Wavelet J-Net: A Frequency Perspective on Convolutional Neural Networks

Linfeng Zhang[1] Xiaoman Zhang[3] Chenglong Bao*[2] Kaisheng Ma*[1]
Institute for Interdisciplinary Information Science, Tsinghua University [1]
YAU Mathematical Sciences Center, Tsinghua University[2]
Institute for Interdisciplinary Information Core Technology[3]
zhang-lf19@mails.tsinghua.edu.cn, xiaomanzhang1998@outlook.com, {clbao, kaisheng}@mail.tsinghua.edu.cn

***Abstract*—It is well acknowledged in image processing domain that the information can be decomposed into different frequency parts and each part has its own merits. However, existing neural networks always ignore the distinctions and straightforwardly feed all the information into neural networks together, treating them equally. In this paper, we propose a novel neural networks framework named J-Net that decomposes images into different frequency bands and then processes them sequentially. Concretely, the images have been decomposed by wavelet transformation and then the wavelet coefficients are fed into neural networks gradually in different depth according to their decomposition levels. An attention module is utilized to facilitate the fusion of neural network features and injected information, yielding significant performance gain. Furthermore, we show how does the information with different frequency impact the accuracy of neural networks. Experiments show that 5.91%, 5.32% and 2.00% accuracy improvements on Caltech 101, Caltech256 and ImageNet, respectively.**

I. INTRODUCTION

Deep convolutional neural networks (CNNs) have been widely deployed in various applications, such as images classification [1]–[3], object detection [4]–[6] and segmentation [7], [8]. Despite their impressive numerical success, obstacles remain on the way towards an efficient and precise human-like vision system. Firstly, CNNs have "black-box" property which is difficult to explain their predictions. Although great efforts have been made for the interpretation of CNNs, demystifying CNNs is still an open problem. Secondly, the internal mechanism of features seems not consistent with human vision systems. CNNs recognize an image firstly by local features and then by global features. In contrast, the human visual system always obtains the global information of images instantly and then observes details.

In this paper, we have made an effective attempt to solve the aforementioned problems from the perspective of frequency. Different from existing deep learning methods where the whole image is inputted to neural networks together, we first decompose the image into different frequency by discrete wavelet transformation and then inject them into neural networks gradually. We further propose the attention concatenation to facilitate the information fusion between feature maps in neural networks and the injected high frequency information to improve the efficiency of J-Net. By restricting the inputs into different bands, it may simplify the analysis and help us understand the internal mechanisms of networks more clearly. Figure 4 shows the architecture of J-Net based on ResNet18 models. Note that the proposed J-Net is not a specific neural network but a general technique that can be applied to all kinds of neural networks.

* Corresponding authors.
This work was partially supported by IIISCT (Institute for interdisciplinary Information Core Technology), National Natural Sciences Foundation of China (No.31970972 and 11901338), and Tsinghua University Initiative Scientific Research Program.

A. Motivation

Although frequency information plays an important role in signal processing, it is not well-studied in the domain of deep learning. We have conducted two experiments to demonstrate the great impact from frequency information to the image classification by convolutional neural networks. Note that the low resolution images indicate images whose high frequency is lost.

Figure 2 shows the accuracy of several binary classification tasks on images with different resolutions. It can be observed that: (i) with more high frequency lost in small resolution images, the accuracy of all tasks inclines to drop; (ii) with the same resolution of images, the accuracy differs for different tasks, e.g. there is more decline on accuracy in "dogs vs. cats" task than that in "dogs vs. wardrobes" task. This phenomenon might be explained by the following reason: compared with wardrobes, the low frequency information of cats, such as shape, is more similar to dogs. As a result, high frequency information is indispensable for distinguishing dogs and cats. Therefore, the classification accuracy depends on the resolution of the images of different categories.

Figure 1 shows the classification accuracy Caltech 101 and Caltech256 with varying resolutions. It is shown that the classification accuracy drops as the resolution decreases in both two datasets. However, the accuracy on Caltech256 shows a more significant degradation tendency than accuracy on Caltech 101, indicating that images which have more hardship on recognition requires a larger resolution, which contains more high frequency information.

