et al. 2021b) | 99.4 ± 0.1 | 94.0 ± 0.6 | 86.6 ± 1.4 | 93.5 ± 0.8 | 91.6 ± 0.4 | 62.3 ± 3.3 |
| ARPL+CS (Chen et al. 2021b) | **99.5 ± 0.1** | 94.3 ± 0.3 | 87.9 ± 1.5 | 94.7 ± 0.7 | 92.9 ± 0.3 | 65.9 ± 3.8 |
| AKPF (Xia et al. 2023) | 99.4 ± NR | 94.3 ± NR | 88.1 ± NR | 94.9 ± NR | 93.0 ± NR | 67.8 ± NR |
| ODL (Liu et al. 2022) | 99.4 ± 0.1 | 93.4 ± 0.7 | 84.8 ± 1.4 | 92.5 ± 1.0 | 89.8 ± 0.7 | 64.3 ± 3.2 |
| FreSH | 99.4 ± 0.1 | **96.3 ± 0.3** | **91.9 ± 1.2** | **96.1 ± 1.0** | **94.4 ± 0.6** | **77.9 ± 4.1** |

Table 3: The OSCR results of open-set recognition. Results are the average of five randomized trials. Results with 'NR' indicate that the standard deviation was not released by the original paper.

| Method | IMGN-C | IMGN-R | LSUN-C | LSUN-R |
|---|---|---|---|---|
| Softmax | 63.9 | 65.3 | 64.2 | 64.7 |
| Openmax | 66.0 | 68.4 | 65.7 | 66.8 |
| CROSR | 72.1 | 73.5 | 72.0 | 74.9 |
| C2AE | 83.7 | 82.6 | 78.3 | 80.1 |
| CGDL | 84.0 | 83.2 | 80.6 | 81.2 |
| GFROSR | 75.7 | 79.2 | 75.1 | 80.5 |
| RPL | 81.1 | 81.0 | 84.6 | 82.0 |
| PROSER | 84.9 | 82.4 | 86.7 | 85.6 |
| CVAE | 85.7 | 83.4 | 86.8 | 88.2 |
| BCR | 87.6 | 86.9 | 88.0 | 87.7 |
| ODL | 85.6 | 85.2 | 86.5 | 82.6 |
| FreSH | **92.0** | **90.1** | **91.2** | **92.6** |

Table 4: Macro F1-Score of open-set classification on CIFAR-10 with various unknown datasets added in the test phase. The results of other methods are from ODL (Liu et al. 2022) and C2AE (Oza and Patel 2019).

*e.g.*, we push forward 10.1% than SOTA method AKPF (Xia et al. 2023) on Tiny-ImageNet. Notably, the improvement on OSCR is more significant than that on AUROC and Acc, indicating that the proposed framework effectively limits open space risk while balancing it with empirical risk.

**Open-Set Classification.** Following (Perera et al. 2020; Yoshihashi et al. 2019), we conduct the open-set classification experiment to evaluate the robustness. Similar to the setting of out-of-distribution detection, $K$ classes in the original dataset are used for training, and the unknown classes from other datasets are regarded as the (K+1)-th open class. Specifically, 10 categories of CIFAR10 are known classes, and unknown samples are from ImageNet (Russakovsky et al. 2015) and LSUN (Yu et al. 2015). The test samples are resized and cropped to the same size as training samples, obtaining ImageNet-crop, ImageNet-resize, LSUN-crop and LSUN-resize datasets. The Macro F1-Score is calculated using the ResNet34 backbone, as shown in Table 4. The proposed method consistently outperforms existing methods by a large margin (4% on average), demonstrating its insensitivity to domain shifts. It is because the FreSH enhances the robustness of recognition that we manage to handle open-set classes from various domains and generalize well.

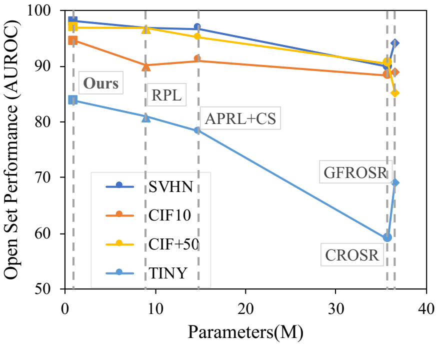

Figure 5: Comparison of the number of parameters.

