QFormer: An Efficient Quaternion Transformer for Image Denoising

Bo Jiang[1] , **Yao Lu**[2*] , **Guangming Lu**[2*] and **Bob Zhang**[3]
[1]College of Mechanical and Electronic Engineering, Northwest A&F University, China
[2]Department of Computer Science, Harbin Institute of Technology at Shenzhen, China
[3]Department of Computer and Information Science, University of Macau, China
jiangbo_PhD@gmail.com, luyao2021@hit.edu.cn, luguangm@hit.edu.cn, bobzhang@um.edu.mo

Abstract

Since Deep Convolutional Neural Networks (DCNNs) and Vision Transformer perform well in learning generalizable image priors from large-scale data, these models have been widely used in image denoising tasks. However, vanilla DCNNs and Transformer suffer from two problems. First, the vanilla DCNNs and Transformer only accumulate the output along the channel axis, ignoring the internal relationship among channels. This results in the severely inadequate color structure representation retrieved from color images. Secondly, the DCNNs or Transformer-based image denoising models usually have a large number of parameters, high computational complexity, and slow inference speed. To resolve these issues, this paper proposes a highly-efficient Quaternion Transformer (QFormer) for image denoising. Specifically, the proposed Quaternion Transformer Block (QTB) simplifies the typical Transformer from a multi-branch structure to an elaborately sequential structure mainly with quaternion transformations, to alternately capture both long-range dependencies and local contextual features with color structure information. Furthermore, the proposed QTB can also avoid considerable element-wise multiplications of computing the self-attention matrices. Thus, our QTB can significantly reduce the computational complexity and its sequential structure can further improve the practical inference speed. Comprehensive experiments demonstrate that the proposed QFormer produces state-of-the-art results in both denoising performance and efficiency. We hope that our work will encourage further research to explore the Quaternion Transformer architecture for image denoising tasks.

1 Introduction

Image denoising aims to remove noise from the noisy images to recover high-quality images. At the same time, the image denoising task is an ill-posed problem, thus it is extremely challenging. Although the image denoising performance of Deep Convolutional Neural Networks (DCNNs)-based methods have been significantly improved compared to traditional model-based methods [Joshi *et al.*, 2005], they usually suffer from a critical problem in that vanilla convolution retrieves local spatial information but severely lacks the long-range dependencies. Transformer [Kolesnikov *et al.*, 2021], an alternative to DCNNs, can capture long-range pixel dependencies. This drives the development of Transformer-based methods [Liang *et al.*, 2021] for image denoising, achieving state-of-the-art performances.

Although both DCNNs and Transformer can bring powerful generalization capability to image denoising models, they severely suffer from two main problems. ***(a) The convolution operator and Transformer only simply add the outputs along the channel axis, resulting in ignoring the complicated interrelationship among channels.*** The vanilla convolutional layer in DCNNs and the input/output projection layer in Transformer both independently process each color channel, which is also known as a monochromatic model. A significant limitation of the monochrome model is that the correlation among the three color channels is completely retrieved. This probably leads to color distortion within denoised color images. Concatenation models are proposed to simply weigh the three color channels to exploit channel correlation [Xu *et al.*, 2017]. This approach may not adequately explore the complex inter-relationships among channels. Hence, existing denoising methods are not effective to model such important color structural information *i.e.*, the correlation information among the three color channels. ***(b) Additionally, DCNNs-based or Transformer-based image denoising models are usually inefficient with considerable parameters, high computational complexity, and slow inference speed.*** Although DCNNs can significantly improve image denoising performances with increasing depth and width (number of channels), it can cause a corresponding increase in parameters and computational cost. Furthermore, the main computational overhead of Transformer-based methods comes from the self-attention mechanism, and its complexity grows quadratically with the spatial resolution, leading to a serious inefficiency in processing high-resolution images. Therefore, it is urgent to study an efficient image denoising method considering the complex mutual information along the channel axis.

Quaternion algebra provides an elegant way of working

*Corresponding author

with vector signals [Moxey *et al.*, 2003]. Since the quaternion unit contains one real part and three imaginary parts, color images can be represented as quaternion matrices, enabling multi-channel information to be processed in a parallel manner. Hence, the quaternion transformation is exactly suitable for color image denoising tasks. Considering the powerful representation of color images by quaternions and the powerful ability of modeling long-range pixel dependencies by Transformer, the key difficulties and challenges in designing an efficient image denoising network based on quaternions and Transformer are mainly from two aspects.

(1) Primitive quaternion operator is non-strict identity mapping for the image denoising task. Strict identity mapping can not only preserve overall texture and color information but also generate high-frequency detail information in image processing methods [Song *et al.*, 2021]. From a mathematical point of view, we have demonstrated that primitive quaternion operators embedded in deep neural networks for image denoising tasks are non-strict identity mappings. Therefore, this becomes a challenge to the application of the primitive quaternion operator in the image denoising tasks. To overcome this limitation, we modify the primitive quaternion operator by using short skip connections in the quaternion operator to alleviate this problem.

(2) How to redesign the Transformer framework with a simple structure, low computational complexity, few parameters, and fast inference speed. We propose the core Quaternion Transformer Block (QTB) integrating the improved quaternion operator with strict identity mapping to modularly formulate our framework of highly-efficient Quaternion Transformer (QFormer) Network to capture both long-range dependencies and local context features with abundant color structure information. The proposed QTB reduces the computational complexity through avoiding the element-wise multiplications in computing self-attention matrices. Furthermore, our QTB is further simplified from a multi-branch structure to an elaborately sequential structure. Due to these mechanisms, the proposed QTB can significantly reduce the computational complexity and especially improve the practical inference speed.

