HyperFLFN: An Image Super-Resolution Model Based on Attention Mechanism and Meta-Learning

Nianze Du, and Ji Zhao*

Abstract—**Image super-resolution, a critical task in the field of computer vision, aims to restore high-quality, high-resolution images from low-resolution inputs. Traditional interpolation methods perform adequately in some scenarios but suffer significant performance degradation when dealing with images containing intricate textures. Deep learning-based models, on the other hand, often exhibit poor super-resolution effects, long training times, and weak generalization performance. This paper proposes a novel image super-resolution model named HyperFLFN. The model adopts an innovative structure by introducing a self-attention mechanism into traditional residual connections, resulting in a new Conformer module. This module can simultaneously learn global and local features, thereby capturing image details and contextual information more effectively during the super-resolution reconstruction process. Compared to traditional methods, the HyperFLFN model achieves significant improvements in enhancing image super-resolution quality and preserving image details. Additionally, this paper introduces a meta-learning mechanism to enhance the model's generalization capability. Extensive experiments conducted on multiple datasets demonstrate that HyperFLFN achieves an average improvement of 4.94% in PSNR and SSIM compared to other models. This study demonstrates that the HyperFLFN model achieves excellent performance across various scenarios. Notably, the model exhibits strong generalization capabilities when handling different types of images, highlighting its potential and feasibility in practical applications. In conclusion, the proposed HyperFLFN model not only makes significant progress in the image super-resolution task but also shows great potential in improving image quality and generalization capabilities. This research provides new insights and methods for further development in the field of image processing.**

Index Terms—**Image Super-Resolution, Deep Learning, Self-Attention Mechanism, RLFN**

I. Introduction

With the rapid advancement of computer vision and image processing, image super-resolution has become a significant and well-studied research direction. The primary goal of image super-resolution is to enhance the visual quality and detail of images by restoring high-resolution images from low-resolution counterparts. In recent years, various methods have emerged in the field of image super-resolution, including traditional interpolation-based methods and deep learning-based methods.

Interpolation-based methods perform simple mathematical operations, such as nearest-neighbor interpolation[1], bilinear interpolation[2], and bicubic interpolation[3]. In bilinear interpolation, the value of each target pixel is determined by the weighted average of the four nearest pixels in the low-resolution image, typically the four closest pixels surrounding the target pixel. Bicubic interpolation, another common interpolation method, considers more neighboring pixels when calculating pixel values, thereby better preserving image details and textures. However, these methods generally fail to capture high-frequency details of the image, leading to limited effectiveness in some scenarios.

Deep learning-based image superresolution methods have become a major direction of development in recent years. These methods utilize deep neural networks to learn the mapping relationship between low-resolution and high-resolution images, thereby achieving image super-resolution reconstruction. Classical examples include SRCNN (Super-Resolution Convolutional Neural Network)[4], which uses three convolutional layers to learn the mapping from low-resolution to high-resolution images; ESPCN (Efficient Sub-Pixel Convolutional Neural Network)[5], which employs sub-pixel convolution layers to enhance network efficiency; SRGAN (Super-Resolution Generative Adversarial Network)[6], which incorporates adversarial network mechanisms to improve image reconstruction quality; and RLFN (Residual Local Feature Network)[7], which combines global and local features for image super-resolution.

Despite their success, deep learning-based image super-resolution models share some common shortcomings[8–10]. First, these models typically require substantial computational resources and memory for training and operation, especially for deep networks or large datasets, making them difficult to deploy in resource-constrained environments. Secondly, their generalization capability is often poor; some models may perform inadequately when handling images of different types, scales, or noise levels, potentially due to insufficient consideration of image diversity during training. Lastly, while some models perform well in restoring global structures and main features of images, they often struggle with complex textures, fine details, or local structures, leading to unnatural or distorted images in certain cases.

To address these challenges, many researchers are focusing on novel breakthroughs such as attention mechanisms and adversarial networks. Therefore, this paper proposes a new image super-resolution model named HyperFLFN, characterized by the following features:

1) Integration of an attention mechanism to enhance the signal-to-noise ratio between the original and super-resolved images;

2) Stacking multiple residual modules to improve computational speed and the similarity in brightness and contrast between the original and super-resolved images;

3) Simultaneous extraction of global and local information, effectively enhancing the quality of the super-resolved images;

Manuscript received Jun 3, 2024; revised Dec 21, 2024.

This work was supported by the Special Fund for Scientific Research Construction of University of Science and Technology Liaoning.

Nianze Du is a Postgraduate of University of Science and Technology Liaoning, Anshan, Liaoning, China. (e-mail: bengbeng315@163.com).

Ji Zhao* is a Professor of University of Science and Technology Liaoning, Anshan, Liaoning, China. (corresponding author to provide phone: +086-139-9808-6167; e-mail: 319973500069@ustl.edu.cn).