**Number of Parameters.** Many existing open-set recognition methods use different types of backbones, such as VGG32 used by us and ARPL (Chen et al. 2021b), WRN-40-4 used by RPL (Chen et al. 2020), WRN-28-10 used by GFROSR (Perera et al. 2020) and DHRNet used by CROSR (Yoshihashi et al. 2019). Furthermore, some generative-based models (Chen et al. 2021b) introduce additional components, such as GANs and AEs. Therefore, training costs and the number of parameters from different methods vary significantly. For a fair comparison, we detail the line charts of AUROC and parameter number, as shown in Figure 5. For methods with an equal number of parameters, we choose the best one to conduct comparisons. The comparison highlights the superiority of the proposed FreSH framework — we achieve the best open-set performance with the smallest parameter amount. As our primary contributions lie in the special design of frequency bands and the reorganization of network connections, we can significantly improve the original lightweight backbone without auxiliary networks.

## Ablation Study and Further Analysis

**Ablation Study.** To investigate the contribution of high-frequency shuffling and low-frequency residual learning for our FreSH framework, we conduct an ablation study on both components. Specifically, we randomly select 15 classes as known classes from CIFAR100, and unknown classes are selected from the remaining 85 classes. The number of un-

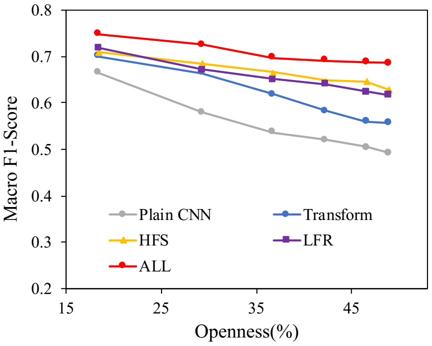

Figure 6: Ablation study against various openness.

| Methods | DFT-12 | DFT-16 | DCT-12 | DCT-16 | DWT |
|---|---|---|---|---|---|
| AUROC | <u>83.6</u> | 83.5 | 83.2 | 83.0 | **83.8** |
| ACC | <u>87.5</u> | 87.4 | 87.1 | 86.7 | **87.6** |

Table 5: Compare DWT with other frequency transformations from different high- and low- frequency thresholds.

known classes increases from 15 to 85, resulting in openness ranging from 18.4% to 48.9%. As shown in Figure 6, we employ Macro F1-Score to evaluate the performance. 'Plain CNN' is the baseline model with cross-entropy loss, which is trained in the same way as PROSER (Zhou, Ye, and Zhan 2021). 'HFS' and 'LFR' represent the addition of high-frequency shuffling or low-frequency residual connections to the baseline, and 'ALL' uses both strategies above. 'Transform' adds data augmentations used in (Hendrycks et al. 2019) to the baseline, considering that HFS introduces image transformation to generate high-frequency material. Overall, our framework benefits from both components and removing either of them will lead to a decline in recognition performance as both of them calibrate the biased perception.

**Compare with Other Frequency Transformations**. We compare DWT with other commonly used frequency transformations including Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT), which is a special form of DFT. The thresholds for separating high and low frequencies are crucial for DFT and DCT. After carefully searching, we ultimately chose 12 and 16 as the optimal thresholds, as shown in Table 5. The success of DWT could be attributed to its **1)** stable subbands decomposition without search of hyperparameter, as denoted in Eq (3); **2)** preservation of both frequency and spatial domain features to promote comprehensive representations, as the subbands generated by classical Fourier transform cannot retain the spatial component.

**Justification for HFS**. The High-frequency shuffling strategy aims to decrease high-frequency (HF) sensitivity by creating diverse HF views. An array of alternatives to HFS, such as employing Gaussian blur, directly deleting HF, replacing HF with Gaussian noise, or other images' HF are compared in Table 6. These strategies damage the HF information to some extent and fail to reserve complete textures. In contrast, HFS does not discard HF information, and it keeps and enriches HF via image transformation. The excellent results in Table 6 demonstrate the representation capabilities of HFS. Moreover, we integrate HFS into existing fine-grained recognition methods (Chou, Lin, and Kao 2022; He et al. 2021), as shown in Table 7, which indicates its effectiveness on higher-resolution CUB dataset. Notably, the experiments utilize ViT-B-16 and Swin-T, demonstrating the generality of HFS on transformer backbones.

| Methods | Blur | Low | H-noise | H-other | HFS |
|---|---|---|---|---|---|
| CIF+50 | 91.8 | 83.7 | 90.6 | <u>92.1</u> | **94.4** |
| TINY | 73.1 | 64.7 | 68.3 | <u>73.8</u> | **77.9** |

Table 6: OSCR results of our HFS and its alternatives.

| Swin-T | | ViT-B-16 | |
|---|---|---|---|
| FGVC-PIM | + HFS | TransFG | + HFS |
| 92.15 | **92.61** | 90.73 | **91.35** |

Table 7: Apply HFS to fine-grained recognition methods.

| Methods | CIFAR10 | | CIFAR+50 | |
|---|---|---|---|---|
| | AUROC | Acc | AUROC | Acc |
| VSR | 93.8 | 97.7 | 94.5 | **98.0** |
| LFR(Ours) | **95.2** | **97.9** | **96.9** | 97.7 |

Table 8: Comparisons of vanilla residual learning and LFR.