We believe that our theoretical statements and empirical results lay the foundation for new Transformers in the super-complex domain. These Transformers can grasp the internal input relationship and reduce the computational cost. As far as we know, this is a promising exploration that a Transformer framework has been defined in a super-complex domain. The contributions of this paper can be summarized as follows:

- We propose a highly-efficient Quaternion Transformer (QFormer) Network using Quaternion Transformer Block (QTB) for the image denoising task. The proposed QTB alternately captures both the long-range dependencies and local contextual features by sufficiently retrieving the color structure information.
- We analyze that quaternion operators are not strictly identically mapped in embedding deep neural networks for image denoising tasks from the mathematical perspective, and thus propose utilizing short skip connections to alleviate this obstacle.
- In terms of model efficiency, the proposed sequential-structured QTB is concise and can avoid considerable element-wise multiplications of computing self-attention matrices, significantly reducing the computation complexity and improving the practical inference speed.
- The proposed QFormer has produced state-of-the-art results from extensive and comprehensive experiments on image denoising tasks, demonstrating satisfactory superiorities in both the denoising performance and efficiency.

2 Related Work

2.1 Quaternion-Based Methods

Since quaternion-based methods can well characterize the cross-channel correlation of color images, they have attracted increasing attention in various vision tasks, such as vector projection [Ell and Sangwine, 2007], and face recognition [Zou *et al.*, 2016], image denoising [Xu *et al.*, 2015]. A model for sparse representation of the quaternion vector was proposed in [Xu *et al.*, 2015] for color image denoising. Several quaternion wavelet methods [Yin *et al.*, 2012] were proposed to achieve a more efficient spatial-spectral image analysis.

The Quaternion Convolutional Neural Network (QCNN) [Zhu *et al.*, 2018] was proposed to extend the real-valued convolutional networks to the complex-valued convolutional networks. Zeng *et al.* [Zeng *et al.*, 2016] proposed a color image classification network based on quaternary principal component analysis. Parcollet *et al.* [Parcollet *et al.*, 2019] proposed studying the influence of the Hamilton product on the task of color image reconstruction based on QCNN only from grayscale images. In general, the main advantage of quaternions for color image representations over vanilla convolutional methods is generalization ability and low computational cost [Xu *et al.*, 2015; Parcollet *et al.*, 2019].

2.2 Vision Transformer

Transformer was originally proposed by [Vaswani *et al.*, 2017] for Natural Language Processing (NLP) tasks, and it achieved amazing performances in language pre-training tasks [Devlin *et al.*, 2018]. Inspired by the application of Transformer in NLP tasks, it has been applied to numerous vision tasks such as image recognition [Kolesnikov *et al.*, 2021], segmentation [Wang *et al.*, 2021a], and object detection [Carion *et al.*, 2020]. Vision Transformer (ViT) [Kolesnikov *et al.*, 2021] embedded hard patches (16×16) to compute correlation and achieved excellent results in image classification. However, the computational complexity of self-attention in Transformer is positively correlated with the square of the number of image patches, resulting in high computational complexity on high-resolution image reconstruction tasks [Liang *et al.*, 2021]. For example, super-resolution [Liang *et al.*, 2021], image colorization [Kumar *et al.*, 2021], denoising [Wang *et al.*, 2021b], and deraining [Wang *et al.*, 2021b]. Much work has focused on improving the transformer's self-attention methods by means of shifting windows [Liang *et al.*, 2021], relative position encod-

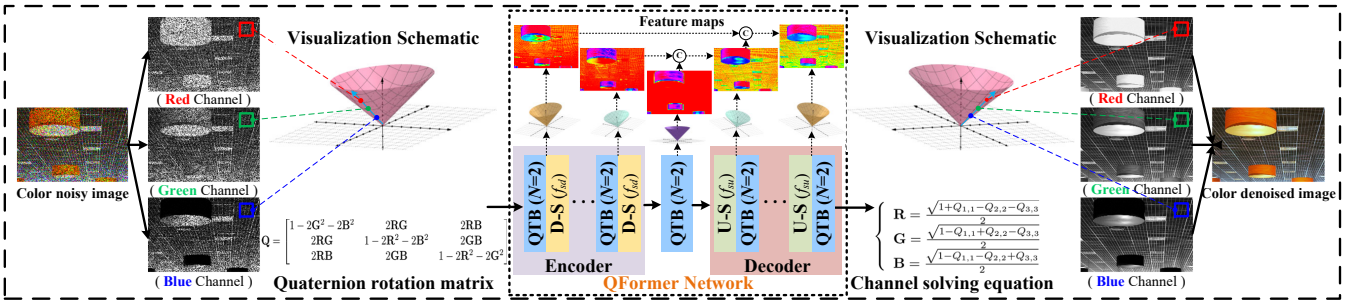

Figure 1: Overall structure of the proposed QFormer Network. The QTB denotes Quaternion Transformer Block. D-S is a downsampling layer with a downsampling factor $f_{sd} = 2$, and U-S is an upsampling layer with an upsampling factor $f_{su} = 2$. © represents the operation of tensor concatenation along the channel dimension.

ing [Wu *et al.*, 2021b], combining convolutions [Wu *et al.*, 2021a], *etc*. However, just utilizing pooling as an alternative to the basic self-attention in MetaFormer [Yu *et al.*, 2021] can still achieve competitive or even state-of-the-art performance. This implies that the most effective part in Transformer may be the Transformer-unique framework rather than the self-attention part.

3 Some Preliminaries

In this section, to easily understand the analysis, we mainly define the mathematical notations and preliminaries of quaternion algebra. Following [Chen *et al.*, 2020], scalars, vectors, matrices, and tensors are denoted as lowercase letters, boldface lowercase letters, boldface capital letters, and boldface italic script letters, *e.g.*, x, $\boldsymbol{x}$, $\mathbf{X}$, and $\boldsymbol{X}$, respectively.