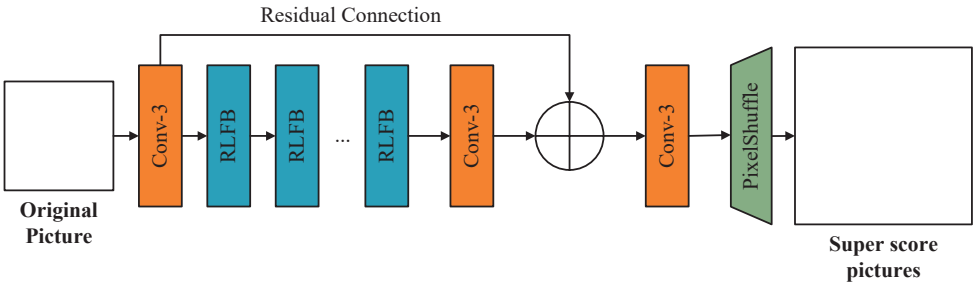

Fig. 1. The Structure of Residual Local Feature Network

4) Automatic calculation of image gradients to improve training efficiency.

II. Related Work

A. Residual Local Feature Network

RLFN (Residual Local Feature Network) is a deep learning-based image super-resolution model. Its goal is to effectively restore high-resolution details and textures in images while reducing blurriness and distortion. The core concept of this model is to combine global and local features for image super-resolution. It enhances the quality of image reconstruction by incorporating residual learning and local feature extraction modules. The structural diagram of the model is shown in Fig.1.

The core components of RLFN include the Global and Local Feature Extractor, the Residual Learning Module, and the Fusion Module. The Local Feature Extractor aims to extract local detail features from the input low-resolution image, typically implemented using deep neural networks composed of convolutional and pooling layers. The Global Feature Extractor is designed to capture the global information of the input image, helping to preserve the overall structure and content of the image, thereby avoiding over-processing and distortion.

The Residual Learning Module is a key component of the RLFN model. It receives local feature representations from the Local Feature Extractor and connects them with the input image through residual connections. This helps the model learn the subtle differences in the image, thereby improving the accuracy of super-resolution reconstruction. The Fusion Module is used to combine local and global features to comprehensively consider the local details and global structure of the image, typically achieved through operations such as convolution and up-sampling.

By adopting the concept of residual learning, the RLFN model effectively preserves the detail information of the input image, resulting in super-resolution reconstructions that are clearer and more natural[11]. By simultaneously considering global and local information, the RLFN model can better understand the structure and content of the image, thereby enhancing the accuracy and quality of the reconstruction.

B. Residual Local Feature Block

The core concept of the Residual Learning Module is to introduce residual connections by adding the feature maps extracted by the Local Feature Extractor to the input image. This design allows the model to directly learn the residuals between the input image and the target high-resolution image, rather than directly learning the mapping from low resolution to high resolution[12–14]. This helps the model to better capture image details, thereby improving the quality of reconstruction.

By stacking multiple Residual Learning Modules, the Residual Local Feature Block (RLFB) is obtained. After the residual connections, one or more convolutional layers and activation functions are typically applied to further process the feature maps. These convolutional layers usually adopt a shallow network structure to reduce computational complexity and decrease the number of parameters. The specific structure is shown in Fig. 2. The design of these layers aims to further extract local features of the image and provide richer information for subsequent fusion. RLFB uses only a few stacked CONV+RELU layers for local feature extraction. Specifically, each feature refinement module in the RLFB includes a 3×3 convolutional layer followed by a ReLU activation function layer. Given the input feature F_{in}, the entire structure is described in Equation (1).

$$F_{refined_i} = RM_i(F_{refined_{i-1}}) \tag{1}$$

In this context, RM_i represents the i-th refinement module, and $F_{refinedi}$ is the result of the i-th refinement module. Subsequently, following the method of RFDB, $F_{refined}$ is fed into a 1×1 convolutional layer and a subsequent ESA block to obtain the final output.

C. Pixel Shuffle

Pixel Shuffle is a commonly used up-sampling technique[15], widely applied in image super-resolution tasks, including deep learning models such as RLFN. The core idea of PixelShuffle is to achieve up-sampling by rearranging the pixel values in the feature map. Specifically, it divides the feature map into several small blocks and then rearranges the pixel values within each block such that adjacent pixel values in the output feature map are separated by a number of pixel positions. Consequently, by stacking these rearranged small blocks, an output feature map with increased resolution is obtained.

Firstly, the input feature map is divided into small blocks of size $r \times r$, where r is the up-sampling factor (typically 2 or 3). Then, the pixel values within each block are rearranged so that adjacent pixel values in the output feature map are separated by r pixel positions. Specifically, if a block contains $n \times n$ pixels, these pixels will be rearranged in the output feature map into a block of size $\frac{n}{r} \times \frac{n}{r}$. Finally, all the rearranged blocks are stacked together to form the final output feature map. This process achieves the up-sampling of the feature map, thereby increasing its resolution.

