Enhanced Channel Attention Network With Cross-Layer Feature Fusion for Spectral Reconstruction in the Presence of Gaussian Noise

Changwu Zou, Can Zhang, Minghui Wei, and Changzhong Zou

Abstract—Spectral reconstruction from RGB images has made significant progress. Previous works usually utilized the noise-free RGB images as input to reconstruct the corresponding hyperspectral images (HSIs). However, due to instrumental limitation or atmospheric interference, it is inevitable to suffer from noise (e.g., Gaussian noise) in the actual image acquisition process, which further increases the difficulty of spectral reconstruction. In this article, we propose an enhanced channel attention network (ECANet) to learn a nonlinear mapping from noisy RGB images to clean HSIs. The backbone of our proposed ECANet is stacked with multiple enhanced channel attention (ECA) blocks. The ECA block is the dual residual version of the channel attention block, which makes the network focus on key auxiliary information and features that are more conducive to spectral reconstruction. For the case that the input RGB images are disturbed by Gaussian noise, cross-layer feature fusion unit is used to concatenate the multiple feature maps at different depths for more powerful feature representations. In addition, we design a novel combined loss function as the constraint of the ECANet to achieve more accurate reconstruction result. Experimental results on two HSI benchmarks, CAVE and NTIRE 2020, demonstrate that the effectiveness of our method in terms of both visual and quantitative over other state-of-the-art methods.

Index Terms—Cross-layer feature fusion (CLFF), combined loss function, enhanced channel attention (ECA), Gaussian noise, spectral reconstruction.

I. INTRODUCTION

HYPERSONTICAL imaging is an analytical process based on spectroscopy in which the spectral signature can be densely sampled to hundreds or even thousands of narrow bands, not only in visible spectrum but also in near-infrared spectrum. Compared with RGB images, hyperspectral images (HSIs) provide not only 2-D spatial information, but also abundant spectral information. The rich spectral signals in HSIs have been verified to be very effective in various signal processing fields, such as object tracking [1], [2], image classification [3], [4], face recognition [5], [6], and scene segmentation [7]. However, traditional HSI acquisition relies on spatial-spectral scanning to capture 3-D signals, which increases the exposure time and limits its application scope in dynamic scenes. To address this problem, a series of methods based on compressed sensing are proposed, such as computed tomographic imaging spectrometry (CTIS) [8], coded aperture snapshot spectral imager (CASSI) [9], and spatial-spectral encoded compressive HSI imager (SSCSI) [10]. The approaches based on compressed sensing use the sparsity of spectral signals for compressive sampling, so as to recover 3-D spectral data. However, expensive hardware systems and complex reconstruction algorithms make these methods hard to obtain HSIs. Conversely, RGB images are relatively easy to obtain. Therefore, accurate spectral reconstruction from RGB images becomes the research focus for scholars recently.

In fact, spectral reconstruction is to obtain 31 channels of HSIs through three channels of RGB images, which can be regarded as the inverse process of the traditional RGB image acquisition. Since this task needs to estimate the information that cannot be collected in RGB imaging system, the HSIs that can be obtained from RGB images are usually not unique, and the solution is not stable, which is a serious ill-posed problem. At the same time, various noises that may be generated in the process of acquiring and transmitting RGB images also need to be considered. Fortunately, it has been found that there is a close correspondence between pixels of RGB and HSI pairs in a given natural scene. Thus, it is still feasible to recover spectral signals from an RGB image by establishing an optimal problem with some image priors as the regularization term, which have been widely used in the signal processing field. The required regularization term can explore the image structure feature and restrict the solution space for the ill-posed problem. The commonly used image priors are the sparse coding, total variation, etc. Arad et al. [11] adopted the K-singular value decomposition (K-SVD) algorithm to process HSI prior and estimated a dictionary for hyperspectral signatures and the corresponding sparse representation. Aeschbacher et al. [12] implemented the reconstruction approach of Arad [11] and introduced a shallow method A+ [13] for better efficiency. Nevertheless, with the gradual improvement of the scale of HSI datasets, the poor expression capacity of this method limits its application in general domains.
In recent years, with the successful application of CNNs in the field of computer vision [14], [15], [16], [17], [18], [19], researchers have carried out a series of explorations of CNN-based model for spectral reconstruction. Galliani et al. [20] trained a Tiramisu-based CNN to learn the end-to-end mapping and achieved the promising results. Xiong et al. [21] preprocessed RGB images through simple interpolation for spectral up-sampling, and adopted the VDSR network [22] for effective spectral reconstruction. Can et al. [23] introduced residual learning into shallow CNN to avoid overfitting in the training stage. Shi et al. [24] first put forward a novel learning network HSCNN-R, which was composed of a set of standard residual blocks. Then, by replacing the residual blocks with dense blocks, they designed a deep network named HSCNN-D to improve the performance further. In the NTIRE 2020 challenge on spectral reconstruction from RGB images [25], various CNN-based algorithms were proposed. Li et al. [26] constructed an adaptive weighted attention network based on a second-order nonlocal module to boost the representational power, which achieved the first and third place results on the “Clean” and “Real World” track, respectively. Zhao et al. [27] built a hierarchical regression network (HRNet) with four levels, and introduced residual dense and global blocks to improve the visual effect. Although satisfactory reconstruction performances have been achieved, there are still some shortcomings in the aforementioned methods. On the one hand, the input of the aforementioned methods can be categorized into two types: the clean images generated by a specific spectral response function or the ground-truth images captured by an RGB camera. The former is apparently noise-free, while the latter only suffers from slight noise pollution. In fact, the capture of RGB images is limited by imaging instruments and atmospheric conditions, and may suffer from Gaussian noise, and the aforementioned methods lack consideration of such case. On the other hand, most of the CNN-based models design deeper network architectures by stacking convolution layers, lacking the full exploration of cross-layer feature information, which restricts the representational capacity of neural networks.