3.1 Quaternion Representation of Color Images

Quaternion was first proposed by Hamilton in 1843 [Hamilton, 2022] as an extension of real space $\mathbb{R}$ and complex space $\mathbb{C}$ to describe the position information of points in three-dimensional space. The definition [Hamilton, 2022] of a quaternion $\dot{p}$ ($\dot{p} \in \mathbb{H}$) has one real part and three imaginary parts as follows:

$$\dot{p} = q_0 + q_1\boldsymbol{i} + q_2\boldsymbol{j} + q_3\boldsymbol{k}, \tag{1}$$

where q_0, q_1, q_2, q_3 ($q_0, q_1, q_2, q_3 \in \mathbb{R}$) are the coefficients of the real and imaginary parts, respectively. $\boldsymbol{i}, \boldsymbol{j}, \boldsymbol{k}$ are three imaginary units obeying the quaternion rules that $\boldsymbol{i}^2 = \boldsymbol{j}^2 = \boldsymbol{k}^2 = \boldsymbol{ijk} = -1$. In addition, if the real part $q_0 = 0$, $\dot{p}$ is called pure quaternion.

Let $\dot{\mathbf{I}}_q$ be an RGB image matrice with the quaternion representation. RGB channel matrices can be represented as a pure quaternion:

$$\dot{\mathbf{I}}_q = \mathbf{R}\boldsymbol{i} + \mathbf{G}\boldsymbol{j} + \mathbf{B}\boldsymbol{k}, \tag{2}$$

where $\mathbf{R}, \mathbf{G}$ and $\mathbf{B}$ indicate the red, green and blue channel of the color image, respectively.

To directly establish the relationship among the $\mathbf{R}, \mathbf{G}$ and $\mathbf{B}$ channels, as shown in Fig. 1, we use a quaternion rotation matrix to closely connect the three channels of the color image represented by pure quaternions. Using the Rodrigues formula [Murray *et al.*, 1994], the quaternion rotation matrix $\mathbf{Q}$ can be obtained from Eqn. 2:

$$\mathbf{Q} = \begin{bmatrix} 1-2\mathbf{G}^2-2\mathbf{B}^2 & 2\mathbf{RG} & 2\mathbf{RB} \\ 2\mathbf{RG} & 1-2\mathbf{R}^2-2\mathbf{B}^2 & 2\mathbf{GB} \\ 2\mathbf{RB} & 2\mathbf{GB} & 1-2\mathbf{R}^2-2\mathbf{G}^2 \end{bmatrix}. \tag{3}$$

Furthermore, in order to compute the $\mathbf{R}, \mathbf{G}$ and $\mathbf{B}$ channels of the reconstructed color image from the quaternion rotation matrix, we used Eqn. 4 to compute the information for the three channels:

$$\begin{cases} \mathbf{R} = \frac{\sqrt{1+Q_{1,1}-Q_{2,2}-Q_{3,3}}}{2} \\ \mathbf{G} = \frac{\sqrt{1-Q_{1,1}+Q_{2,2}-Q_{3,3}}}{2} \\ \mathbf{B} = \frac{\sqrt{1-Q_{1,1}-Q_{2,2}+Q_{3,3}}}{2} \end{cases}, \tag{4}$$

where $Q_{i,j}$ represents the element at the i^{th} row and the j^{th} column in quaternion rotation matrix $\mathbf{Q}$.

3.2 Identity Mapping of QNNs

The essence of image denoising methods based on traditional end-to-end DCNNs is to learn the mapping relationship from noisy images to clean images. Since the difference between the input and output feature maps of each convolutional layer is very small, this phenomenon is called strict identity mapping, which is greatly important for deep learning-based image processing methods [Song *et al.*, 2021]. Strict identity mapping not only preserves overall texture and color information but also produces high-frequency detail information. Although the mutual information among the color channels can be sufficiently preserved by the quaternion neural networks, it cannot always produce high-frequency detail information, such as the relation described in **Theorem 1** below.

Theorem 1. *For the arbitrary noisy image $\dot{\mathbf{I}}_\mathbf{n}$, and an quaternion filter with its weight parameters $\dot{\mathbf{W}}$. When $\Delta\zeta = 0$, there exists a $\dot{\mathbf{W}}$ that satisfies Eqn. 5, which means that the quaternion neural network cannot produce high-frequency detail information.*

$$\dot{\mathbf{I}}_\mathbf{n} \approx \dot{\mathbf{I}}_\mathbf{n} \circledast \dot{\mathbf{W}} = \dot{\mathbf{I}}_\mathbf{n} + \Delta\zeta, \tag{5}$$

where $\Delta\zeta$ denotes the high-frequency detail information and $\circledast$ represents the quaternion convolution operation defined in [Zhu et al.*, 2018]. Eqn. 5 can be rewritten as Eqn. 6 according to the above constraints:*

$$\dot{\mathbf{I}}_\mathbf{n} = \dot{\mathbf{I}}_\mathbf{n} \circledast \dot{\mathbf{W}}, \tag{6}$$

Proof. Here we consider a quaternion fully-connected layer for simplicity. Furthermore, the following proof can be easily extended to quaternion convolution layers. Then, let $\dot{\mathbf{I}}_\mathbf{n}$ be the identity quaternion matrix, we have:

$$\dot{\mathbf{I}}_\mathbf{n} = \dot{\mathbf{I}}, \tag{7}$$

where $\dot{\mathbf{I}}$ is an identity matrix whose elements are all 1. Then,

$$\dot{\mathbf{I}}_\mathbf{n} \circledast \dot{\mathbf{W}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{s_i}\hat{w}_i\hat{a}_i\hat{w}_i^*, \tag{8}$$

where $\hat{a}_i = 1$ denotes an element of the identity quaternion matrix $\dot{\mathbf{I}}_\mathbf{n}$, $s_i > 0$ ($s_i \in \mathbb{R}$) is the magnitude of $\hat{w}_i$, and N is the number of elements in the quaternion matrix. According to literature [Zhu *et al.*, 2018], $\hat{w}_i$ follows the quaternion convolution weight:

$$\hat{w}_i = s_i\left(\cos\frac{\theta_i}{2} + \sin\frac{\theta_i}{2}\boldsymbol{\epsilon}\right), \tag{9}$$

where $\theta_i \in [-\pi, \pi]$ is the rotation angle parameter of the quaternion convolution operation. $\boldsymbol{\epsilon}$ is the unit basis vector, *i.e.*, $\boldsymbol{\epsilon} = \frac{\sqrt{3}}{3}(\boldsymbol{i} + \boldsymbol{j} + \boldsymbol{k})$.