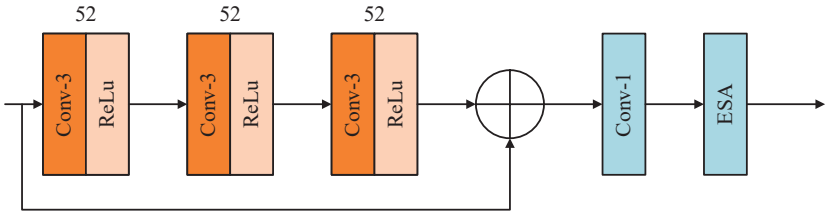

Fig. 2. The Structure of Residual Local Feature Block

III. Method

This paper proposes a novel model to better address the task of image super-resolution, referred to as HyperFLFN, the specific structural diagram of which is illustrated in Fig. 3.

A. Self-Attention Mechanism

The self-attention mechanism improves image super-resolution quality by learning the relationships between different positions within an image, enabling the model to selectively focus on important features and better preserve the image's structure and content during reconstruction[16].

After applying the linear transformation matrix $W_L \in \mathbb{R}^{L\times N}$ to transform the original data, this mechanism adds a position-related vector to the input data by assigning different encodings to each position and each embedded dimension. A common method for position encoding uses sine and cosine functions. For each position pos and each embedded dimension i, the position encoding $\mathrm{PE}_{(pos,i)}$ can be calculated according to Equation (2):

$$\begin{aligned} PE_{(pos,\ \ 2i)} &= \sin\left(\frac{pos}{1000^{\frac{2i}{d_k}}}\right) \\ PE_{(pos,\ \ 2i+1)} &= \cos\left(\frac{pos}{1000^{\frac{2i}{d_k}}}\right) \end{aligned} \tag{2}$$

In these equations, pos is the current position in the entire input sequence and i is the dimension index after linear transformation. After position encoding, the input for the self-attention mechanism, $X_{\text{embedding}}$, is obtained. Following the matrix transformation and position encoding, the Transformer model uses learnable weight matrices $W_Q \in \mathbb{R}^{D\times d_k}$, $W_K \in \mathbb{R}^{L\times d_k}$, and $W_V \in \mathbb{R}^{L\times d_k}$ to convert each input X_i (i.e., the position-encoded $X_{\text{embedding}}$) into three matrices: Query (Q), Key (K), and Value (V), where d_k represents the embedding dimension. The conversion process is described in Equation (3):

$$\begin{aligned} Q_j &= W_Q^j X_i^T,\ j = 1,\ 2,\ \ldots,\ 8 \\ K_j &= W_K^j X_i^T,\ j = 1,\ 2,\ \ldots,\ 8 \\ V_j &= W_V^j X_i^T,\ j = 1,\ 2,\ \ldots,\ 8 \end{aligned} \tag{3}$$

After obtaining Q, K, and V, the model calculates the attention weights. For each position, it computes the dot product of Q with all K, scales the dot products, and obtains attention scores. The scores are then normalized using the Softmax function to obtain the attention weights, which is the core of the self-attention mechanism and the Transformer model, as shown in Equation (4):

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

The advantage of the self-attention mechanism lies in its ability to establish relationships between different positions within an image, independent of the image size, thus capturing dependencies effectively.

B. Conformer Block

The task of image super-resolution requires the utilization of both global and local features of the image. Local features are compact vector representations of local image neighborhoods, while global features include contour representations, shape descriptions, and long-range feature representations. In the original RLFB within RLFN, only convolutional layers were used to extract features, and convolutional layers have limited capability for extracting global features[17].

To leverage both local features and global representations, this paper designs the Conformer Block to replace the CONV+RELU layers in RLFB. The structure is illustrated in Fig. 4. Specifically, the method involves feeding the global features from the self-attention mechanism branch into the convolutional layers to enhance the global perception capability of the convolutional branch. Similarly, the local features from the CNN branch are fed into the self-attention mechanism to enhance the local perception capability of the self-attention mechanism. Through this process, the Conformer Block simultaneously acquires both global and local features of the image.

C. ESA Block

Focusing on edge features is also crucial. The Edge-Selective Attention (ESA) module is a key attention mecha-