At present, attention mechanism is an excellent and commonly used structure in signal processing field [28], [29], [30], [31], [32], [33], [34], [35]. The main idea of attention mechanism is to achieve the effect of information screening by emphasizing key features and ignoring other unnecessary features so that the value of the important area of the image is enlarged, and the value of the unimportant part is reduced. Especially, the relative attention mechanisms have been applied in the field of HSI processing [30], [31], [32], [33], [34], [35]. In the HSI super-resolution problem, Dong et al. proposed a residual channel attention block to carry out feature extraction work [30]. Yu et al. proposed the feedback spatial attention module and feedback spectral attention module and applied for the HSI classification [39]. These existing channel attention mainly determines which channel information to pay attention to by learning the spectral features of the target HSIs. In order to illustrate it clearly, we give out the main structure of the channel attention as shown in Fig. 1. In addition, some work proposed a self-attention mechanism used on the transformer such as [33], [34], and [35]. The self-attention mechanism is different from channel attention in that it determines which part to focus on and which part to ignore through the relationship between different areas in the input image.

Different from the aforementioned existing attention mechanism, for our spectral reconstruction problem in case of Gaussian noise for HSI, in this article, a novel enhanced channel attention network (ECANet) with fully mining the cross-layer feature information for spectral reconstruction is proposed. The main idea is reflected in two aspects. One aspect is that an enhanced channel attention (ECA) block with dual skip connection and residual connection is proposed. Different from other networks that use cross-layer feature fusion, we use it more frequently to mine useful information in images and reduce the influence of image noise. Specifically, the intrablock residual connection and interblock residual connection at intervals are designed in the ECA block to connect features at different levels while preserving the spatial information of the input image, which effectively ensures the stability of the training process. In another aspect, atrous convolution is used instead of ordinary convolution in the ECA block to expand the feature extraction range, reduce the influence of local noise on training, and enhance the feature extraction ability. Additionally, the utilization of the channel attention mechanism makes the network model the dependence between channels and focus on features conducive to reconstruction. To make full use of the multilevel differences of feature maps at different depths, we proposed a cross-layer feature fusion (CLFF) unit for integral concatenation, so as to generate more representative features. In the reconstruction stage, instead of directly reconstructing the clean HSI, the proposed ECANet uses the global residual learning strategy to estimate the residual feature maps, which helps to reduce the traditional degradation problem in training. Besides, we design a novel loss function combining MSE and SSIM loss as the constraint to further improve the reconstruction accuracy of the proposed ECANet. The effectiveness of the proposed network structure is demonstrated in Experiment.

In general, the main contributions of this article can be enumerated as follows.

1) A novel ECANet for spectral reconstruction in the presence of Gaussian noise is proposed. To our best knowledge, this is the first work in spectral reconstruction that considers additive noise. Specifically, we put forward the ECA block with dual skip connection based on the traditional channel attention block to fully utilize the features of the adjacent blocks. The channel attention mechanism emphasizes the informative features and improves the reconstruction accuracy of the network.
2) Since various image feature information exist in different layers of the ECANet, in order to take full advantage of these feature maps at different depths, we design a CLFF unit to fuse these feature information from the different layers. The proposed CLFF unit further improves the reconstruction accuracy of the network.

3) We explore the advantages of the global residual learning strategy and combined loss function. Furthermore, we evaluate the proposed ECANet on three hyperspectral datasets with Gaussian noise. Compared with the typical spectral reconstruction approaches, the ECANet has achieved promising reconstruction results in visual effects and quantitative indexes.

The rest of this article is organized as follows. The proposed network framework and our ECA module are introduced in detail in Section II. Sections III and IV show the experimental setup, ablation analysis, and experimental results and discussions. Finally, Section V concludes this article.

II. METHODOLOGY

A. Problem Formulation

Generally, to match the perception of three kinds of cone cells in the human visual system, most commercial cameras can merely capture limited information from the visible spectrum, i.e., red, green, and blue channels. The RGB images inevitably show defects in performing various tasks since the indication of the material provided by the images is relatively shallow. On the contrary, the HSIs are rich in spectral information, which can intrinsically reflect the structure and composition of the object. Commonly, assuming that \( X \in \mathbb{R}^{N \times F} \) is an HSI and \( Y \in \mathbb{R}^{N \times 3} \) is the corresponding RGB image, for a given spectral response function \( S \in \mathbb{R}^{F \times 3} \), the RGB image can be generated through a typical process of spectral degradation [36], [37]

\[
Y = XS
\]  

where \( N \) is the number of pixels in each channel, and \( F(N, \lambda) \geq 3 \) is the number of spectral bands in \( X \).