Then we have:

$$\begin{aligned}
\dot{\mathbf{I}}_\mathbf{n} \circledast \dot{\mathbf{W}} &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{s_i}\hat{w}_i\hat{w}_i^* \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{s_i}\left(\cos\frac{\theta_i}{2} + \sin\frac{\theta_i}{2}\boldsymbol{\epsilon}\right)\left(\cos\frac{\theta_i}{2} - \sin\frac{\theta_i}{2}\boldsymbol{\epsilon}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{s_i}\left[\left(\cos\frac{\theta_i}{2}\right)^2 - \left(\sin\frac{\theta_i}{2}\boldsymbol{\epsilon}\right)^2\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{s_i}\left[1 - 2\left(\sin\frac{\theta_i}{2}\right)^2\boldsymbol{\epsilon}^2\right]
\end{aligned} \tag{10}$$

Since the value range of $\left[1 - 2\left(\sin\frac{\theta_i}{2}\right)^2\boldsymbol{\epsilon}^2\right]$ is $[-1, 1]$, combining Eqns. 7 and 10, we have $\frac{1}{N}\sum_{i=1}^{N}\frac{1}{s_i}\hat{w}_i\hat{w}_i^* = \dot{\mathbf{I}}$, which is obviously possible.

□

According to the above theorem and analysis, although the quaternion neural networks can fully retain the mutual information of the color channels, they cannot always produce high-frequency detail information. This indicates the primitive quaternion operator lacks a strict identity mapping relationship. While stacking more quaternion layers may alleviate this problem, it can significantly increase the model size and computational cost. To address this issue, in practice, we provide a short skip connection operation for quaternion convolutional units or fully connected units, *i.e.*:

$$\dot{\mathbf{I}}_\mathbf{y}^m \approx \dot{\mathbf{I}}_\mathbf{x}^m + \underbrace{\dot{\mathbf{W}}^m \circledast \dot{\mathbf{I}}_\mathbf{x}^m}_{\Delta\zeta}, \tag{11}$$

where $\dot{\mathbf{W}}^m$ is the weights of quaternion filters in the m^{th} layer, $\dot{\mathbf{I}}_\mathbf{x}^m$ and $\dot{\mathbf{I}}_\mathbf{y}^m$ are the input and output data, respectively.

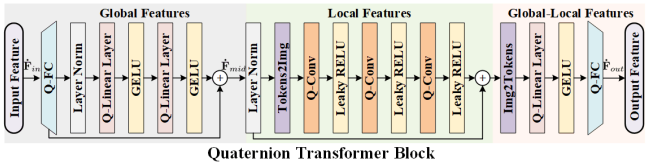

Figure 2: Quaternion Transformer Block (QTB). $\otimes$ denotes matrix multiplication, $\oplus$ and $\odot$ are element-wise multiplication and addition, respectively. MSA is multi-head self-attention layer.

Considering $\dot{\mathbf{W}}^m \circledast \dot{\mathbf{I}}_\mathbf{x}^m$ as $\Delta\zeta$, that is, high-frequency detail information, according to the analysis of **Theorem 1**, the short skip connection operation can effectively alleviate the shortage of strict identity mapping of quaternion operators.

4 QFormer Networks

4.1 Quaternion Transformer Block (QTB)

As shown in Fig.1, the overall structure of the proposed QFormer is a symmetric network mainly composed of encoder and decoder, with the Quaternion Transformer Block (QTB) as its core component. As shown in Fig.2, QTB is an elaborately sequential structure mainly with quaternion transformations. Our QTB and the multi-head self-attention within the traditional Transformer are different in the following two aspects.

(1) QTB captures long-range dependencies also containing local features with abundant color structure information. Capturing long-range dependencies aims to improve the receptive field of deep neural networks to improve model performance. The self-attention mechanism brute forcibly captures long-range dependencies by computing the interaction between any two patches using element-wise multiplication. Diametrically, our QTB first refines global features through several fully connected layers based on the quaternion operator and GELU activation function for the purpose of global information extraction. Then the quaternion operator is used to further extract local features with color structure information from the refined global features to obtain local information. Finally, the produced global information is supplemented with the local information by element-wise aggregating them together using skip connections. Such captured long-term dependencies contain abundant local features with color structure information.

(2) The proposed QTB is more efficient in both computational complexity and practical inference speed than that of the self-attention mechanism. Because the self-attention mechanism adopts element-wise multiplication to capture long-range dependencies, its computational complexity increases quadratically with the size of the input feature. This greatly increases the training and inference costs of the model. Compared with the self-attention mechanism, the proposed QTB uses element-wise addition to aggregate global features with local features to capture long-range dependencies to avoid excessive computational complexity. Therefore, our QTB can significantly reduce the computational complexity. Furthermore, the sequential structure of the proposed

QTB can further promote the practical inference speed, evaluated in Section 5.4.

Specifically, the QTB input $\dot{\mathbf{F}}_{in} \in \mathbb{H}^{H\times W\times C}$ is first reshaped to tokens, which are fed into a quaternion fully connected layer to extract global features, as shown in Eqn. 12:

$$\dot{\mathbf{F}}_c = f_{qfc}\left(\dot{\mathbf{F}}_{in} \in \mathbb{H}^{H\times W\times C} \rightarrow \dot{\mathbf{F}}_{in} \in \mathbb{H}^{HW\times C}\right), \quad (12)$$

where f_{qfc} denotes quaternion fully connected layer, and $\dot{\mathbf{F}}_c \in \mathbb{H}^{HW\times C}$ is the global features. Then, after normalizing $\dot{\mathbf{F}}_c$, we use the quaternion linear layer and the GELU non-linearity activation layer to further capture the global feature. The computation of the above process is represented as:

$$\dot{\mathbf{F}}_{mid} = \phi\left(f_{ql}^2\left(\phi\left(f_{ql}^1\left(\mathrm{LN}\left(\dot{\mathbf{F}}_c\right)\right)\right)\right)\right) + \dot{\mathbf{F}}_c, \quad (13)$$

where LN represents the layer normalization [Ba *et al.*, 2016], f_{ql}^1 and f_{ql}^2 both are quaternion convolutional layers, and ϕ denotes GELU non-linearity activation layer [Hendrycks and Gimpel, 2016].