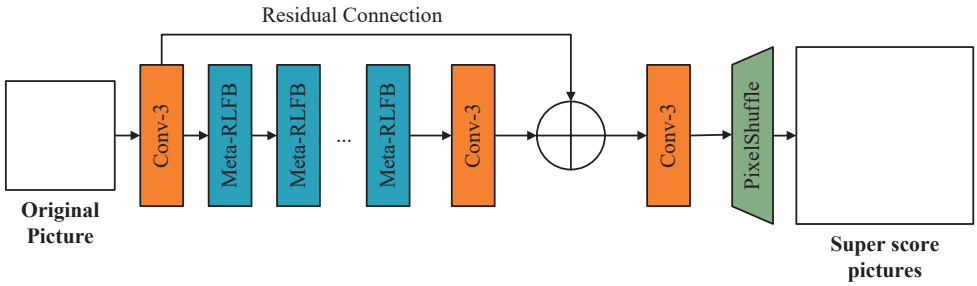

Fig. 3. The Structure of HyperFLFN

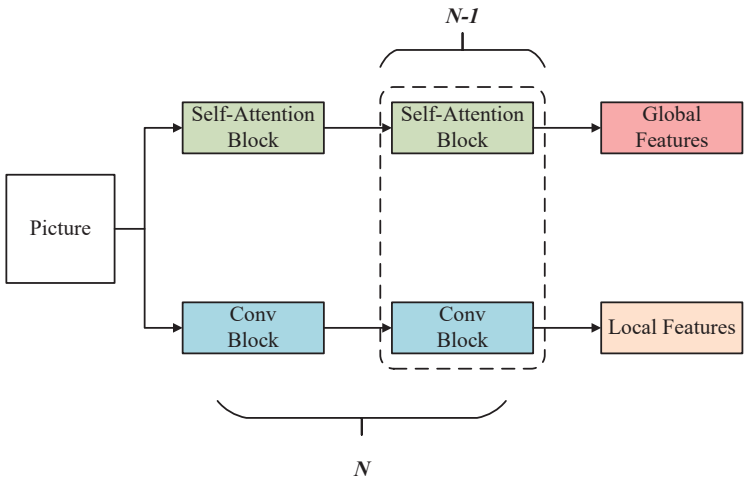

Fig. 4. The Structure of Conformer

nism designed to selectively focus on important edge features within local regions, thereby enhancing the model's ability to perceive image details and structures[18]. Its structural diagram is shown in Fig. 5.

Firstly, the ESA module receives feature maps from the previous layer as input, typically local features extracted by a convolutional neural network. Subsequently, the ESA module employs a set of convolutional filters to extract edge features from the image. These filters are usually designed as sharpening filters or edge detection filters to capture high-frequency edge information in the image. Next, the ESA module calculates the edge responsiveness at different positions in the image through average pooling to determine which positions possess significant edge features. Based on the computed edge responsiveness, the ESA module weights the original feature maps, emphasizing important edge features and suppressing information from non-edge regions. This is usually achieved through element-wise multiplication of the original feature maps. Finally, the ESA module fuses the weighted feature maps with the original feature maps to obtain the final feature representation. This fusion process typically employs residual connections to retain the information from the original features.

The working principle of the ESA module is to enhance the model's ability to perceive image details and structures by selectively focusing on important edge features in the image. By introducing the edge-selective attention mechanism, the ESA module can more effectively capture critical information within the image and maintain the image's clarity and structure during reconstruction.

D. MetaRLFB

By leveraging the Conformer Block for effective learning of both global and local image features, and the ESA module for edge feature learning, the core mechanism of the Meta-RLFB, known as meta-learning, can be constructed. Meta-learning is a machine learning paradigm aimed at enabling the model to quickly learn and adapt using a small number of training samples, thereby achieving generalization to new tasks[19].

The role of the meta-learning module in Meta-RLFB is to learn the initialization and updating rules for adaptive parameters, allowing the model to quickly learn and adapt when faced with different super-resolution tasks. Specifically, the details are as follows:

1) Adaptive Parameters: These are the core components of the meta-learning module, responsible for dynamically adjusting the model's parameters during training to adapt to different super-resolution tasks. These adaptive parameters are typically designed as learnable variables, with their initialization and updating rules being the focus of the meta-learning module.

2) Update Rule: The meta-learning module also needs to design an update rule for adjusting the adaptive parameters. This update rule is usually designed based on feedback from the training samples and the specific form of the meta-learning algorithm, aiming to dynamically adjust the adaptive parameters according to the task's requirements.

During the meta-learning phase, the model learns how to quickly adapt to new super-resolution tasks using a small number of training samples through extensive training samples and the meta-learning algorithm. It learns how to effectively utilize the training samples for adjusting the adaptive parameters. In practical applications, when the model encounters a new super-resolution task, it quickly adjusts the adaptive parameters using a small number of training samples and the meta-learning module. It then employs recursive feature fusion and back-projection networks for

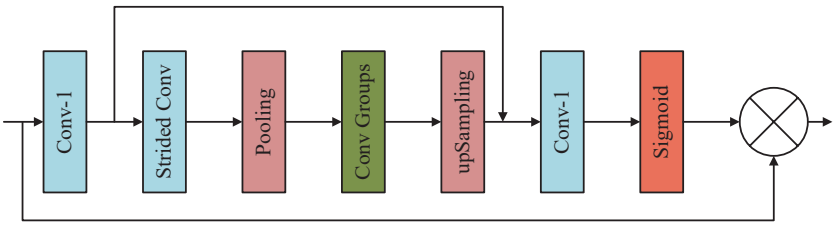

Fig. 5. The Structure of Edge-Selective Attention

multi-scale feature fusion and enhancement. Finally, the features are transformed into high-resolution images through a reconstructor, completing the image super-resolution reconstruction.