These results motivate us to process different frequency information by different methods rather than inputting them as a whole. Besides, in recent years, F-principle, a significant breakthrough has been attained on the study of frequency

978-0-7381-3366-9/21/$31.00 ©2021 IEEE

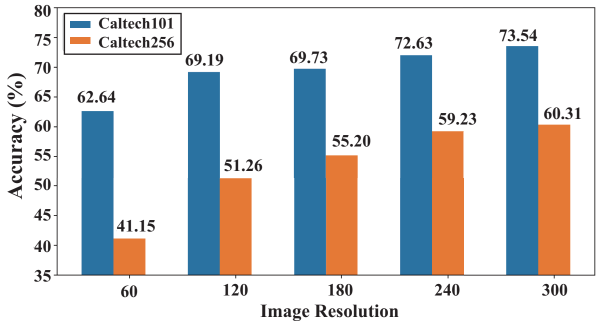

Fig. 1. Experiments on various datasets with different image resolution.

in neural networks training period with Fourier analysis [9]: Neural networks fit target functions from low frequency to high frequency gradually. Besides, the energy landscape of low resolution images seems to smooth the original objective function which increases the possibilities of attaining the stationary point with good generalization ability. Although the frequency information is important for image analysis and processing, it is still not clear how to use it with high efficiency and effectiveness.

Compared with existing neural network architectures, the proposed architecture has several differences: (i) All the information is reserved for processing and on the contrary, the resized images lost some high frequency which reduces the classification accuracy as shown in Figure 2; (ii) The network is relatively easy to explain and debug as the frequency has been input gradually; (iii) By processing the band information independently, this decomposition provides the parallel implementation and shows the acceleration effects. In this paper, we show that processing the frequency gradually can improve the classification accuracy with a large margin.

It is noted that the architecture of J-Net is motivated by U-Net [10], which is the state-of-the-art method in many image processing tasks. As shown in Figure 3, J-Net can be viewed as the right half part of U-Net, targeting at classification instead of segmentation. The U-Net for image segmentation can be seen as a $\mathbb{R}^N \rightarrow \mathbb{R}^N$ mapping for processing pixel-level tasks, but the classification in J-Net is a $\mathbb{R}^N \rightarrow \mathbb{R}^1$ function which is relatively easier to be addressed. J-Net replaces

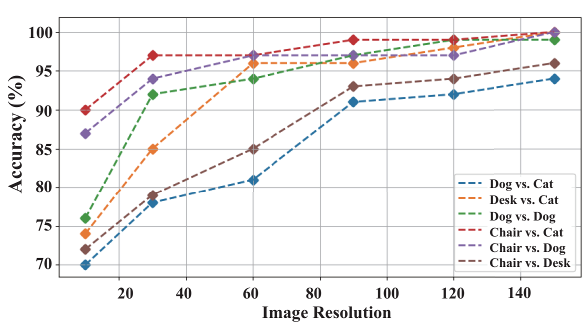

Fig. 2. Experiments on various binary classification tasks with different image resolution. The images of these tasks are samppled from ImageNet.

the convolution layers in the left part with discrete wavelet transformation, which has less parameters and computation. Another splendid advantage of J-Net lies in its interpretability. The parameters of convolution layers in the left part of U-Net are learned from data, which is generally acknowledged as black boxes and it is difficult to understand the internal mechanism of neural networks.

In contrast, the discrete wavelet transformation decomposes the image into different bands and each band is well understood in the time and frequency domain. Therefore, the band information processed by the neural networks is controllable when the training images have high resolutions. Based on this property, an ablation study of frequency bands has been conducted in the discussion section.

In summary, the main contributions of this paper are listed as follows:

- A novel frequency inspired neural network architecture is proposed to utilize the information at different frequency levels with various methods. The proposed J-Net firstly processes the low frequency information, and then the high frequency information. This procedure is more analogous to human perception. To the best of our knowledge, this is the first attempt by controlling the frequency inputs for classification which yields significant accuracy gains.
- Extensive experiments have been conducted to prove the generality and effectiveness of J-Net in three state-of-the-art neural networks: SENet, ResNeXt, ResNet, three benchmark datasets: ImageNet, Caltech101, Caltech256, and four wavelet basis: Haar, ReverseBior, Biorthogonal, Daubechies. On average, 5.91%, 5.32% and 2.00% accuracy gain can be observed in Caltech101, Caltech256 and ImageNet, respectively.

The rest paper is organized as follows. Section II introduces the related work of J-Net. Section III demonstrates the details of J-Net. Section IV shows the experiments results on different datasets and neural networks. Section V reveals the influence of frequency on neural networks. Finally, a conclusion is brought forth in section VI.

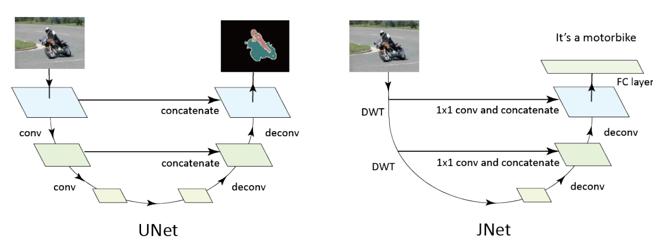

Fig. 3. Diagrams of U-Net and the proposed J-Net. Their main difference is that (i) U-Net targets at image segmentation while J-Net aims at image classification. (ii) Information with different resolutions and frequency is obtained by convolutional layers in U-Net yet discrete wavelet transformation in J-Net. (iii) The feature merging method is native concatenation. In contrast, the proposed J-Net merges features in neural networks and high frequency information with the proposed attention concatenation.