To enhance the ability of QTB to capture local contextual feature, several quaternion convolutional layers are attached to process $\dot{\mathbf{F}}_{mid} \in \mathbb{H}^{H\times W\times C}$. As shown in Fig. 2, we first apply a normalization layer to each token. Next, the tokens are reshaped into 2D feature maps, and then three sets of quaternion convolutions with the kernel size of 3×3 and GELU activation layers are employed to capture local information. Then, the local features are flattened into tokens and the number of local feature channels is reduced to the same number as the global feature channels using a quaternion linear layer. Finally, element-wise addition of global and local features is conducted to capture long-range dependencies.

4.2 Loss Function

The Charbonnier loss is applied as the ultimate loss function for the proposed QFormer, formulated as follows:

$$\mathcal{L} = \frac{1}{N_t}\sum\sqrt{\|\mathbf{I}_d - \mathbf{I}_{traget}\|^2 + \epsilon^2}. \quad (14)$$

where N_t denotes the number of training samples, $\mathbf{I}_{traget}$ represents the groundtruth corresponding to the input noise image, ϵ^2 is a constant that is empirically set to 1×10^{-6}.

5 Experiments

5.1 Architecture Scales

In order to illustrate the high efficiency of the proposed QFormer in image denoising, three different scales of parameters for QFormer are constructed in our experiments, *i.e.*, QFormer-T (Tiny, $C = 16$), QFormer-S (Small, $C = 32$) and QFormer-B (Base, $C = 44$). The difference between the above settings is only the number of feature channels C, and other settings remain unchanged, such as the depth N of QTB is set to 2.

5.2 Evaluation on Real-world Noisy Images

The QFormer's performance on real-world noisy images, containing complex and unknown noise, is evaluated, highlighting its practical value in real-world denoising applications. Table 1 presents the results obtained from denoising real noisy image datasets. The proposed QFormer is subject to a comparative analysis against fourteen state-of-the-art denoising methods. It is evident that the proposed QFormer significantly enhances the PSNR/SSIM results when compared to the other fourteen state-of-the-art methods across the three real noisy image datasets. In the case of the SIDD dataset, QFormer-T demonstrates similar gains in PSNR and SSIM as Uformer and Restormer. Furthermore, it achieves competitive performance alongside HPDNet on the PolyU datasets and with APD-Nets on the Nam datasets. Additionally, QFormer-S surpasses HPDNet and Restormer with an average PSNR improvement of **0.26 dB** on PolyU and **0.04 dB** on SIDD, respectively. Moreover, QFormer-B outperforms APD-Nets by an average PSNR gain of **0.01 dB** on Nam. These findings underscore the effectiveness of the proposed QFormer network structure in denoising real noisy images. QFormer leverages the sequential structure of QTB to refine global features, retrieve channel-related local features, and aggregate global-local features for capturing long-term dependencies. This further implies that QFormer adeptly utilizes these long-range (non-local) feature to enhance denoising performance.

Dataset	Nam		PolyU		SIDD	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DnCNN-B [Zhang *et al.*, 2017]	36.08	0.903	35.74	0.878	38.56	0.910
FFDNet+ [Zhang *et al.*, 2018]	37.85	0.938	37.19	0.939	38.60	0.909
TWSC [Xu *et al.*, 2018]	38.37	0.952	37.63	0.954	35.89	0.838
CBDNet [Guo *et al.*, 2019]	38.51	0.957	37.85	0.956	38.68	0.909
RIDNet [Anwar and Barnes, 2019]	38.72	0.960	38.07	0.957	38.71	0.913
VDN [Yue *et al.*, 2019]	39.16	0.965	38.43	0.960	39.29	0.911
PAN-Net [Ma *et al.*, 2021]	40.18	0.978	39.91	0.971	39.33	0.912
AINDNet [Kim *et al.*, 2020]	39.21	0.966	38.78	0.963	39.45	0.915
MIRNet [Zamir *et al.*, 2020]	39.88	0.973	39.25	0.971	39.71	0.959
HPDNet [Ma *et al.*, 2022]	40.26	0.979	39.89	0.970	39.72	0.958
APD-Nets [Jiang *et al.*, 2022]	40.36	0.989	N/A	N/A	39.75	0.959
Uformer [Wang *et al.*, 2021b]	N/A	N/A	N/A	N/A	39.77	0.959
Restormer [Zamir *et al.*, 2021]	N/A	N/A	N/A	N/A	40.02	0.960
QFormer-T	40.03	0.975	39.90	0.971	39.72	0.959
QFormer-S	40.19	0.980	40.15	0.973	40.06	0.961
QFormer-B	40.37	0.982	40.31	0.974	40.18	0.963

Table 1: Average PSNR and SSIM of the denoised real images from Nam, PolyU and SIDD datasets. PSNR and SSIM are positively correlated with visual quality.

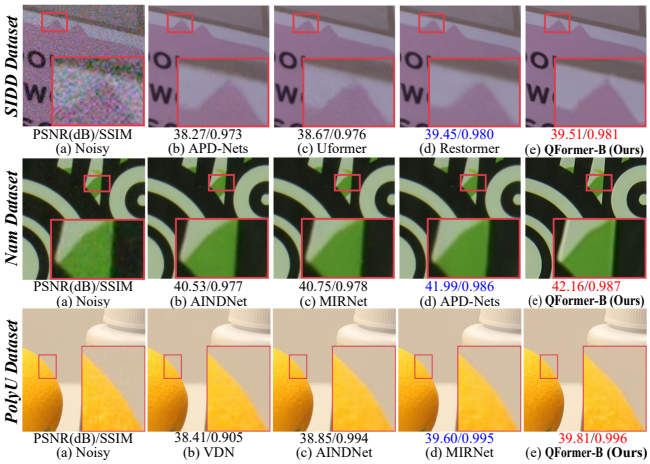

Figure 3: Visual comparisons between QFormer and its competitors in the evaluation of real-noisy image denoising.