IV. Experiment

This section is organized into subheadings to provide a succinct and precise depiction of experimental results, their interpretation, and the empirical conclusions drawn.

A. Dataset

In the experimental section, the training dataset comprises the training subset of the DIV2K dataset. The model's performance is tested on four benchmark datasets: Set5, Set14, BSD100, Urban100 and Manga109. The following is a brief introduction to these datasets:

1) DIV2K[20]: DIV2K (Diverse 2K Resolution Dataset) is a large-scale dataset widely used in image super-resolution research. It consists of various types of images covering diverse scenes and contents, exhibiting rich diversity. Firstly, the DIV2K dataset includes a large number of high-resolution image samples, with resolutions reaching 2K (2048x1080) or higher. These images typically contain abundant details and textures, making them suitable for training and evaluating super-resolution algorithms. Secondly, the DIV2K dataset covers various types of images, including natural landscapes, urban scenes, portraits, and more. These images demonstrate rich diversity, enabling effective assessment of algorithm performance across different scenarios. Lastly, in addition to high-resolution images, the DIV2K dataset also provides corresponding low-resolution images, typically obtained through downsampling. These low-resolution images can be paired with high-resolution images for training and evaluating super-resolution algorithms.

2) Set5: Set5 is a small-scale dataset composed of 5 classic images, used for testing and evaluating the performance of image super-resolution algorithms. These images typically depict common scenes such as natural landscapes, buildings, etc., exhibiting different textures and structures. Due to its small scale, Set5 can be quickly employed for initial validation and evaluation of algorithms.

3) Set14: Set14 is a larger-scale dataset consisting of 14 classic images, used for more comprehensive testing and evaluation of image super-resolution algorithms. Compared to Set5, Set14 contains more image samples covering a wider range of scenes and contents. Evaluating with Set14 provides better understanding of algorithm performance across different scenarios, offering more reliable performance evaluation.

4) Urban100: Urban100 is a large dataset comprising 100 urban landscape images, used for evaluating the performance of image super-resolution algorithms in complex scenes. These images typically feature complex structures and textures, such as buildings, streets, vehicles, etc., posing higher challenges. Evaluation using the Urban100 dataset offers a more comprehensive understanding of algorithm performance in real-world application scenarios, providing richer performance assessment.

5) BSDS100: BSDS100 is a large dataset consisting of 100 natural images, commonly used for evaluating the performance of image super-resolution algorithms in natural scenes. These images encompass rich textures, structures, and content, covering various natural scenes such as landscapes, animals, plants, etc. Evaluating with the BSDS100 dataset offers better insight into algorithm applicability and generalization in natural scenes, providing a more comprehensive performance assessment.

5) Manga109[21]: Manga109 is a large-scale dataset comprising 106,000 images from 106 different manga titles, designed for evaluating and testing performance in the realm of manga scenes. Each image is meticulously annotated, providing essential metadata such as page numbers and chapter identifiers. The diversity of content within Manga106 makes it a valuable resource for advancing techniques in computer vision and natural language processing, particularly in the field of graphic storytelling.

These datasets are commonly employed for training, validation, and testing of image super-resolution algorithms, aiming to evaluate algorithm performance across different scenarios and datasets, thereby promoting algorithm development and improvement.

B. Evaluation Indicators

In this paper, PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) are employed as two commonly used image quality evaluation metrics to assess the performance of image reconstruction or compression algorithms. These metrics are also frequently used in the field of image super-resolution to evaluate the quality of reconstructed images. Below is a brief introduction to these two evaluation metrics:

1) PSNR: PSNR is a measure of image quality that is typically used to evaluate the degree of similarity between the reconstructed image and the original image. It quantifies image distortion based on the mean squared error (MSE) between the pixel values of the two images. The formula for its calculation is shown in Equation (5).

$$\mathrm{PSNR} = 10 \cdot \log_{10} \left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}} \right) \tag{5}$$

where MAX represents the maximum possible pixel value (usually 255), and MSE is the mean squared error, indicating the average squared difference in pixel values between the original and reconstructed images. Higher PSNR values indicate greater similarity between the super-resolved and original images.

2) SSIM: SSIM is a measure of structural similarity between images, considering not only image luminance, contrast, and structural information but also the perceptual characteristics of the human visual system. Its calculation process is described by Equation (6).

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{6}$$

where x and y are the original and reconstructed images, respectively, μ_x and μ_y are the mean values of the images, σ_x^2 and σ_y^2 are the variances of the images, σ_{xy} is the covariance of the images, and C_1 and C_2 are constants used to stabilize the calculation. The SSIM value ranges from -1 to 1, with values closer to 1 indicating higher similarity between the images.