As we can observe from (1), spectral reconstruction from clean RGB images is the inverse process of the aforementioned spectral degradation, which is obviously a severely underconstrained problem. Up to now, scholars have proved that it is indeed possible to address this issue by building a learning or prior-based model to learn a nonlinear mapping to reconstruct HSIs from noise-free RGB images [38]. However, due to the influence of instrument or atmosphere, the capture of RGB images commonly introduces the Gaussian noise. In this case, the spectral degradation from HSI to RGB image can be described as

\[
\tilde{Y} = XS + V
\]  

where \( \tilde{Y} \) represents the noisy RGB image, and \( V = \xi(0, \sigma_n) \) represents the Gaussian noise, which is based on a Gaussian distribution with zero mean and unit variance \( \sigma_n \) to change the noise intensity. The signal to noise ratios (SNRs) of each band in the RGB images are defined as

\[
\text{SNR}_{rgb,i} = 10 \log \left( \frac{\| (XS)_i \|_F^2}{\sigma_n} \right), i = 1, 2, 3. 
\]  

In this case, the goal of our work is to reconstruct a corresponding clean HSI of high quality from a noisy RGB image. That is, we need to learn a function \( F \) so that \( X = F(Y) \) in order to make the error between \( X \) and \( \tilde{X} \) as small as possible. The introduction of the Gaussian noise further increases the difficulty of high-quality spectral reconstruction so that the optimization of the model is particularly significant.

B. Network Architecture

The overall architecture of the ECANet is illustrated in Fig. 2. First, the noisy RGB input image has only three spectral bands, for which we extract the shallow features through a simple convolution layer

\[
G_0 = \mathcal{H}_{FE}(I_{RGB})
\]  

where \( \mathcal{H}_{FE}(\cdot) \) denotes shallow features extraction.

The backbone of the ECANet is stacked with multiple ECA groups for deep feature extraction. Each ECA group consists of a set of ECA blocks in which the attention mechanism is used to redistribute the response weight for each feature channel. The ECA block is an improvement of the typical channel attention block. Concretely, there are two skip connections, respectively, acted on intrablock and adjacent blocks in the ECA block. For the former, it not only allows the block to bypass the rich low-level information from feature maps, but also alleviates the traditional problem of vanishing and exploding gradient. For the latter, it makes the best of the correlation of feature maps and increases the interaction in the adjacent blocks. To exploit these features in different depth layers without direct attenuation, we design CLFF unit to concatenate the multiple feature maps on the spectral dimension, and such CLFF unit also makes the network stable and easy to train. Finally, we utilize global residual learning rather than straightforward prediction to construct the output for the sake of superior spectral reconstruction.

C. Attention Mechanism

The traditional stacked flat structures (e.g., Conv-BN-ReLU connection structure) lack the discrimination between channels in feature maps, which limits the representation capacity of neural networks. Indeed, for the spectral reconstruction in the presence of Gaussian noise, network learning should focus on the significant features, which can be reflected by the differences between channels to a large extent. In other words, the capacity of discrimination learning across channels is quite beneficial to the final spectral reconstruction result. Hu et al. [39] first put forward an architectural unit named squeeze-and-excitation (SE) block that can be regarded as a tool to redistribute the weight using the independencies between spectral bands. Various network modules [26], [38] based on the SE block have been proved to be feasible in the task of spectral reconstruction. However, the structure of SENet formed by stacking SE blocks is still in series from front to back, which does not fully utilize the features of
adjacent blocks. In view of this defect, we propose a novel ECA block with a dual skip connection. The structure of our ECA block is shown in Fig. 2(b). To facilitate the expression of the cross-layer feature fusion, we stack multiple ECA blocks as an ECA group. We weigh the model parameters and stack four ECA groups to form the network backbone. The ECA group is represented as

\[ G_i = H_i(G_{i-1}) = H_i(H_i(...(H_1(G_0)))) \] (5)

where \( G_{i-1} \) and \( G_i \) denote the input and output of the \( i \)-th ECA group, and \( H_i \) denotes the function of \( i \)-th ECA block. The ECA group provides quite large receptive field so as to capture rich contextual information adequately. For a single ECA group, the \( k \)-th ECA block can be represented as

\[ B_k = W_{ECA} \times Z_1 + B_{k-1} \] (6)

where \( B_{k-1} \) and \( B_k \) denote the input and output feature map, respectively, \( W_{ECA} \) is the learned weight vector.

Considering that the distribution of Gaussian noise generally has global characteristics, we first adopt the global average pooling on \( Z_1 \), which can be regarded as the global statistical information of all spatial pixels in a single channel. Then, two stacked convolutional layers with the filter size of \( 1 \times 1 \) are exploited to model the complexity between channels of feature maps. The former is used to downsampling the channel number with ratio \( r \), whereas the latter makes the feature map reverts to the original channel number with the same ratio \( r \). Ultimately, we use the sigmoid activation function to normalize the vector into [0,1]. The aforementioned gating mechanism can be formulated as follows:

\[ W_{ECA} = f(W_u \times \delta(W_d \times H_{GAP}(Z_1))) \] (7)

where \( H_{GAP}(...) \) represents the global average pooling, \( f(...) \) and \( \delta(...) \) denote sigmoid and PReLU activation respectively, and \( W_u \) and \( W_d \) denote the weight set of two stacked convolutional layers, respectively. \( Z_1 \) denotes the production component acquired through

\[ Z_1 = \delta(W_{dil} \times Z_0) \] (8)

where \( W_{dil} \) denotes the weight set of an atrous convolution with a dilation rate of 2. As shown in Fig. 3, the atrous convolution can enlarge the receptive field of the ECA block without increasing the number of parameters. \( Z_0 \) can be acquired through

\[ Z_0 = W_1 \times \delta(W_0 \times B_{k-1}) + R_{k-1} \] (9)
where \( W_0 \) and \( W_1 \) denote the weight set of the first and the second convolution layer in the ECA block, and \( R_{k-1} \) denotes the residual component provided by the previous block.