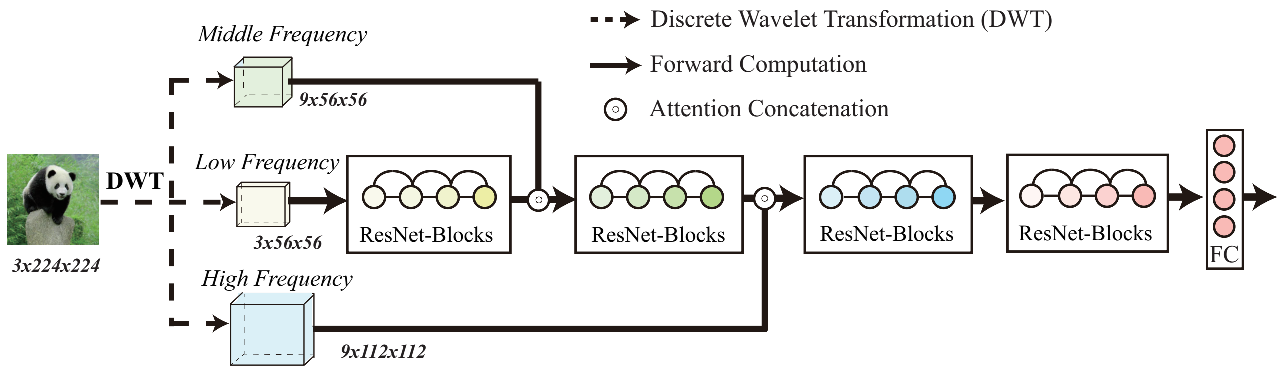

Fig. 4. The architecture of J-Net on ResNet backbone:(i) Information of three frequency groups from DWT are fed into neural networks from low frequency to high frequency. (ii) An attention concatenation module is proposed to facilitate the fusion of new frequency information and features in neural networks. (iii) Deconvolution and 1x1 convolution layers are introduced to adjust the size of features and high frequency information for merging.

II. Related work

1) Frequency Methods in Neural Networks.: With the prosperity of convolutional neural networks, fruitful works have been proposed to enhance the performance of neural networks via frequency methods. OctConv separates the low and high spatial frequency signal in images and then processes them in different groups of convolution layers, yielding consistently higher accuracy with less memory and computation in both training and inferent period [11]. Wang *et al.*, Chen *et al.* and Vogel *et al.* compressed the weights and activation of neural networks in frequency domain via discrete cosine transformation [12]–[14]. Targeting at promoting the convergence of neural networks, discrete Fourier transformation is utilized to allow models to directly learn their weights in frequency domain [15]. Moreover, Xu *et al.* studied the learning procedure of deep neural networks through the lens of frequency, discovering the F-principle in neural networks: DNN fits target functions in training procedure from low frequencies gradually to high frequencies [9].

2) Wavelet Methods in Neural Networks.: As one of the common methods in numerical analysis and functional analysis, wavelet transformation can capture both frequency and location information. Many works have explored its usage in deep neural networks. Chen *et al.* proposed WAE which decomposes the information in images via auto-encoder, accelerating the inference of convolutional neural networks [16]. Wavelet pooling is proposed to replace the traditional max and average pooling with a 2D discrete wavelet transformation, obtaining global information of images during downsampling computation [17]. Fujieda *et al.* applied discrete wavelet factorization on texture classification problem, yielding state-of-the-art results by multi-resolution analysis [18]. Liu *et al.* merged convolutional neural networks with wavelet transformation, which outperforms prior works on multiple restoration tasks [19]. Moreover, wavelet methods are also widely utilized in the super resolution domains. In wavelet SRNet [20], an embedding layer and a wavelet prediction layer are exploited to predict the coefficients of the wavelets of high resolution images by that of low resolution images.

3) Multi-resolution Analysis in Neural Networks.: Substantial research has been proposed to improve the performance of neural networks with muti-resolution analysis. The coarse information in high layers and the fine information in shallow layers were combined by addition or concatenation in FCN [21], PSPN [22], yielding dramatic performance gain on segmentation. In UNet [10], [23], [24], skipping connections between features with the same resolution from the downsampling to upsampling were designed to improve the effectiveness of images reconstruction. MSDenseNet [25] studies the influence of coarse and fine-grained features to neural networks. Moreover, multi-resolution analysis is also broadly utilized in objection detection such as SSD [4], FPN [8],SPPNet [26] where the fusion of information with various resolution leads to significant increment of recognition accuracy, especially on small objects.