**Compare LFR with Vanilla Spatial Residual (VSR) Learning.** The proposed Low-Frequency Residual (LFR) aims to enhance low-frequency global structure. So it establishes connections from shallow to deep layers across different convolutional stages. For a fair comparison, we add vanilla spatial residual connections (He et al. 2016) at the same location as the LFR. As shown in Table 8, LFR improves unknown detection performance while maintaining closed-set accuracy. The robust and accurate performance of LFR can be attributed to the focus on global features conveyed by low-frequency subbands.

## Conclusion

In this paper, we provide a seminal insight into frequency inductive bias in the original spatial domain. We propose the Frequency Shuffling and Enhancement (FreSH) framework, consisting of High-Frequency Shuffling (HFS) to alleviate over-perception of subtle texture and Low-Frequency Residual (LFR) learning to enhance global structure perception. As a result, decoupled frequency subbands improve the robustness and accuracy without introducing extra parameters or complex optimization. Both HFS and LFR paradigms can be flexibly incorporated into existing frameworks. In the future, we will delve into generalized frequency representations for more challenging visual tasks.

## Acknowledgments

This work is supported in part by the National Natural Science Foundation of China Under Grants No.62176253 and No.U20B2066.

## References

Bae, W.; Yoo, J.; and Chul Ye, J. 2017. Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. In *CVPRW*, 145–153.

Bendale, A.; and Boult, T. E. 2016. Towards open set deep networks. In *CVPR*, 1563–1572.

Chen, G.; Peng, P.; Ma, L.; Li, J.; Du, L.; and Tian, Y. 2021a. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *ICCV*, 458–467.

Chen, G.; Peng, P.; Wang, X.; and Tian, Y. 2021b. Adversarial reciprocal points learning for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8065–8081.

Chen, G.; Qiao, L.; Shi, Y.; Peng, P.; Li, J.; Huang, T.; Pu, S.; and Tian, Y. 2020. Learning open set network with discriminative reciprocal points. In *ECCV*, 507–522. Springer.

Cho, W.; and Choo, J. 2022. Towards Accurate Open-Set Recognition via Background-Class Regularization. In *ECCV*, 658–674. Springer.

Chou, P.-Y.; Lin, C.-H.; and Kao, W.-C. 2022. A novel plug-in module for fine-grained visual classification. *arXiv preprint arXiv:2202.03822*.

Dhamija, A. R.; Günther, M.; and Boult, T. 2018. Reducing network agnostophobia. *NeurIPS*, 31.

Fritsche, M.; Gu, S.; and Timofte, R. 2019. Frequency separation for real-world super-resolution. In *ICCVW*, 3599–3608. IEEE.

Ge, Z.; Demyanov, S.; Chen, Z.; and Garnavi, R. 2017. Generative openmax for multi-class open set classification. *arXiv preprint arXiv:1707.07418*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, volume 27, 2672–2680.

Guo, C.; Frank, J. S.; and Weinberger, K. Q. 2018. Low frequency adversarial perturbation. *arXiv preprint arXiv:1809.08758*.

Guo, Y.; Camporese, G.; Yang, W.; Sperduti, A.; and Ballan, L. 2021. Conditional variational capsule network for open set recognition. In *ICCV*, 103–111.

Haar, A. 1911. Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen*, 71(1): 38–53.

He, J.; Chen, J.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C.; and Yuille, A. 2021. Transfg: A transformer architecture for fine-grained recognition. arXiv 2021. *arXiv preprint arXiv:2103.07976*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2019. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*.

Huang, H.; Wang, Y.; Hu, Q.; and Cheng, M.-M. 2022. Class-Specific Semantic Reconstruction for Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 448–456. pmlr.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kodama, Y.; Wang, Y.; Kawakami, R.; and Naemura, T. 2021. Open-set recognition with supervised contrastive learning. In *MVA*, 1–5. IEEE.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84–90.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

LeCun, Y.; Cortes, C.; Burges, C.; et al. 2010. MNIST handwritten digit database.

Li, J.; Xie, H.; Li, J.; Wang, Z.; and Zhang, Y. 2021. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *CVPR*, 6458–6467.