To introduce more nonlinear and accelerate convergence, the activation function in the ECA block adopts parametric rectified linear unit (PReLU) [40] rather than ReLU, the formula of which is shown as

\[
f(x_i) = \begin{cases} x_i, & \text{if } x_i > 0 \\ ax_i, & \text{if } x_i \leq 0 \end{cases}
\]

\[(10)\]

### D. Cross-Layer Feature Fusion

As shown in Fig. 4, there are various levels of feature information at different depths of the proposed ECANet. To fully exploit these diverse features between ECA groups, we put forward a CLFF unit to concatenate different-level feature maps. The CLFF unit in the proposed model is utilized to merge the multiple feature maps belonging to different ECA groups, as shown in Fig. 2. In addition, the CLFF unit can be considered as a set of skip connections, which has been proved to be feasible for solving the problem of vanishing and exploding gradient. In the field of computer vision, the concatenation of multilevel feature maps has been adopted reasonably for superior performance. Yuan et al. [40] achieved excellent results using concatenated representation for restoration in the task of HSI denoising. In the model of depth map super-resolution, Song et al. [41] exploited a multistage fusion module to reuse features, which efficiently improves the super-resolution performance. The proposed CLFF unit is defined as follows:

\[
F_c = H_{\text{cat}}(G_2, G_3, G_4)
\]

\[(11)\]

where \( H_{\text{cat}}(\cdot) \) denotes the concatenation operation, and \( G_2, G_3, \) and \( G_4 \) denote the output feature maps of different ECA groups, as illustrated in Fig. 2.

Let \( C \) be the number of feature channels, and the combined feature \( F_c \in N \times C \times L \times W \) is restored to the original size through a simple convolution layer

\[
F_{\text{res}} = H_{\text{conv}}(F_c)
\]

\[(12)\]

where \( F_{\text{res}} \in N \times C \times L \times W \) denotes the output of the convolution layer, which is used to reconstruct the noise-free HSI through the global residual connection.

### E. Global Residual Learning Strategy

Rather than the straightforward reconstruction, our ECANet first generates a residual noise \( F_{\text{res}} \), as shown in (12). In the case of the sparsity of the spatial distribution of Gaussian noise in feature maps, the residual mode can provide a smoother hypersurface for gradient descent of the ECANet, which effectively reduces the risk of network degradation. Besides, in terms of the spectral reconstruction task, the global residual learning strategy makes it feasible to ease the training process by increasing the layer depth. The restored clean feature map can be represented as

\[
F_1 = G_0 + F_{\text{res}}
\]

\[(13)\]

where the \( F_1 \) is the output of global residual learning, which is also used for the reconstruction. Ultimately, we reconstruct the noise-free HSI through a \( 3 \times 3 \) convolution layer

\[
I_{\text{SR}} = H_{\text{conv}}(\delta(F_1)).
\]

\[(14)\]

### F. Combined Loss Function

To better optimize the proposed ECANet and preserve high-frequency details, we adopt a weighted combination of the mean-squared error (MSE) and the structural similarity index measurement (SSIM) as the loss function. The combined loss function can be expressed as

\[
L_{\text{SR}} = L_{\text{MSE}} + \lambda L_{\text{SSIM}}
\]

\[(15)\]

where \( \lambda \) is the weighting factors assigned to SSIM loss \( L_{\text{SSIM}} \). The MSE loss is a pixel-wise loss function where each pixel in the reconstructed HSI \( R \) is directly compared to that of in the ground-truth HSI \( X \). The MSE loss can be represented as

\[
L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (R_i - X_i)^2
\]

\[(16)\]

where \( N \) is the number of pixels in the HSI, \( R_i \) and \( X_i \) represent the \( i \)th pixel value of the recovered and the ground-truth HSI, respectively. Furthermore, the SSIM loss is described as

\[
L_{\text{SSIM}} = \frac{1}{N} \sum_{i=1}^{N} (1 - \text{SSIM}(R_i, X_i))
\]

\[(17)\]

where \( \text{SSIM}(\cdot) \) denotes the SSIM value between the reconstructed and the ground-truth HSI, and we further provide its calculation formula

\[
\text{SSIM}(R_i, X_i) = \frac{(2\mu_{R_i}\mu_{X_i} + c_1)(\sigma_{R_i}\sigma_{X_i} + c_2)}{(\mu_{R_i}^2 + \mu_{X_i}^2 + c_1)(\sigma_{R_i}^2 + \sigma_{X_i}^2 + c_2)}
\]

\[(18)\]