III. The proposed J-Net

A. Wavelet transformation

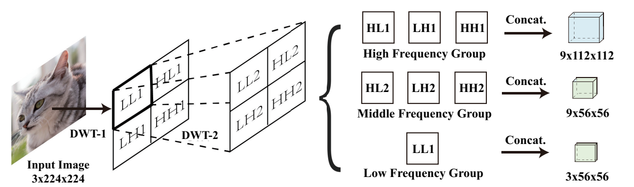

Fig. 5. Decomposition of images with multi-level wavelet discrete transformation: (i) The two-level discrete wavelet transformation (DWT-1 and DWT-2) is applied to the input images to decompose its frequency. (ii) The first discrete wavelet transformation is applied to the original image and the second wavelet discrete transformation is applied on the LL1 frequency of the first decomposition, which indicates the information with the lowest frequency of the original image.(iii) The decomposition results are divided and concatenated into three groups according to their frequency.

For a given function ψ, let $\chi(\psi)$ be the collection of the dilations and shift of ψ:

$$\chi(\psi) = \{\psi_{jk} = 2^{j/2}\psi(2^j x - k), j, k \in Z\}.$$

ψ is the orthonormal wavelet if $\chi(\psi)$ forms a basis in L_2 spaces. The wavelet transformation is good in time resolution of high frequencies, while for slowly varying functions, the frequency resolution is remarkable. This advantage allows us to process the frequency information gradually by using the wavelet transformation. The main wavelets are Haar wavelets, Daubechies wavelets, etc. Please refer to [27] for more details. In the discrete setting, it corresponds to the filtering operations for which one is the low pass filter and the other one is the high pass filter and satisfies the perfect reconstruction property.

In J-Net, a multi-level wavelet transformation is utilized to acquire information with diverse frequency in images. As depicted in Figure 5, we first apply a discrete wavelet transformation on input images, and continually decompose the low frequency part in the result of the first transformation. The obtained information is divided into three groups: The low frequency group, which is inputted into neural networks firstly, consists of the lowest frequency information (LL2 in Figure 5). The middle frequency group, which is injected into neural networks in the second phase, consists of the second transformation results except for the one with the lowest frequency (HL2, LH2, HH2). The left information (LH1, HL1, HH1) forms the highest frequency level, is injected into the neural networks as the last inputs. The R, G, B channels of input images are transformed individually and concatenated together eventually.

Although there are also other frequency decomposition methods such as discrete Fourier transformation (DFT) and discrete cosine transformation (DCT), experiments show that wavelet is a better method in J-Net, which is introduced in section V.

B. Neural networks architecture

1) Adjusting the first convolution layer with a large stride.: In many classical neural networks architectures, such as ResNet [2], AlexNet [1], VGG [28], a convolution layer and pooling layer with strides larger than 1 is attached at the beginning of the whole neural networks to decrease the width and height of input images. In J-Net, the stride of these layers should be adjusted to 1 because the width and height of input information have been already decreased by 4 times in discrete wavelet transformation.

2) Choosing the suitable injection position.: Another critical problem in J-Net is how to choose the suitable injection position for middle and high frequency information. Most existing neural networks can be divided into several stages according to their downsampling operation such as ResNet [2], DenseNet [29]. Inspired by U-Net where the skip connection is applied to the features before downsampling, We inject the high frequency information into neural networks before the 1_{st} and 2_{nd} downsampling of neural networks, which has been certified as an effective method in experiments.

3) Merge the features and the high frequency information.: When the high frequency information is injected into the neural networks, the features in neural networks and the high frequency information are merged. Unfortunately, there are several obstacles in the way of merging: (i) The high frequency information and features in neural networks have different sizes so they can not be merged straightforwardly. For example, as shown in Figure 6, the size of high frequency is 9x112x112 yet the size of feature map is 256x28x28. (ii) The high frequency information has much fewer channels than the features in neural networks. So the high frequency information will have very limited influence on the neural network if they are merged straightforwardly. (iii) It's hard for neural networks to directly merge the frequency information and representative features together.

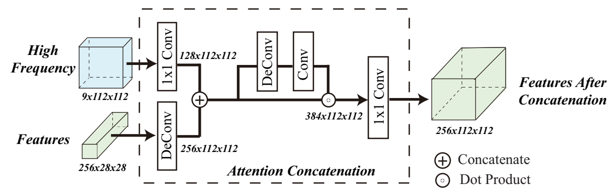

Fig. 6. The architecture of the proposed attention concatenation module. "DeConv" indicates deconvolution layers. The additional convolutional layers are introduced to adjust the size of features and high frequency information.