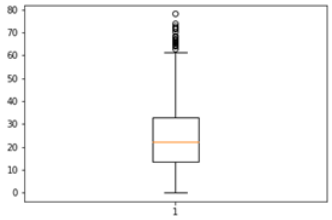

Fig. 6. Tile Gradient Distribution

C. Data Cleaning and Experimental Setup

To enhance the model's generalization capability, this study cropped the training set into patches of size 1024. During the model training process, data augmentation techniques such as image flipping and rotation were employed to fully utilize all available data. To expedite the training process, a random cropping strategy was also implemented. This approach, however, resulted in numerous blank patches and patches with minimal information, which do not contribute to the network's ability to learn deblurring. Consequently, data cleaning was performed to address this issue. Given that the cropped data exceeded 70,000 patches, screening became quite challenging, especially when training on Aistudio, which requires continuous data reloading.

To address this, the average gradient of each patch was computed to measure the complexity of the textures contained within each patch, and the statistics were visualized, as shown in Fig. 6. It is evident from the statistical graph that the majority of patches have gradient values around 20, with few patches having gradient values below 10. Upon further comparison of the cropped data, it was found that patches with an average gradient less than 10 were mostly blank or had minimal textures. Consequently, patches with an average gradient greater than 10 were written to a text file, and data were read from this file to construct the training set.

This study employed a progressive training strategy to accelerate the training process. The training was conducted in stages with varying batch sizes and patch sizes as follows: batch size of 8 and patch size of 192 for 184,000 iterations, batch size of 5 and patch size of 256 for 128,000 iterations, batch size of 4 and patch size of 320 for 96,000 iterations, batch size of 2 and patch size of 384 for 72,000 iterations, batch size of 1 and patch size of 512 for 72,000 iterations, and batch size of 1 and patch size of 1024 for 48,000 iterations. A cosine annealing learning rate strategy was used to optimize the network, with initial learning rate adjustments at 184,000 and 416,000 iterations. The learning rate was set to 2e-4, and the optimizer used was AdamW.

D. Baseline

To better evaluate the performance of HyperFLFN, this study selected several mainstream image super-resolution models for comparison, including RCAN, CAN, IGNN, HAN, RLFN, and SwinIR. Below is a brief introduction to these models:

1) RCAN (Residual Channel Attention Network)[22]: RCAN is an image super-resolution model based on residual learning and channel attention mechanisms. By incorporating residual learning modules and channel attention modules, it achieves efficient extraction and reconstruction of image features.

2) CAN (Context Aggregation Network)[23]: CAN is an image super-resolution model based on context aggregation mechanisms. It leverages multi-scale feature fusion and context-aware mechanisms to fully utilize both global and local information of images. The CAN model exhibits strong adaptability and generalization capabilities, achieving notable results in the field of image super-resolution.

3) IGNN (Iterative Gradient-based Nearest Neighbour) [24]: IGNN is an image super-resolution model based on iterative gradient descent and nearest neighbor interpolation. It iteratively optimizes the difference between the reconstructed image and the original image, gradually enhancing image details and textures using nearest neighbor interpolation. The IGNN model is simple yet effective, achieving good reconstruction results in certain scenarios.

4) HAN (Hybrid Attention Network)[25]: HAN is an image super-resolution model that integrates spatial attention and channel attention mechanisms. By simultaneously considering spatial and channel information of images, it achieves dual attention to image details and structures. The HAN model demonstrates good generalization capabilities and performance, finding applications in image super-resolution tasks.

5) RLFN (Residual Local Feature Network)[6]: RLFN is an image super-resolution model based on residual learning and local feature extraction. It employs residual learning modules and local feature fusion mechanisms to fully utilize both local and global information of images. The RLFN model excels in reconstructing image details and textures, achieving good performance on several datasets.

6) SwinIR (Swin Transformer-based Image Restoration) [26]: SwinIR is an image restoration model based on

TABLE I
COMPARISON OF ABLATION EXPERIMENTS

Different module combinations			Set5		Set14		BSD100		Urban100		Manga109	
ESA Channels	Conformer Block	metaRLFB	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
64	2	4	38.29	0.962	34.15	0.921	32.39	0.901	33.38	0.938	35.49	0.944
64	3	4	38.30	0.962	34.13	0.922	32.39	0.901	33.38	0.939	35.49	0.945
64	4	4	38.32	0.961	34.17	0.920	32.40	0.901	33.38	0.939	35.44	0.945
64	4	6	38.38	0.964	34.22	0.923	32.51	0.902	33.42	0.943	35.46	0.947
64	6	6	38.42	0.963	34.43	0.930	32.53	0.903	33.47	0.952	35.50	0.949
128	6	8	38.30	0.951	34.72	0.910	32.56	0.901	33.53	0.959	35.51	0.950

TABLE II
COMPARISON AT TWICE THE SUPER SCORE

Method	Datasets									
	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RCAN	38.27	0.9614	34.12	0.920	32.41	0.902	33.34	0.938	39.44	0.978
SAN	38.31	0.962	34.07	0.921	32.42	0.897	33.10	0.937	39.32	0.979
IGNN	38.24	0.9613	34.07	0.922	32.41	0.900	33.23	0.938	39.35	0.978
HAN	38.27	0.9614	34.16	0.922	32.41	0.901	33.35	0.939	39.35	0.977
RLFN	38.29	0.9618	34.15	0.922	32.42	0.901	33.38	0.939	39.34	0.979
SwinIR	38.35	0.962	34.14	**0.924**	32.44	**0.903**	33.40	0.939	39.58	0.979
HyperFLFN	**38.38**	**0.964**	**34.22**	0.923	**32.50**	0.902	**33.42**	**0.943**	**39.60**	**0.980**