For the weighted factors in the combined loss function in (15), we set \( \lambda \) to 0.1 according to experience.
and decays by 40% every 600 for testing. As shown in Table II, when the ϵ β is 0.1, the effect is the best. Our proposed ×5, 10, and 25. For each RGB and HSI pair in the and normalize them to range [0, 1]. The batch size is S ECAN Xmax(R) = log10 R denote the reconstructed and E of the two L×λ 512 F OF THE convolution layers is set to 8. In addition, UANTITATIVE The root mean square error 64. In the ECA block used as quantitative measures to evaluate the performance of 3]. Specifically, AWGN is applied to the simulated RGB images is added to the noise-free RGB images with the noise level 512 small amount of unknown noise), of which the spatial size is World” (unknown spectral response function and including a training images and 10 test images. The dataset contains RGB datasets to evaluate our ECANet, i.e., CA VE [42] A. Experimental Setup 1) Hyperspectral Datasets: In this article, we use two public hyperspectral datasets to evaluate our ECANet, i.e., CA VE [42] and NTIRE 2020 [11]. The CA VE dataset consists of 32 HSIs collected by the Apogee Alta U260 camera with a spatial resolution of 512 × 512. All images have 31 channels in the wavelength range from 400 to 700 nm at 10-nm intervals. We use the CIE 1964 color matching functions to simulate the corresponding RGB images. We randomly extract 22 image pairs for training and the remaining ten image pairs for testing. In the training phase, a set of patches with a spatial size of 64, as input, are extracted from the RGB images. There is an overlap of four pixels between adjacent patches to avoid possible artifacts on the border. Furthermore, we expanded training samples by dicing the HSI and rotating it, flipping it horizontally, etc. The other public images in NTIRE 2020 dataset captured by the Specim IQ mobile hyperspectral camera include: 450 training images and 10 test images. The dataset contains RGB images of “Clean” (known spectral response function) and “Real World” (unknown spectral response function and including a small amount of unknown noise), of which the spatial size is 512 × 482. Besides, the additive white Gaussian noise (AWGN) is added to the noise-free RGB images with the noise level σn set to 5, 10, and 25, where n denotes a generic band n ∈ [1, 2, 3]. Specifically, AWGN is applied to the simulated RGB images for CA VE dataset and the “Clean” RGB images in NTIRE 2020 dataset. 2) Quantitative Measures: The root mean square error (RMSE), peak signal to noise ratio (PSNR), mean relative absolute error (MRAE), and spectral angle mapper (SAM) are used as quantitative measures to evaluate the performance of the proposed spectral reconstruction method. RMSE represents the pixel-level error between the recovered and the ground-truth images. PSNR is an objective image evaluation metrics, but an inaccurate representation of human perception. Hence, MRAE is also provided for a better comparison. Besides, SAM adopted in this dissertation is used to estimate the spectral distortion between the reconstructed and the ground-truth HSI. The quantitative measures RMSE, PSNR, MRAE, and SAM are calculated as (19) and (20), where R and X denote the reconstructed and the ground-truth HSI, respectively.

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (R_i - X_i)^2}, \quad \text{SAM} = \arccos \left( \frac{<R, X>}{|R||X|} \right) 
\]

(19)

\[
\text{PSNR} = 10 \log_{10} \left( \frac{\max(X)^2}{\text{MSE}(R, X)} \right), \quad \text{MRAE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|R_i - X_i|}{X_i} 
\]

(20)

3) Parameter Setting: For each RGB and HSI pair in the training dataset, we randomly crop a set of patches of size 64 × 64 and normalize them to range [0, 1]. The batch size is set to 64. The ECANet is trained for 3000 epochs overall. The initial learning rate is 2 × 10^{-4} and decays by 40% every 600 epochs. We adopt Adam optimizer with β1 = 0.9, β2 = 0.999, and ϵ = 10^{-8}. As for the loss function, we tried to take different values of parameter λ for testing. As shown in Table II, when the value of parameter λ is 0.1, the effect is the best. Our proposed ECANet is implemented on the PyTorch framework using an NVIDIA RTX 3090 GPU, and the training time of a typical model is approximately 48 h.

B. Ablation Analysis for Different Modules of the Proposed Network

To verify the effect of different modules, we carry out ablation analysis for different modules of the proposed network on the CA VE dataset with noise intensity σn = 5, 10, and 25.
The network framework of the baseline method is composed of stacked residual blocks, the depth of which is equal to the proposed ECANet. The detail is given in Table III.

1) **Attention Mechanism:** Different from the residual blocks or dense blocks commonly used in convolutional neural networks, the proposed ECA block greatly improves the performance in solving the problem of spectral reconstruction. With the introduction of the attention mechanism to the baseline, the experimental results compared with the only using baseline decrease RMSE by 0.1762, 0.2526, and 0.1910, respectively, and increase PSNR by 0.73, 0.88, and 0.79 dB, respectively, which demonstrates the feasibility of the proposed ECA block.
2) Cross-Layer Feature Fusion: In order to show the effect of cross layer feature fusion model, the experimental results of the model “Baseline+ECA+CLFF” compared with the model “Baseline+ECA” decreases 0.1035, 0.0862, and 0.0340 in RMSE, and increase 0.09 dB, 0.20 dB, and 0.24 dB in PSNR with the noise level $\sigma_n = 5, 10,$ and 25, respectively.

3) Combined Loss Function: Furthermore, the proposed loss function can improve the reconstruction performance to a certain extent. The extra experiment shows that the model “Baseline+ECA+CLFF+L_{SR}” compared with the model “Baseline+ECA+CLFF” decreases 0.0650, 0.0604, and 0.0578 in RMSE, and increases 0.13 dB, 0.16 dB, and 0.14 dB in PSNR with the noise level $\sigma_n = 5, 10,$ and 25, respectively.