The best information merging ratio between features and high frequency information may vary from different images so a fixed merging scheme may not be the best solution.

To address these problems, we propose a novel feature merging method named attention concatenation. Concretely, a downsample-upsample attention module is utilized to learn 0-1 attention mask for both features and high frequency information. Different from the common attention modules which aim to enable neural networks to pay more attention to important pixels or channels, the proposed attention concatenation targets to learn the best information merging ratio for different input adaptively. The architecture of attention concatenation is shown in Figure 6. First, a deconvolution layer is applied to the features to adjust their width and heights. Second, a pointwise (1x1) convolution is applied to the high frequency information to improve the number of channels. Then, the features after deconvolution and high frequency information after pointwise (1x1) convolution are concatenated together and finally input into an attention module to facilitate their merging.

A series of experiments have been conducted to show the effectiveness of the proposed attention concatenation method. As shown in Table I:(i) In all of the three feature merging methods, the proposed J-ResNet outperforms the baseline by a large margin. (ii) On both Caltech101 and Caltech256 datasets, the proposed attention concatenation achieves higher accuracy than the other two feature merging methods.

IV. Experiments

In this section, we introduce the detailed experiment information and corresponding explanation.

A. Experiment settings

The proposed J-Net is evaluated on three datasets: ImageNet2012 [30], Caltech101 and Caltech256 [31] with various neural networks architectures: ResNet [2], SENet [32], and ResNeXt [3].

TABLE I
COMPARISON BETWEEN THE PROPOSED ATTENTION CONCATENATION (OUR APPROACH) AND THE OTHER TWO KINDS OF FEATURE MERGING METHODS ON J-RESNET18 AND CALTECH DATASETS.

Merging Strategy	Caltech101	Caltech256
Baseline	68.34	59.35
Add	71.21	61.39
Concat.	71.04	61.25
Ours	**72.84**	**63.53**

TABLE II
EXPERIMENTS RESULTS OF TOP 1 AND TOP 5 ACCURACY (%) ON THE VALIDATION SET OF IMAGENET. NEURAL NETWORKS WITH "J-" INDICATES TRAINED WITH THE PROPOSED J-NET.

Neural Networks	Top1 Accuracy	Top5 Accuracy
ResNet18	69.21	89.01
J-ResNet18	71.47	90.29
ResNet50	75.35	92.41
J-ResNet50	76.41	93.29
SENet18	70.19	89.76
J-SENet18	72.88	91.34

In the training period, common techniques including learning rate schedule, l_2 regularizer and data argumentation including horizon flipping and random cropping are utilized to improve model accuracy. The batch size of experiments on Caltech and ImageNet is 128 and 256, respectively. In all the reported experiments, the initialization of the learning rate is 0.1 and neural networks are trained by 90 epochs, with the learning rate divided by 10 at the 30_{th}, 60_{th}, 85_{th} epochs. The Haar wavelet basis is utilized in experiments of Table II and Table III, implemented with PyWavelet.

The Caltech101 and Caltech256 datasets are randomly divided into the training set and test set in the proportion of 8:2. And all the reported accuracy of ImageNet is evaluated on the validation set. In the following tables, the neural networks' names started with "J" indicate neural networks equipped with proposed J-Net. All the experiments are conducted with PyTorch, running on Tesla V100 and RTX 2080Ti GPU devices. Codes of J-Net will be released on github.

B. Results on ImageNet

As depicted in Table II, significant accuracy gain can be observed on various neural networks. On average, J-Net leads to 2.00% accuracy increment on Top1 accuracy, ranging from 2.69% on SENet18 as maximum to 1.06% on ResNet50 as a minimum, and 1.25% accuracy increment op Top5 accuracy, ranging from 1.58% on SENet18 as maximum to 0.88% on ResNet50 as a minimum.

C. Results on Caltech101 and Caltech256

Experiments on Caltech101 and Caltech256 are depicted in Table III. It can be observed that: (i) Significant and consistent accuracy increment can be observed on all the neural networks and both datasets. (ii) On Caltech101, an increment of 5.91% is attained on average, ranging from 9.66% on ResNet50 as a maximum and 1.68% on SENet18 as a minimum. (iii) On Caltech256, an increment of 5.32% is attained on average, ranging from 7.36% on ResNeXt as a maximum and 3.71% on SENet18 as a minimum. (iv) 6.68% accuracy gain is observed on 50 layers neural networks, which is larger than the 4.02% accuracy gain on 18 layers neural networks, indicating that deeper neural networks benefit more from the proposed J-Net.