Li, J.; You, S.; and Robles-Kelly, A. 2018. A frequency domain neural network for fast image super-resolution. In *IJCNN*, 1–8. IEEE.

Liu, Z.-g.; Fu, Y.-m.; Pan, Q.; and Zhang, Z.-w. 2022. Orientational Distribution Learning with Hierarchical Spatial Attention for Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Lu, J.; Xu, Y.; Li, H.; Cheng, Z.; and Niu, Y. 2022. Pmal: Open set recognition via robust prototype mining. In *AAAI*, volume 36, 1872–1880.

Luo, C.; Lin, Q.; Xie, W.; Wu, B.; Xie, J.; and Shen, L. 2022. Frequency-driven imperceptible adversarial attack on semantic similarity. In *CVPR*, 15315–15324.

Mallat, S. G. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7): 674–693.

Moon, W.; Park, J.; Seong, H. S.; Cho, C.-H.; and Heo, J.-P. 2022. Difficulty-Aware Simulator for Open Set Recognition. In *ECCV*, 365–381. Springer.

Neal, L.; Olson, M.; Fern, X.; Wong, W.-K.; and Li, F. 2018. Open set learning with counterfactual images. In *ECCV*, 613–628.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Oza, P.; and Patel, V. M. 2019. C2ae: Class conditioned auto-encoder for open-set recognition. In *CVPR*, 2307–2316.

Perera, P.; Morariu, V. I.; Jain, R.; Manjunatha, V.; Wigington, C.; Ordonez, V.; and Patel, V. M. 2020. Generative-discriminative feature representations for open-set recognition. In *CVPR*, 11814–11823.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115: 211–252.

Scheirer, W. J.; de Rezende Rocha, A.; Sapkota, A.; and Boult, T. E. 2012. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772.

Sharma, Y.; Ding, G. W.; and Brubaker, M. 2019. On the effectiveness of low frequency perturbations. *arXiv preprint arXiv:1903.00073*.

Sun, X.; Yang, Z.; Zhang, C.; Ling, K.-V.; and Peng, G. 2020. Conditional gaussian distribution learning for open set recognition. In *CVPR*, 13480–13489.

Van den Branden Lambrecht, C. J.; and Kunt, M. 1998. Characterization of human visual sensitivity for video imaging applications. *Signal Processing*, 67(3): 255–269.

Vaze, S.; Han, K.; Vedaldi, A.; and Zisserman, A. 2021. Open-set recognition: A good closed-set classifier is all you need. *arXiv preprint arXiv:2110.06207*.

Wang, H.; Wu, X.; Huang, Z.; and Xing, E. P. 2020. High-frequency component helps explain the generalization of convolutional neural networks. In *CVPR*, 8684–8694.

Wang, Y.; Yu, K.; Chen, C.; Hu, X.; and Peng, S. 2023. Dynamic Graph Learning With Content-Guided Spatial-Frequency Relation Reasoning for Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7278–7287.

Xia, Z.; Wang, P.; Dong, G.; and Liu, H. 2023. Adversarial Kinetic Prototype Framework for Open Set Recognition. *IEEE Transactions on Neural Networks and Learning Systems*.

Xiao, M.; Zheng, S.; Liu, C.; Wang, Y.; He, D.; Ke, G.; Bian, J.; Lin, Z.; and Liu, T.-Y. 2020. Invertible image rescaling. In *ECCV*, 126–144. Springer.

Yang, H.-M.; Zhang, X.-Y.; Yin, F.; and Liu, C.-L. 2018. Robust classification with convolutional prototype learning. In *CVPR*, 3474–3482.

Yang, H.-M.; Zhang, X.-Y.; Yin, F.; Yang, Q.; and Liu, C.-L. 2020. Convolutional prototype network for open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5): 2358–2370.

Yao, T.; Pan, Y.; Li, Y.; Ngo, C.-W.; and Mei, T. 2022. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *ECCV*, 328–345. Springer.

Yin, D.; Gontijo Lopes, R.; Shlens, J.; Cubuk, E. D.; and Gilmer, J. 2019. A fourier perspective on model robustness in computer vision. *NeurIPS*, 32.

Yoshihashi, R.; Shao, W.; Kawakami, R.; You, S.; Iida, M.; and Naemura, T. 2019. Classification-reconstruction learning for open-set recognition. In *CVPR*, 4016–4025.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zhang, H.; Li, A.; Guo, J.; and Guo, Y. 2020. Hybrid models for open set recognition. In *ECCV*, 102–117. Springer.

Zhou, D.-W.; Ye, H.-J.; and Zhan, D.-C. 2021. Learning placeholders for open-set recognition. In *CVPR*, 4401–4410.