C. Ablation Analysis for Different Parameters
To verify the effect of the different network parameters including the number of ECA blocks in each ECA group and

### TABLE V
Quantitative Comparison With Other Three Methods on the CAVE and NTIRE 2020 Datasets

| Dataset         | Noise level | Index | Method      | Sparse Coding [11] | HSCNN-R [24] | HRNet [27] | DsTer [34] | Proposed |
|-----------------|-------------|-------|-------------|--------------------|--------------|-------------|------------|----------|
| CAVE [42]       | $\sigma_n = 5$ | RMSE  | 5.1650      | 3.4603             | 3.3165       | 3.2874      | 3.1123     |
|                 |             | MRAE  | 0.8476      | 0.2495             | 0.4811       | 0.4019      | 0.3741     |
|                 |             | PSNR  | 30.54       | 35.04              | 35.24        | 35.37       | 35.93      |
|                 |             | SAM   | 11.4811     | 6.4547             | 6.1756       | 5.9214      | 5.6456     |
|                 | $\sigma_n = 10$ | RMSE  | 6.8636      | 3.7353             | 3.5022       | 3.4823      | 3.3312     |
|                 |             | MRAE  | 1.2464      | 0.2961              | 0.5462       | 0.4517      | 0.4243     |
|                 |             | PSNR  | 28.24       | 34.10              | 34.32        | 34.58       | 35.19      |
|                 |             | SAM   | 14.3471     | 7.0001             | 6.8007       | 6.5723      | 6.0583     |
|                 | $\sigma_n = 25$ | RMSE  | 8.4646      | 4.5027             | 4.4317       | 4.3812      | 4.2085     |
|                 |             | MRAE  | 1.5871      | 0.3808             | 0.6671       | 0.4219      | 0.3794     |
|                 |             | PSNR  | 26.42       | 31.63              | 32.18        | 32.24       | 32.72      |
|                 |             | SAM   | 16.9929     | 9.3395             | 8.5447       | 8.3142      | 8.0137     |
| NTIRE 2020 [25] | $\sigma_n = 5$ | RMSE  | 4.5733      | 2.9371             | 2.8769       | 2.8021      | 2.6600     |
|                 |             | MRAE  | 0.1122      | 0.0720             | 0.0651       | 0.0673      | 0.0642     |
|                 |             | PSNR  | 34.1024     | 37.8348            | 38.0370      | 38.2714     | 39.1384    |
|                 |             | SAM   | 5.5520      | 2.9525             | 2.8799       | 2.8142      | 2.3387     |
|                 | $\sigma_n = 10$ | RMSE  | 5.7042      | 3.3664             | 3.4856       | 3.3218      | 3.3017     |
|                 |             | MRAE  | 0.1587      | 0.0867             | 0.0805       | 0.0822      | 0.0804     |
|                 |             | PSNR  | 31.7134     | 36.1735            | 36.3539      | 36.5917     | 36.8447    |
|                 |             | SAM   | 7.2518      | 3.6830             | 3.4242       | 3.3899      | 3.3248     |
|                 | $\sigma_n = 25$ | RMSE  | 7.0283      | 4.5143             | 4.4927       | 4.4015      | 4.3328     |
|                 |             | MRAE  | 0.2083      | 0.1171             | 0.1105       | 0.1162      | 0.1146     |
|                 |             | PSNR  | 29.7738     | 33.2942            | 33.5190      | 33.5526     | 33.7261    |
|                 |             | SAM   | 9.0049      | 5.1577             | 4.8301       | 4.8066      | 4.6917     |
| Real World      | $\sigma_n = 5$ | RMSE  | -           | 3.1430             | 2.7640       | 2.7362      | 2.5118     |
|                 |             | MRAE  | -           | 0.0738             | 0.0620       | 0.0613      | 0.0597     |
|                 |             | PSNR  | -           | 36.87              | 38.07        | 38.25       | 38.96      |
|                 |             | SAM   | -           | 3.1135             | 2.7282       | 2.6844      | 2.3741     |

### TABLE VI
Quantitative Comparison With Other Three Methods on the Image Feathers _ms With Uniformly Distributed Noise and Poisson Noise

| Images        | Noise parameters | Index | Method      | Sparse Coding [11] | HSCNN-R [24] | HRNet [27] | DsTer [34] | Proposed |
|---------------|------------------|-------|-------------|--------------------|--------------|-------------|------------|----------|
| feathers_ms   | $[-10, 10]$      | RMSE  | 6.9756      | 3.7838             | 3.0701       | 3.6055      | 3.4267     |
|               |                   | MRAE  | 1.3623      | 0.3062             | 0.5986       | 0.4823      | 0.4402     |
|               |                   | PSNR  | 27.64       | 33.92              | 34.12        | 34.51       | 35.02      |
|               |                   | SAM   | 15.4512     | 7.6970             | 7.0227       | 6.8174      | 6.1420     |
|               | $[-25, 25]$       | RMSE  | 8.8421      | 4.7590             | 4.5601       | 4.4912      | 4.3102     |
|               |                   | MRAE  | 1.6008      | 0.4010             | 0.6853       | 0.4513      | 0.3802     |
|               |                   | PSNR  | 25.21       | 30.23              | 31.82        | 32.06       | 32.61      |
|               |                   | SAM   | 17.0121     | 9.4001             | 8.6127       | 8.4511      | 8.0489     |