TABLE III
EXPERIMENTS RESULTS OF ACCURACY (%) ON CALTECH101 AND CALTECH256. NEURAL NETWORKS WITH "J-" INDICATES TRAINED WITH THE PROPOSED J-NET.

Neural Networks	Caltech101	Caltech256
ResNet18	68.34	59.35
J-ResNet18	72.84	65.53
ResNet50	54.38	59.41
J-ResNet50	64.04	64.35
SENet18	73.64	58.27
J-SENet18	75.32	61.98
SENet50	67.90	60.41
J-SENet50	76.06	64.83
ResNeXt50	63.07	58.37
J-ResNeXt50	68.60	65.73

TABLE IV
ACCURACY (%) OF J-RESNET18 WITH FOUR KINDS OF WAVELET BASIS. NUMBERS IN BOLD ARE THE BEST.

Wavelet basis	Caltech101	Caltech256	SubImageNet
Baseline	68.34	58.95	76.89
Haar	**72.84**	**63.53**	81.45
RbioNr.Nd	72.43	63.04	80.16
Biorthgonal	71.45	62.98	**81.74**
Daubechies	72.23	63.29	80.77

As depicted in table III, The proposed J-Net brings additional parameters compared with the original models (For example, 1.49M in J-ResNet18). However, experiments results demonstrate that the benefits of J-Net is more than the bonus from parameters. Because there are very limited images in Caltech datasets (less than 100 images in most of the categories), the deep neural networks are very inclined to overfit the training set, leading to the phenomenon that the deeper neural networks with more parameters achieve equal and even lower accuracy than the shallow neural networks. For example, the accuracy of ResNet50 is 13.96% than ResNe18 in Caltech101. In contrast, the proposed J-Net achieves significantly higher accuracy than the baseline, indicating the proposed J-Net is more than simply increasing the depth of neural networks.

V. DISCUSSION

In this section, we discuss and analyze the performance of our proposed neural networks with more justifications.

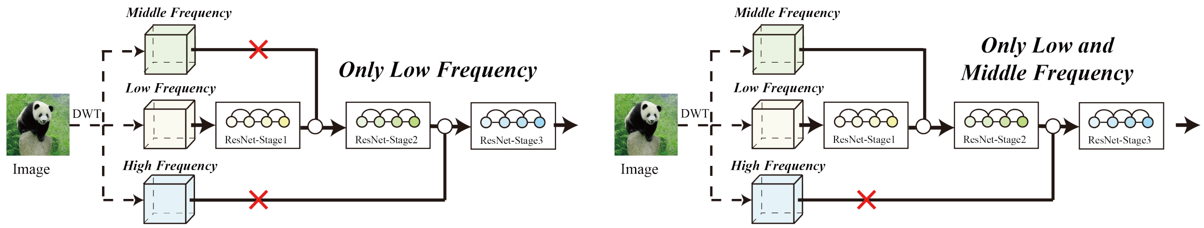

Fig. 7. Description for the ablation study on image frequency. (i)"Only Low Frequency Information" indicates only the low frequency information of images is input into the neural networks. (ii) "Low and Middle Frequency Information" indicates only the low frequency and middle frequency information of images is input into the neural networks.

TABLE V
EXPERIMENT RESULTS OF THE FREQUENCY ABLATION STUDY IN CALTECH101 DATASETS.

Neural Networks	Baseline	Low Frequency	Low+Mid Frequency	All Frequency
J-ResNet18	68.34	66.74	74.88	72.84
J-ResNet50	54.38	49.10	57.33	64.04
J-SENet18	73.64	71.60	76.87	75.32
J-SENet50	67.90	62.37	72.57	76.06
J-ResNeXt50	63.07	57.38	58.36	68.60

TABLE VI
EXPERIMENT RESULTS OF THE FREQUENCY ABLATION STUDY IN CALTECH256 DATASETS.

Neural Networks	Baseline	Low Frequency	Low+Mid Frequency	All Frequency
J-ResNet18	59.35	58.60	60.89	63.53
J-ResNet50	59.41	53.34	54.49	64.35
J-SENet18	58.27	58.75	61.69	61.98
J-SENet50	60.41	59.94	64.29	64.83
J-ResNeXt50	58.37	49.70	57.82	65.73

A. Generality of wavelet basis

Four different types of wavelet transformations are chosen for grouping the frequency information. Table IV shows the effectiveness of J-Net with four kinds of wavelet basis (Haar basis, ReverseBior basis, Biothogonal basis, Daubechies basis, Daubenchies basis) on three kinds of datasets. Note that Sub-ImageNet denotes a dataset that consists of 100 categories randomly chosen from ImageNet. It is observed that (i) On three kinds of datasets, J-Nets with all kinds of wavelet basis achieves significant accuracy gain: 3.90% on Caltech101, 4.26% on Caltech256, 4.14% on Sub-ImageNet. (ii) The accuracy range of J-Net with different wavelet basis is smaller than 1.17%, indicating that the choice of wavelet basis doesn't impact the accuracy of J-Net severely.