Dataset	Nam		PolyU		SIDD	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Case.I: QFormer (convolution operator)	39.28	0.967	38.80	0.963	39.53	0.925
Case.II: QFormer (short skip *x*)	39.15	0.962	38.73	0.959	39.51	0.921
Case.III:QFormer (short skip ✓)	**40.37**	**0.982**	**40.31**	**0.974**	**40.18**	**0.963**

Table 2: Performance effect of identity mapping.

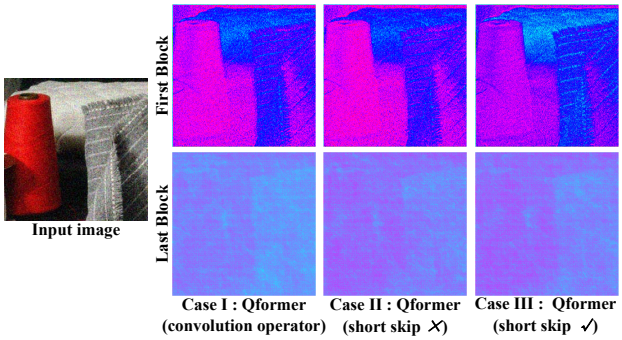

Figure 4: Effect of identity mapping on the output feature map of QTB in QFormer network.

To visually demonstrate the superiority of our method, Fig. 3 shows a visual comparison of different methods on different datasets for denoising real noisy images. We can observe that our QFormer achieves the best visual results in terms of noise removal and detail preservation. For example, neither Uformer nor Restormer can preserve the boundary contours of the fine pink regions, while the proposed QFormer can reconstruct and preserve the underlying edges. This further demonstrates that the correlation of color channel and long-range dependencies can improve denoising performance.

5.3 Ablation Study

To visually illustrate the impact of quaternion operation and identity mapping on image denoising, we use the following case to conduct experiments. Case.I means that only the convolution operator is used in the QFormer structure. Case.II means that only the primitive quaternion operator (*i.e.*, without short skip connections) is used in the QFormer structure. Case.III means that only the proposed quaternion operator (*i.e.*, using short skip connections) is used in the QFormer structure. It is worth noting that the convolution operator is an identity map of [Song *et al.*, 2021].

Effect of Quaternion Operation As shown in Fig. 4, compared to QFormer (Case.I) consisting of pure convolution operators, QFormer (Case.III) with short skip connections retains and enhances more detailed information (*e.g.*, boundary, edges). The experimental environment, experimental settings, and network structure of Case.I and Case.III are consistent. Therefore, the reason why the feature maps in Case.I and Case.III have large differences are that the quaternion operator and the vanilla convolution operator only model the correlation between the input noise image channels differently. This fully demonstrates that, compared with the vanilla convolution operator, the quaternion convolution operator can effectively and adequately model the color structure information of the input image to improve the feature representation ability. To quantify the impact of quaternion operation on image denoising performance, the PSNR/SSIM of Case.I, Case.II, and Case.III on real datasets are reported in Table 2. Compared with QFormer based on pure vanilla convolution operator (Case.I), QFormer with short skip connections (Case.III) improves PSNR by **1.09 dB**, **1.51 dB** and **0.65 dB** on Nam, PolyU and SSID datasets. This illustrates that the quaternion operator can improve the image denoising performance by modeling the color structure information of the input image.

Method	Case	Quaternion	Nam	PolyU	SIDD
PSNR	a	✗(UNet)	38.87	38.09	39.14
	b	✗(QFormer-B)	39.28	38.80	39.53
	c	✓(UNet)	38.92	38.56	38.84
	d	✓(QFormer-B)	40.37	40.31	40.18
SSIM	e	✗(UNet)	0.959	0.952	0.906
	f	✗(QFormer-B)	0.967	0.963	0.925
	g	✓(UNet)	0.961	0.957	0.911
	h	✓(QFormer-B)	0.982	0.974	0.963
Speed (s)	i	✗(UNet)	0.063	0.060	0.061
	j	✗(QFormer-B)	0.057	0.049	0.056
	k	✓(UNet)	0.060	0.058	0.058
	l	✓(QFormer-B)	0.018	0.016	0.018
FLOPs (G)	m	✗(UNet)	18.28	18.28	18.28
	n	✗(QFormer-B)	15.53	15.53	15.53
	o	✓(UNet)	8.24	8.24	8.24
	p	✓(QFormer-B)	6.30	6.30	6.30
#Params (MB)	q	✗(UNet)	15.24	15.24	15.24
	r	✗(QFormer-B)	13.01	13.01	13.01
	s	✓(UNet)	9.62	9.62	9.62
	t	✓(QFormer-B)	8.35	8.35	8.35

Table 3: Ablation study of quaternion operators in terms of FLOPs, inference time, the number of parameters, PSNR (dB) and SSIM on Nam, PolyU and SIDD test datasets. ✗indicates that the model consists only of vanilla convolution operators, ✓denotes that the model consists only of quaternion operators.