### Poisson noise

| Images        | Poisson noise peak | Index | Method      | Sparse Coding [11] | HSCNN-R [24] | HRNet [27] | DsTer [34] | Proposed |
|---------------|-------------------|-------|-------------|--------------------|--------------|-------------|------------|----------|
| feathers_ms   | 2                 | RMSE  | 5.3723      | 3.5341             | 3.3814       | 3.3513      | 3.2417     |
|               |                   | MRAE  | 1.0344      | 0.2728             | 0.5102       | 0.4354      | 0.4015     |
|               |                   | PSNR  | 29.6701     | 34.39              | 34.76        | 34.81       | 35.35      |
|               |                   | SAM   | 13.8810     | 7.7917             | 7.2982       | 7.1502      | 6.8316     |
the size of kernel, we carry out ablation experiment on the CAVE dataset with noise intensity $\sigma_n = 10$. From the result shown in Table IV, it is easy to find that the performance of the reconstruction improves as the number of ECA blocks increases. However, when the number of ECA blocks becomes larger than 8, it hardly promotes the performance, but increases a lot of training time. When the number of ECA blocks in each ECA group $K = 8$ and the size of kernel $= 3 \times 3$, it obtains the best result. This conclusion holds for NTIRE 2020 dataset as well.

IV. RESULTS AND DISCUSSIONS

We demonstrate the superiority of the proposed ECANet method from both objective and subjective aspects. We first evaluate the reconstruction performance of the proposed method and other mainstream methods through typical quantitative indexes. Specifically, we selected four representative models Sparse Coding [11], HSCNN-R [24], HRNet [27], and DsTer [34] in the field of hyperspectral reconstruction in recent years to compare with our method. Also, we provide the absolute error maps of the aforementioned methods to measure the detail reconstruction and noise reduction level of the methods.

A. Quantitative Evaluation

Table V provides the comparison of the average indexes of the proposed ECANet reconstruction method and the other four methods on CAVE and NTIRE 2020 datasets. The table uses bold and underline in order to sign the best and second best results.

From the quantitative results of experiments, it can be seen that the proposed method has achieved satisfactory results in the given quantitative indexes. For CAVE dataset, when $\sigma_n = 5$, 10, and 25, the SNR of the original image is 22.34, 20.68, and 17.45 dB, respectively. The PSNR of the reconstruction results of our method is 35.93, 35.19, and 32.72 dB, respectively. It is easy to find that the performance improvement of the reconstruction is very significant. As shown in Table V, in case of different noise levels, the four CNN-based methods are obviously better than the sparse coding method, which benefits from the powerful context information utilization capacity of neural networks. Among the
CNN-based methods, the reconstruction index of the proposed ECANet method is obviously superior to the other three methods. The reconstruction index of the proposed ECANet method is almost all-around ahead for all noise levels. Specifically, in the case of CAVE dataset, the proposed method is significantly better than the second best method DsTer in indicator PSNR with an average improvement of +0.55 dB. Furthermore, the RMSE, MRAE and SAM are decreased in average −0.1663, −0.0326, and −0.3484 by our method compared with the DsTer, respectively. For the NTIRE 2020 dataset, the proposed method is also significantly better than DsTer method in the all indicators, except that the MRAE when $\sigma_n = 25$ is slightly worse than HRNet. Furthermore, considering the universality of the model, two different kinds of noise are also added to the noise-free RGB image feathers _ ms on CAVE dataset for test. One is the uniform white noise with two different distribution parameters $[-10, 10]$ and $[-25, 25]$, and the other is Poisson noise. As shown in (21), we apply “poissrnd” function in MATLAB to add Poisson noise to the image of data type uint8.

$$I_{\text{noise}} = \text{poissrnd}(I_{\text{clean}} * p) / p$$

(21)

where $I_{\text{clean}}$ and $I_{\text{noise}}$ represent the clean image and the noised image, respectively, and $p$ represents the Poisson noise peak, which is set to 2 in our experiment. The indexes of the spectral reconstruction of several methods are shown in Table VI. It can be seen that the proposed method also obtains better results than the other methods in case of different noise.

B. Visual Results

To show our high fidelity of the reconstruction result with the test dataset, we visualize the reconstruction result of each method by the error image, which is the heat maps between the ground truth and the recovered HSI to display glaring textures as shown in Figs. 5–7. The error map of the CAVE dataset adopts the 18th band of “feathers _ ms,” and the 12th band of “watercolors _ ms,” and the error map of NTIRE 2020 dataset adopts the 11th band of “ARAD _ HS_0464” and 14th band of “ARAD _ HS_0462.” From Figs. 5 and 6, it is easy to see that our method is more superior to other methods. In detail, the results reconstructed by the sparse coding method have a poor visual effect and contain strong noise, reflecting its weak antinoise capacity. This is because the nonlinear mapping of sparse coding is pixel level, and the noise information in images cannot be corrected by accurate context information. The visual effects of the reconstructed results by the three methods HSCNN-R, HRNet, and DsTer are slightly better than sparse coding method, but the restoration of some detailed textures is insufficient, such as the feather texture of the “feathers _ ms” and the edge details of the “watercolors _ ms.” However, the proposed ECANet shows obvious advantages in reconstruction results. In case of heavy noise intensity, the visual effect of our method is still best. So, it can obviously be seen that the proposed method outperforms the comparison methods.

Besides, we utilize the “Real World” RGB images in the NTIRE 2020 dataset for comparison, as shown in Fig. 7. The reconstruction results of the HSCNN-R and the HRNet methods
have low-fidelity information. However, the proposed method restores HSIs well and shows more satisfactory visual effect and better reconstruction fidelity. It further proves good anti-noise capacity of the proposed method for real-world spectral reconstruction. In general, the ECANet method is superior to the other methods.