B. Ablation study of frequency

An ablation study on the influence of information with different frequency is introduced in this section by comparing the accuracy of models with different frequency information. In this study, we compare model accuracy in four different kinds of situation:(i) Baseline, indicating that models are trained by standard methods without the proposed J-Net. (ii) Low Frequency, indicating that models are trained by only the low frequency with J-Net. (iii) Low & Middle Frequency, indicating that models are trained by both low frequency and middle frequency information with J-Net. (iv) All frequency, indicating that models are trained with the whole frequency information with J-Net. By comparing the accuracy achieved by models with different frequency, the importance of frequency to neural networks can be observed. The second and the third situation are depicted in Figure 7.

Table V and VI show the results under three frequency settings, five kinds of neural networks and two datasets in detail. It is observed that: (i) *Low frequency is crucial, but not decisive.* Accuracy of J-Net with only low frequency information sometimes can achieve the corresponding baseline, such as SENet18 on Caltech256. However, due to the loss of middle and higher frequency, its average accuracy of two datasets is 4.03% and 3.10% lower than the average baseline, respectively. (ii) *Middle frequency information always leads to accuracy increment.* Compared with J-Net with only low frequency information, middle frequency information leads to 6.60% accuracy gain on Caltech101 and 3.77% accuracy gain on Caltech256 on average, indicating that middle information is also indispensable for precise prediction. (iii) *High frequency is beneficial in most situations.* Experiments show J-Net injected with high frequency achieves another 3.33% and 4.25% accuracy increment on Caltech101 and Caltech256, respectively.

However, the accuracy of J-ResNet18 and SENet18 on

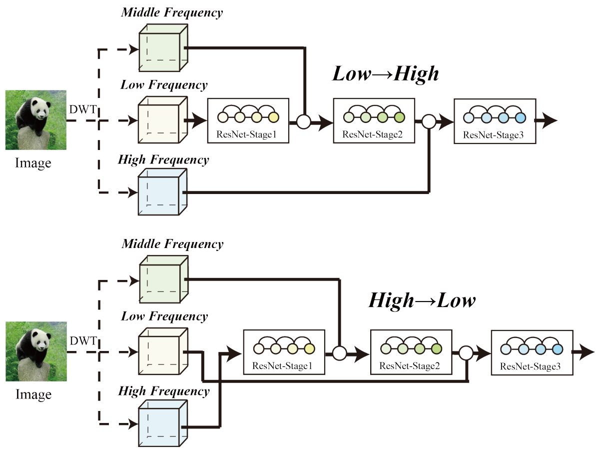

Fig. 8. The proposed J-Net with different frequency injection order. (i) **Low→High** indicates first input low frequency information and then inject high frequency information.(ii) **High→Low** indicates first input high frequency information and then inject low frequency information.

Caltech101 with all frequency is lower than that without high frequency, indicating that the benefit of high frequency is not consistent. This phenomenon is only observed on shallow (18 layers) neural networks on Caltech101, which may be explained by following two reasons: (i) High frequency is injected into J-Net at the last several layers. In shallow neural networks, left layers for high frequency information are insufficient for processing them. (ii) Compared with Caltech256, the classification on Caltech101 is much easier, which results in less importance of high frequency information.

TABLE VII
ACCURACY (%) COMPARISON BETWEEN TWO FREQUENCY ORDERS OF J-RESNET AND J-SENET ON CALTECH101 AND CALTECH256 DATASETS. NUMBER IN BOLD ARE THE BEST.

Models	Baseline	High→Low	Low→High
J-ResNet18	68.34	72.22	**72.84**
J-SENet18	73.64	71.77	**75.32**

TABLE VIII
ACCURACY (%) COMPARISON BETWEEN TWO FREQUENCY ORDERS OF J-RESNET AND J-SENET ON CALTECH101 AND CALTECH256 DATASETS. NUMBER IN BOLD ARE THE BEST.