Effect of Identity Mapping In this section, as shown in Fig. 4, we perform a more intuitive visualization of feature maps within the proposed QFormer interiors on different settings to observe the differences. It can be seen from the figure that, compared with the QFormer without short skip connections (Case.II), high-frequency information (*e.g.*, lines, points) is clearly emphasized in the feature maps generated by QFormer with short skip connections (Case.III). This illustrates that the lack of strict identity mapping of the primitive quaternion operator may be a major obstacle to the improvement of image denoising performance. At the same time, it also demonstrates that the short skip connection operation can overcome the defect of the non-strict identity mapping of the original quaternion operator by superimposing the input features on the output features. To quantify the impact of identity mapping on image denoising performance, the PSNR/SSIM of Case.I, Case.II, and Case.III on real datasets are reported in Table 2. From the table, it can be found that compared with QFormer without short skip connections (Case.II), QFormer

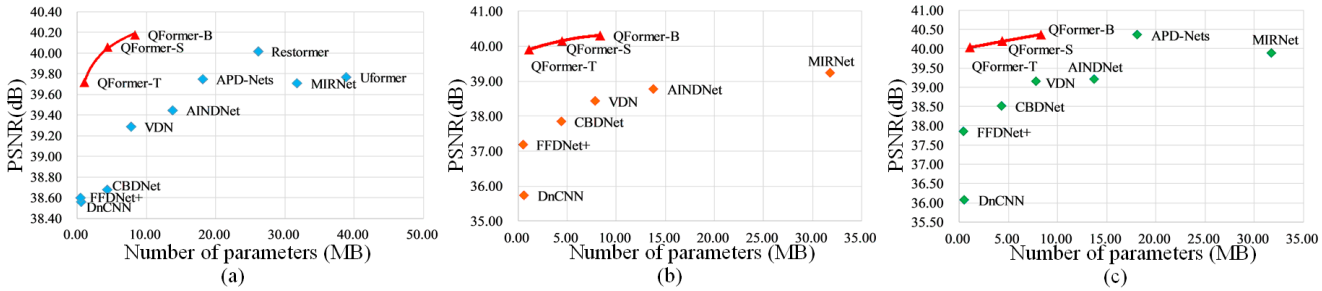

Figure 5: PSNR and parameters on the real noisy image testing sets. (a) On SSID dataset. (b) On PolyU dataset. (c) On Nam dataset.

with short skip connections(Case.III) improves PSNR by **1.22 dB**, **1.58 dB** and **0.67 dB** on Nam, PolyU, and SSID datasets, respectively. By comparing the specific quantitative measurements, it can be observed that short skip connections can rise above the non-strict identity mapping of the primitive quaternion operator and greatly improve denoising performance.

5.4 Efficiency of QFormer

The reported results of the model sizes are shown in Fig. 5. It is worth noting that the proposed QFormer-B achieves the best image denoising performance on the SIDD, PonlyU, and Nam real noisy image datasets. In addition, on the SIDD dataset, compared with the state-of-the-art Restormer, the proposed QFormer-T, QFormer-S, and QFormer-B reduce the number of parameters by about **22.8 times**, and **5.7 times** and **3.0 times**, respectively. Similarly, on the PonlyU dataset, compared with the state-of-the-art MIRNet, the parameters of the proposed QFormer-T, QFormer-S, and QFormer-B are reduced by about **36.0 times**, **9.1 times** and **4.8 times**, respectively. On the Nam dataset, compared with the state-of-the-art APD-Nets, the parameters of the proposed QFormer-T, QFormer-S, and QFormer-B are reduced by about **16.3 times** and **4.1 times**, and **2.2 times**, respectively.

Additionally, we further investigate the advantages of quaternion operators, and QTB structure within QFormer. ***First, we discuss the effect of quaternions within QFormer***, as shown in Table 3. We can find that once all the quaternion convolution and quaternion fully connected layers in QFormer are replaced by vanilla convolution and fully connected layers, the image denoising performance, and computational efficiency drop. Specifically, while keeping the overall structure of QFormer unchanged, compared with the model with quaternion operator (*i.e.*, Case.d and Case.h), the image denoising performance PSNR/SSIM of the model without quaternion operator (*i.e.*, Case.b and Case.f) decreases significantly on the three real noisy image datasets, with an average decrease of about **1.08 dB** and **0.021**, respectively. In comparisons in terms of efficiency, *i.e.*, Case.j vs Case.l, Case.n vs Case.p and Case.r vs Case.t, the number of parameters, inference speed, and FLOPs all decline prominently. This fully demonstrates the great potential of the quaternion operator to achieve a better balance between the performance and efficiency of image denoising. The overall UNet structure is also used and unchanged, and the quaternion operator is introduced to replace the vanilla convolutions (*i.e.*, Case.c, Case.j, Case.k, Case.o and Case.s). Compared with the UNet model using the convolution operator (*i.e.*, Case.a, Case.e, Case.i, Case.m and Case.q), the image denoising performance and computational efficiency of the UNet with the quaternion operator are significantly improved. This further indicates that the quaternion operator can also play an important role in image denoising performance and efficiency without introducing the QTB structure.

Secondly, we discuss the effect of the QTB structure in QFormer. We compare UNet, which employs the quaternion operator (*i.e.*, Case.c, Case.j, Case.k, Case.o and Case.s), with the proposed QFormer. From Table 3, it can be found that the PSNR/SSIM of QFormer with QTB structure on the three test sets of real noisy images is higher than those of Case.c and Case.j by about **1.51 dB** and **0.030**, respectively. In addition, QFormer with QTB structure improves the inference speed by about **70.46**%. This suggests the proposed QTB structure is better at the ability to model the long-range dependencies for heavy noise removal with satisfactory efficiency.

6 Conclusion

In this paper, we propose a highly efficient Quaternion Transformer (QFormer) Network using Quaternion Transformer Block (QTB) for the image denoising task. The proposed QTB alternately captures both the long-range dependencies and local contextual features with sufficiently retrieving color structure information. In addition, from the mathematical perspective, we explained that quaternion operators are not strictly identically mapped in embedding deep neural networks for image denoising tasks, and thus proposed using short skip connections to ease this limitation. The proposed sequential-structured QTB is concise and can avoid considerable element-wise multiplications of computing self-attention matrices, thereby reducing computation complexity and improving practical inference speed. From extensive and comprehensive experiments on denoising tasks, the proposed QFormer achieves state-of-the-art results, demonstrating the satisfactory superiority of denoising and efficiency. We hope this pioneering efficient QFormer structure can encourage further exploration of the Quaternion Transformer architecture for image denoising tasks.