V. CONCLUSION

In this article, we have proposed an ECANet method to recover HSIs from RGB images in the presence of Gaussian noise. First, based on the traditional channel attention block, we have designed the ECA block, in which the dual residual learning is utilized to capture the information of adjacent blocks, and the atrous convolution is used to improve the efficiency. Considering that the noise levels of feature maps in different depth layers are different, we use the CLFF unit to aggregate the hierarchical information to generate more representative features. In addition, the combined loss function is used to optimize the ECANet to improve the reconstruction fidelity. We have analyzed the effectiveness of the proposed components through ablation experiments. The ECANet method demonstrates the improvement over the results of the other representative spectral reconstruction methods in both subjective and objective evaluations. In our future work, we will focus on the spectral reconstruction in the presence of mixed noise in input RGB images, such as stripe noise and pulse noise.

REFERENCES

[1] G. Tochon, J. Chanussot, M. Dalla Mura, and A. L. Bertozzi, “Object tracking by hierarchical decomposition of hyperspectral video sequences: Application to chemical gas plume tracking,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 8, pp. 4567–4585, Aug. 2017.
[2] Z. Li, F. Xiong, J. Zhou, J. Wang, J. Lu, and Y. Qian, “BAE-Net: A band attention aware ensemble network for hyperspectral object tracking,” in Proc. IEEE Int. Conf. Image Process., 2020, pp. 2106–2110.
[3] P. Sindhi, C. Chen, V. Asari, Y. Xu, and W. Li, “Classification of hyperspectral image using multiscale spatial texture features,” in Proc. Workshop Hyperspectral Image Signal Process., Evol. Remote Sens., 2016, pp. 1–4.
[4] S. Zhong, C. Chang, and Y. Zhang, “Iterative support vector machine for hyperspectral image classification,” in Proc. IEEE Int. Conf. Image Process., 2018, pp. 3309–3312.
[5] A. V. Kartha and V. A. Bharadi, “Face recognition using orthogonal transform coefficients of hyperspectral face images,” in Proc. Int. Conf. Inf. Process., 2015, pp. 349–354.
[6] G. Y. Chen, W. F. Xie, S. P. Wang, and H. K. Liu, “Hyperspectral face recognition with minimum noise fraction, log-polar Fourier features and collaborative representation-based classifier,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2016, pp. 809–812.
[7] T. Harakawa, J. Chanussot, and J. A. Benediktsson, “Segmentation and classification of hyperspectral images using watershed transformation,” Pattern Recognit., vol. 43, no. 7, pp. 2367–2379, Jul. 2010.
[8] T. Okamoto and I. Yamaguchi, “Simultaneous acquisition of spectral image information,” Opt. Lett., vol. 16, no. 16, pp. 1277–1279, 1991.
[9] A. A. Wagadarikar, N. P. Pitsianis, X. Sun, and D. J. Brady, “Video rate spectral imaging using a coded aperture snapshot spectral imager,” Optics Express, vol. 17, no. 8, 2009, pp. 6368–6388.
[10] Y. Liu, Y. Liu, I. Wu, and Q. Bai, “Spatial–spectral encoded compressive hyperspectral imaging,” ACM Trans. Graph., vol. 33, no. 6, pp. 233:1–233:11, 2014.
[11] B. Arad and O. Ben-Shahar, “Sparse recovery of hyperspectral signal from natural RGB images,” in Proc. IEEE Eur. Conf. Comput. Vis., 2016, pp. 19–34.
[12] J. Aeschbacher, J. Wu, and R. Timofte, “In defense of shallow learned spectral reconstruction from RGB images,” in Proc. IEEE Int. Conf. Comput. Vis. Workshops, 2017, pp. 471–479.
[38] J. Li et al., “Hybrid 2-D–3-D Deep residual attentional network with structure tensor constraints for spectral super-resolution of RGB images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2321–2335, Mar. 2021.

[39] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[40] Q. Yuan, Q. Zhang, J. Li, H. Shen, and L. Zhang, “Hyperspectral image denoising employing a spatial C-Spectral deep residual convolutional neural network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1205–1218, Feb. 2019.

[41] X. Song et al., “Channel attention based iterative residual learning for depth map super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5630–5639.

[42] F. Yasuma, T. Mitsunaga, D. Iso, and S. K. Nayar, “Generalized assorted pixel camera: Postcapture control of resolution, dynamic range, and spectrum,” *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2241–2253, Sep. 2010.

Changwu Zou received the Ph.D. degree in mathematical sciences from Peking University, Beijing, China, in 2017.

He is currently an Associate Professor with the College of Mathematics and Statistics, Fuzhou University, Fuzhou, China. His research interests include pattern recognition and data analyzing.

Can Zhang received the B.S. degree in software engineering, in 2021, from Fuzhou University, Fuzhou, China, where he is currently working toward the M.S. degree with the College of Computer and Data Science.

His research interests are in the areas of image understanding and hyperspectral image reconstruction.

Minghui Wei received the B.S. degree in software engineering from Fuzhou University, Fuzhou, China, in 2019 and the M.S. degree in computer application from the College of Computer and Data Science, Fuzhou University, in 2022.

His research interests include the areas of computer vision and image super-resolution.

Changzhong Zou received the Ph.D. degree in computer science from Fuzhou University, Fuzhou, China, in 2018.

He is currently an Associate Professor with the College of Computer and Data Science, Fuzhou University. His research interests include deep learning, image understanding and analyzing, and hyperspectral image processing.