Models	Baseline	High→Low	Low→High
J-ResNet18	59.35	63.12	**63.53**
J-SENet18	58.27	59.65	**61.98**

C. Influence of frequency order

The information with different frequency is fed into neural networks according to an ascending order (from low to high frequency). An intuitive question is how does the order of three levels frequency information influence the performance of neural networks. To answer this question, a series of experiments on J-Nets have been conducted with the ascending (Low→High) and descending (High→Low) frequency injection order respectively, which have been shown in Figure 8.

The experiments results is shown in Table VIII. It's observed that (i) J-Net with ascending frequency order achieves the highest accuracy in all situations. (ii) J-Net with descending frequency order outperforms the baseline by a less margin compared with ascending frequency order J-Net, indicating that the shallow layers of neural networks are inclined to process the low frequency information while the deep layers tend to process the high frequency information.

TABLE IX
COMPARISON OF THREE FREQUENCY DECOMPOSITION METHODS ON CALTECH101. NUMBER IN BOLD ARE THE BEST.

Models	Baseline	DWT	DCT	DFT
J-ResNet18	68.34	**72.84**	67.85	66.53
J-SENet18	73.64	**75.32**	68.55	66.42

D. J-Net with other decomposition methods

Discrete wavelet transformation (DWT) is utilized in J-Net to decompose images into information with different frequency, which can also be achieved by discrete cosine transformation (DCT) or discrete Fourier transformation (DFT). We further investigate their effectiveness with J-ResNet18, SENet18 on Caltech101 dataset.

As depicted in Table IX, no accuracy improvements can be observed on ether J-ResNet18 or SENet18 with discrete cosine transformation or discrete Fourier transformation, which indicates that the discrete wavelet transformation is a better frequency decomposition method in J-Net.

VI. CONCLUSION

In this paper, we propose a novel neural network framework for visual tasks by inputting the decomposed images gradually into the neural networks. The proposed J-Net is not only effective but also more interpretable and controllable than vanilla convolutional neural networks. Extensive experiments demonstrate that the proposed J-Net can be easily equipped on various neural networks with excellent generalization ability on various datasets, neural networks and wavelet basis.

Moreover, based on the J-Net, we further reveal how the information with different frequency impacts the prediction of neural networks, and study the influence of frequency injection order and other frequency decomposition methods such as DCT (discrete cosine transform) and DFT (discrete Fourier transform). This work may open up a new opportunity for investigating the frequency information for designing neural network architectures.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.
[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[3] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*. IEEE, 2017, pp. 5987–5995.
[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *ECCV*. Springer, 2016, pp. 21–37.
[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015, pp. 91–99.
[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *CVPR*, 2016, pp. 779–788.
[7] Q. Dou, H. Chen, Y. Jin, L. Yu, J. Qin, and P.-A. Heng, "3d deeply supervised network for automatic liver segmentation from ct volumes," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 149–157.
[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
[9] Z.-Q. J. Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," 2019.
[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
[11] Y. Chen, H. Fang, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, S. Yan, and J. Feng, "Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution," 2019.
[12] Y. Wang, C. Xu, S. You, D. Tao, and C. Xu, "Cnnpack: Packing convolutional neural networks in the frequency domain," in *NeurIPS*, 2016, pp. 253–261.
[13] W. Chen, J. Wilson, S. Tyree, K. Q. Weinberger, and Y. Chen, "Compressing convolutional neural networks in the frequency domain," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1475–1484.
[14] S. Vogel, C. Schorn, A. Guntoro, and G. Ascheid, "Guaranteed compression rate for activations in cnns using a frequency pruning approach," in *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2019, pp. 296–299.
[15] O. Rippel, J. Snoek, and R. P. Adams, "Spectral representations for convolutional neural networks," in *NeurIPS*, 2015, pp. 2449–2457.
[16] T. Chen, L. Lin, W. Zuo, X. Luo, and L. Zhang, "Learning a wavelet-like auto-encoder to accelerate deep neural networks," in *AAAI*, 2018.
[17] T. Williams and R. Li, "Wavelet pooling for convolutional neural networks," in *ICLR*, 2018.
[18] S. Fujieda, K. Takayama, and T. Hachisuka, "Wavelet convolutional neural networks for texture classification," 2017.
[19] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-cnn for image restoration," in *CVPR Workshops*, 2018, pp. 773–782.
[20] H. Huang, H. Ran, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *ICCV*, 2017.
[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.
[23] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "Resunet-a: a deep learning framework for semantic segmentation of remotely sensed data," *arXiv preprint arXiv:1904.00592*, 2019.
[24] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.
[25] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Q. Weinberger, "Multi-scale dense networks for resource efficient image classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
[27] S. Mallat, *A wavelet tour of signal processing*. Elsevier, 1999.
[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708.
[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255.
[31] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer vision and Image understanding*, vol. 106, no. 1, pp. 59–70, 2007.
[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.