Acknowledgements

This work was supported in part by the NSFC fund (NO. 62176077, 62206073), in part by the Shenzhen Key Technical Project (NO. JSGG20220831092805009, JSGG20220831105603006, JSGG20201103153802006, KJZD20230923115117033), in part by the Guangdong International Science and Technology Cooperation Project (NO. 2023A0505050108), in part by the Shenzhen Fundamental Research Fund (NO. JCYJ20210324132210025), in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (NO. 2022B1212010005), and in part by the Guangdong Shenzhen joint Youth Fund under Grant 2021A151511074, in part by the Natural Science Foundation of Guangdong Province under Grant 2023A1515010893, in part by the Shenzhen Doctoral Initiation Technology Plan under Grant RCBS20221008093222010.

References

[Anwar and Barnes, 2019] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3155–3164, 2019.

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[Chen *et al.*, 2020] Yongyong Chen, Xiaolin Xiao, and Yicong Zhou. Low-rank quaternion approximation for color image processing. *IEEE Transactions on Image Processing*, 29:1426–1439, 2020.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Ell and Sangwine, 2007] Todd A Ell and Stephen J Sangwine. Quaternion involutions and anti-involutions. *Computers & Mathematics with Applications*, 53(1):137–143, 2007.

[Guo *et al.*, 2019] Shi Guo, Zifei Yan, K. Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1712–1722, 2019.

[Hamilton, 2022] William Rowan Hamilton. *Elements of quaternions*. BoD–Books on Demand, 2022.

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[Jiang *et al.*, 2022] Bo Jiang, Yao Lu, Jiahuan Wang, Guangming Lu, and David Zhang. Deep image denoising with adaptive priors. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[Joshi *et al.*, 2005] Manjunath V Joshi, Subhasis Chaudhuri, and Rajkiran Panuganti. A learning-based method for image super-resolution from zoomed observations. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(3):527–537, 2005.

[Kim *et al.*, 2020] Yoonsik Kim, Jae Woong Soh, Gu Yong Park, and Nam Ik Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3479–3489, 2020.

[Kolesnikov *et al.*, 2021] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[Kumar *et al.*, 2021] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. *arXiv preprint arXiv:2102.04432*, 2021.

[Liang *et al.*, 2021] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021.

[Ma *et al.*, 2021] Ruijun Ma, Bob Zhang, Yicong Zhou, Zhengmin Li, and Fangyuan Lei. Pid controller-guided attention neural network learning for fast and effective real photographs denoising. *IEEE transactions on neural networks and learning systems*, PP, 2021.

[Ma *et al.*, 2022] Rui Ma, Shuyi Li, Bob Zhang, and Zhengmin Li. Towards fast and robust real image denoising with attentive neural network and pid controller. *IEEE Transactions on Multimedia*, pages 2366–2377, 2022.

[Moxey *et al.*, 2003] C Eddie Moxey, Stephen J Sangwine, and Todd A Ell. Hypercomplex correlation techniques for vector images. *IEEE Transactions on Signal Processing*, 51(7):1941–1953, 2003.

[Murray *et al.*, 1994] Richard M. Murray, S. Shankar Sastry, and Li Ze-xiang. A mathematical introduction to robotic manipulation. 1994.

[Parcollet *et al.*, 2019] Titouan Parcollet, Mohamed Morchid, and Georges Linarès. Quaternion convolutional neural networks for heterogeneous image processing. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8514–8518. IEEE, 2019.

[Song *et al.*, 2021] Dehua Song, Yunhe Wang, Hanting Chen, Chang Xu, Chunjing Xu, and DaCheng Tao. Addersr: Towards energy efficient image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15648–15657, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2021a] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 568–578, 2021.

[Wang *et al.*, 2021b] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021.

[Wu *et al.*, 2021a] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22–31, 2021.

[Wu *et al.*, 2021b] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10033–10041, 2021.

[Xu *et al.*, 2015] Yi Xu, Licheng Yu, Hongteng Xu, Hao Zhang, and Truong Nguyen. Vector sparse representation of color image using quaternion matrix analysis. *IEEE Transactions on image processing*, 24(4):1315–1329, 2015.

[Xu *et al.*, 2017] Jun Xu, Lei Zhang, David Zhang, and Xiangchu Feng. Multi-channel weighted nuclear norm minimization for real color image denoising. In *Proceedings of the IEEE international conference on computer vision*, pages 1096–1104, 2017.

[Xu *et al.*, 2018] Jun Xu, Lei Zhang, and David Dian Zhang. A trilateral weighted sparse coding scheme for real-world image denoising. *ArXiv*, abs/1807.04364, 2018.

[Yin *et al.*, 2012] Ming Yin, Wei Liu, Jun Shui, and Jiangmin Wu. Quaternion wavelet analysis and application in image denoising. *Mathematical Problems in Engineering*, 2012, 2012.

[Yu *et al.*, 2021] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. *arXiv preprint arXiv:2111.11418*, 2021.

[Yue *et al.*, 2019] Zongsheng Yue, Hongwei Yong, Qian Zhao, Lei Zhang, and Deyu Meng. Variational denoising network: Toward blind noise modeling and removal. In *NeurIPS*, 2019.

[Zamir *et al.*, 2020] Syed Waqas Zamir, Aditya Arora, Salman Hameed Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. *ArXiv*, abs/2003.06792, 2020.

[Zamir *et al.*, 2021] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv preprint arXiv:2111.09881*, 2021.

[Zeng *et al.*, 2016] Rui Zeng, Jiasong Wu, Zhuhong Shao, Yang Chen, Beijing Chen, Lotfi Senhadji, and Huazhong Shu. Color image classification via quaternion principal component analysis network. *Neurocomputing*, 216:416–428, 2016.

[Zhang *et al.*, 2017] K. Zhang, W. Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26:3142–3155, 2017.

[Zhang *et al.*, 2018] K. Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27:4608–4622, 2018.

[Zhu *et al.*, 2018] Xuanyu Zhu, Yi Xu, Hongteng Xu, and Changjian Chen. Quaternion convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–647, 2018.

[Zou *et al.*, 2016] Cuiming Zou, Kit Ian Kou, and Yulong Wang. Quaternion collaborative and sparse representation with application to color face recognition. *IEEE Transactions on image processing*, 25(7):3287–3302, 2016.