TABLE III
COMPARISON AT FOUR TIMES THE SUPER SCORE

Method	Datasets									
	Set5		Set14		BSD100		Urban100		Manga109	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RCAN	32.63	0.900	28.87	0.787	27.77	0.744	26.82	0.808	31.22	0.917
SAN	32.64	0.901	28.92	0.788	27.78	0.744	26.79	0.806	31.18	0.917
IGNN	32.57	0.899	28.85	0.789	27.77	0.743	26.84	0.809	31.28	0.918
HAN	32.64	0.902	28.9	0.789	27.80	0.744	26.85	0.809	31.20	0.916
RLFN	32.63	0.897	**29.00**	0.791	27.81	0.745	2688	0.810	31.30	0.920
SwinIR	32.72	**0.903**	28.94	0.791	27.83	0.746	27.07	0.816	31.67	0.922
HyperFLFN	**38.73**	0.901	28.93	**0.810**	**28.01**	**0.760**	**27.28**	**0.943**	**31.70**	**0.925**

the Swin Transformer, encompassing tasks such as super-resolution, denoising, and deblurring. It utilizes the self-attention mechanism and local feature extraction capabilities of the Swin Transformer to achieve efficient image reconstruction and restoration. The SwinIR model demonstrates excellent performance in image reconstruction tasks and leads in performance on multiple datasets

E. Ablation Experiment

To validate the effectiveness of the three proposed mechanisms, ablation experiments were conducted under the conditions of using the DIV2K dataset and a super-resolution scale factor of $2\times$. The experiments were designed with varying numbers of ESA channels, Conformer Blocks, and metaRLFB modules. The specific results are presented in Table I.

As shown in Table I, increasing the number of ESA channels and Conformer Blocks can improve the quality of image super-resolution to a certain extent. This improvement is primarily due to the enhanced learning of global, local, and edge feature information in the images. However, when the number of ESA channels increases to 128, the excessive focus on important edge features introduces too much noise into the super-resolved images. Consequently, the final configuration of the HyperFLFN model includes 6 Conformer Blocks, 6 metaRLFB modules, and 64 ESA channels.

V. RESULT AND ANALYSIS

In this study, the training dataset was uniformly set to DIV2K. The performance of HyperFLFN and the six baseline models mentioned in Section 4.4 was evaluated on the Set5, Set14, BSD100, Urban100 and Manga109 datasets, with super-resolution scales of $2\times$ and $4\times$. All data represent the average of five experimental results, with the best results highlighted in bold and the second-best results underlined. Table II presents the comparison results for $2\times$ super-resolution, while Table III shows the comparison results for $4\times$ super-resolution.

From the data in Table II and Table III, it can be observed that HyperFLFN achieved either the best or second-best results across almost all evaluation metrics in the various datasets. This demonstrates that HyperFLFN has a significant advantage over other models in image super-resolution tasks. Specifically, in the $2\times$ super-resolution comparison, HyperFLFN improved the PSNR metric on the four datasets by 0.08%, 0.18%, 0.18%, 0.06% and 0.05%, respectively, with an average improvement of 0.11% compared to the second-best results. For the SSIM metric, HyperFLFN showed improvements of 0.21%, 0.03%, 0.10%, 0.39% and 0.03% on the four datasets, with an average improvement of 0.15

In the 4x super-resolution comparison, HyperFLFN improved the PSNR metric on the four datasets by 18.37%, 0.03%, 0.65%, 0.78% and 0.03%, respectively, with an average improvement of 4.0%. For the SSIM metric, the improvements were 0.03%, 2.35%, 1.89%, 15.54% and 0.01% on the four datasets, with an average improvement of 4.0%. The improvements in $4\times$ super-resolution were significantly greater than those in $2\times$ super-resolution, indicating that the proposed model performs well in image super-resolution tasks.

VI. Conclusions

Image super-resolution has long been a critical task in the field of computer vision, aiming to reconstruct high-quality, high-resolution images from low-resolution counterparts. Traditional interpolation methods perform well under certain circumstances but suffer significant performance degradation when handling images rich in detail. Although deep learning-based models have made some progress, they generally exhibit suboptimal super-resolution performance, long training times, and insufficient generalization capabilities.

To address these issues, this paper proposes a novel image super-resolution model named HyperFLFN. This model introduces an innovative structure, the Conformer module, which integrates self-attention mechanisms into traditional residual connections. This integration enables the simultaneous learning of global and local features, thereby more effectively capturing image details and contextual information. Compared to traditional methods, the HyperFLFN model achieves significant improvements in both image super-resolution quality and detail preservation. Additionally, this paper incorporates a meta-learning mechanism to enhance the model's generalization ability. Extensive experimental validation shows that HyperFLFN achieves an average improvement of 4.94% in PSNR and SSIM compared to other models. The results indicate that the HyperFLFN model demonstrates superior performance across various scenarios, particularly in handling different types of images, highlighting its potential and feasibility for practical applications.

In summary, the proposed HyperFLFN model not only makes significant advancements in the task of image super-resolution but also shows great potential in enhancing image quality and generalization capability. This research provides new insights and methodologies for further developments in the field of image processing.

References

[1] N. Jiang and L. Wang, "Quantum Image Scaling Using Nearest Neighbor Interpolation," *Quantum Information Processing*, vol. 14, pp. 1559–1571, 2015.

[2] P. Smith, "Bilinear Interpolation of Digital Images," *Ultramicroscopy*, vol. 6, no. 2, pp. 201–204, 1981.

[3] R. Keys, "Cubic Convolution Interpolation for Digital Image Processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[4] C. M. Ward, J. Harguess, B. Crabb, and S. Parameswaran, "Image quality assessment for determining efficacy and limitations of Super-Resolution Convolutional Neural Network (SRCNN)," in *Applications of Digital Image Processing XL*, vol. 10396. SPIE, 2017, pp. 19–30.

[5] M. A. Talab, S. Awang, and S. A.-d. M. Najim, "Super-Low Resolution Face Recognition Using Integrated Efficient Sub-Pixel Convolutional Neural Network (ESPCN) and Convolutional Neural Network (CNN)," in *2019 IEEE International Conference on Automatic Control and intelligent Systems (I2CACIS)*. IEEE, 2019, pp. 331–335.

[6] Y. Xiong, S. Guo, J. Chen, X. Deng, L. Sun, X. Zheng, and W. Xu, "Improved SRGAN for Remote Sensing Image Super-Resolution across Locations and Sensors," *Remote Sensing*, vol. 12, no. 8, p. 1263, 2020.

[7] F. Kong, M. Li, S. Liu, D. Liu, J. He, Y. Bai, F. Chen, and L. Fu, "Residual Local Feature Network for Efficient Super-Resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 766–776.

[8] Z. Wang, J. Chen, and S. C. Hoi, "Deep Learning for Image Super-Resolution: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[9] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep Learning for Single Image Super-Resolution: A brief Review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.

[10] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[12] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 136–144.

[13] Y. Fan, H. Shi, J. Yu, D. Liu, W. Han, H. Yu, Z. Wang, X. Wang, and T. S. Huang, "Balanced Two-Stage Residual Networks for Image Super-Resolution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 161–168.

[14] J. Shi, Q. Liu, C. Wang, Q. Zhang, S. Ying, and H. Xu, "Super-Resolution Reconstruction of MR Image with A Novel Residual Learning Network Algorithm," *Physics in Medicine & Biology*, vol. 63, no. 8, p. 085011, 2018.

[15] C.-K. Huang and H.-H. Nien, "Multi Chaotic Systems Based Pixel Shuffle for Image Encryption," *Optics Communications*, vol. 282, no. 11, pp. 2123–2127, 2009.

[16] H. Zhao, J. Jia, and V. Koltun, "Exploring Self-Attention for Image Recognition," in *Proceedings of the IEEE/CVF Conference on Computer vision and Pattern Recognition*, 2020, pp. 10 076–10 085.

[17] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi *et al.*, "Recent Developments on Espnet Toolkit Boosted by Conformer," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5874–5878.

[18] D. D'Ambrosio, A. Iellem, R. Bonecchi, D. Mazzeo, S. Sozzani, A. Mantovani, and F. Sinigaglia, "Cutting edge: Selective Up-Regulation of Chemokine Receptors CCR4 and CCR8 upon Activation of Polarized Human Type 2 Th Cells," *The Journal of Immunology*, vol. 161, no. 10, pp. 5111–5115, 1998.

[19] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-Learning in Neural Networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5149–5169, 2021.

[20] E. Agustsson and R. Timofte, "Ntire 2017 Challenge on Single Image Super-Resolution: Dataset and Study," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 126–135.

[21] A. Fujimoto, T. Ogawa, K. Yamamoto, Y. Matsui, T. Yamasaki, and K. Aizawa, "Manga109 dataset and creation of metadata," in *Proceedings of the 1st International Workshop on coMics ANalysis, Processing and Understanding*, 2016, pp. 1–5.

[22] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual Attention Network for Image Classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[23] J. Lu, R. Mottaghi, A. Kembhavi *et al.*, "Container: Context Aggregation Networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19 160–19 171, 2021.

[24] R. Timofte and L. Van Gool, "Iterative Nearest Neighbors," *Pattern Recognition*, vol. 48, no. 1, pp. 60–72, 2015.

[25] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid Multiple Attention Network for Semantic Segmentation in Aerial Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2021.

[26] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image Restoration using Swin Transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.