# Edge-enhanced Information Distillation Transformer Network for Lightweight Super-Resolution

**Yike Cui*, Zhonghua Liu**, Weihua Ou***, Yong Liu*, Kaibing Zhang******

* *Information Engineering College, Henan University of Science and Technology, Luoyang, China (e-mail: 740081575@qq.com, luoyangliuyong@163.com).*
** *School of Information Engineering, Zhejiang Ocean University, Zhoushan, China (e-mail: lzhua_217@163.com).*
*** *School of Big Data and Computer Science, Guizhou Normal University, Guizhou, China (e-mail: ouweihua@gznu.edu.cn).*
****College of Electronics and Information, Xi'an Polytechnic University, Xian, China (e-mail: zhangkaibing@xpu.edu.cn).*

**Abstract**: In the recent works, deep learning has been applied to the single-image super-resolution (SISR) with significant results. To enhance performance, most approaches advocate for the construction of deeper and wider networks. Nonetheless, the augmentation of network depth and width could result in considerable memory storage demands and computational expenses. Besides, there is still a lot of room for improvement in edge information processing. In order to tackle these problems, we present an edge-enhanced information distillation transformer network (EDTN) for lightweight super-resolution, in which the edge enhancement block and the transformer module are combined into a unified framework. Specifically, this work adopts an edge-enhanced reparameterization convolution block based on the existing reparameterization methods which can pay more attention to the edge information in the image. Meanwhile, this work proposes a novel gradient-variance loss that can recover edge details and retain sharp edge visuals. In order to use the information of different situations more flexibly from the feature, this work employ an transformer to preserve high-frequency information similar features across long distances. The experimental results demonstrate the effectiveness of the proposed method.

*Keywords*: Super-resolution; Lightweight Network; Deep Learning; Transformer

## 1. INTRODUCTION

Single image super-resolution (SISR) (Freeman et al., 2000) is a comprehensive task concerning low-level computer vision, whose goal is to restore a high-resolution (HR) image from its low-resolution (LR). This is the ill-conditioned issue as one high-resolution (HR) image typically corresponds to lots of low-resolution images. To solve this challenging problem, many classical methods have been proposed. Nevertheless, the majority of them concentrate on improving restoration quality by increasing the model scale, which can make deployment on mobile devices more difficult. Therefore, it has become a hot topic of concern to design lightweight networks for SR.

The super-resolution convolutional neural network (SRCNN) (Dong et al., 2014) is a pioneering work, in which the convolutional neural network is firstly introduced into SISR. Besides, the deeply-recursive convolutional network (DRCN) (Kim et al., 2016) and the deep recursive residual network (DRRN) (Tai et al., 2017a) adopt a recursive network, which can effectively decrease parameters by parameter sharing. The performance degradation caused by recursive modules needs to be compensated by increasing the network's depth and width. In this case, it's not effective for model performance due to their limited representation capability, which need cost more computational resources. Therefore, it has become a focus of attention to create influential modules and dedicated networks for improving SR efficiency. (Liu et al., 2020) introduce information distillation technology and shallow residual convolution blocks in residual feature distillation network (RFDN) to improve performance under limited conditions. Besides, (Zhang et al., 2021) implement a lightweight convolutional network with re-parameterized convolutional blocks to enhance the extraction of edge information. These networks enable real-time applications on standard mobile devices. However, the proposed approaches still have room for improvement in terms of performance. Furthermore, the neural architecture search (NAS) (Chu et al., 2021) and model pruning (Li et al., 2020) are developed to achieve networks that are more lightweight and efficient. By introducing a method that is differentiable to identify a more flexible architecture based on RFDN, Huang et al. present an alternative approach to neural architecture search (NAS) which designs both network-level and cell-level structures with minimal cost. Nevertheless, most of these approaches concentrate on capturing local contextual details and overlook broader similar textures, potentially causing issues like artifacts in the final image. Due to the restricted receptive fields of basic convolution operations, incorporating both global features and local details poses a challenge, which need to find a balance between efficiency and complexity.

Recently, there are significant achievements in image super-resolution approaches based on the transformer architecture. On the one hand, many works on SR transformer networks have prompted the development of lightweight transformer

SISR effectively. On the other one hand, the previous distillation-based methods have achieved impressive performance in SISR, so there is still room for improvement in the channel and spatial extracting abilities. This proposed feature extraction block combines two complementary components, which are respectively the edge enhancement block (Wang, 2022) and the transformer module (Zamir et al., 2021). A transformer module with linear complexity to implement channel-wise self-attention is proposed to capture dependencies within the input features along the channel dimension effectively.

The key contributions of this work are described as follows:

1) This paper proposes a lightweight Edge-enhanced Feature Distillation Transformer network (EDTN) to resolve the image super-resolution problem efficiently, which has significant visual effects under limited resource utilization.

2) An edge-enhanced distillation transformer block (EDTB) is presented, which introduces transformer block into edge enhancement block. The proposed block preserves the high-frequency edge information effectively to enhance image edge information reconstruction.

3) The proposed method introduces a loss function based on gradient variance to preserve edge information, enhancing the capability to restore edge details and maintain sharp visual edges.

## 2. RELATED WORK

### *2.1 Image super-resolution models based on CNN*

Recently, convolutional neural networks (CNNs) have made significant strides in advancing low-level computer vision tasks (Shafiq et al., 2022). In the field of super-resolution, (Dong et al., 2016b) introduce SRCNN, which is the earliest CNN-based approach, and has achieved good performance compared with traditional methods. Building upon this foundation, (Kim et al., 2016) introduce more intricate DRCN and VDSR networks using residual learning principles. (Tai et al., 2017a) further enhance the architecture by introducing DRRN with recursive blocks and the innovative memory-based approach, MemNet (Tai et al., 2017b). However, these techniques rely on bicubic interpolation for preprocessing LR images, resulting in potential loss of details and increased computational complexity. To address this issue, (Dong et al., 2016a) propose the accelerating the Super-Resolution convolutional neural network (FSRCNN), integrating deconvolution layers for image upsampling at the network's conclusion, thereby reducing computational burden. Subsequently, a lot of CNN-based SR networks (Lim et al., 2017) have emerged to refine reconstruction outcomes. For instance, (Lim et al., 2017) propose the enhanced deep residual network (EDSR) by stacking residual blocks and eliminating batch normalization layers. Additionally, the additional networks such as the non-local neural network (NLRN) (Wang et al., 2018), the residual channel attention network (RCAN) (Zhang et al., 2018), and the second-order attention network (SAN) (Dai et al., 2019) have been developed to enhance performance by capturing feature correlations in spatial and channel dimensions. Nonetheless, these networks typically exhibit a high parameter count, which are often difficult to implement in practical applications due to equipment limitations and other factors.

### *2.2 Lightweight SR models*

Currently, in order to address the challenge of implementing super-resolution (SR) on edge devices, many lightweight networks have been proposed. These lightweight networks can be roughly categorized into 3 types. The methods based on knowledge distillation (Gao et al., 2018), architectural design (Gao et al., 2023), and neural architecture search (NAS) (Chu et al., 2021). (Kim et al., 2016b) propose Deeply-Recursive Convolutional Network (DRCN) and (Tai et al., 2017b) propose Deep Recursive Residual Network (DRRN) introduce recursive layer parameter sharing to reduce parameters. However, challenges such as a large number of network layers still prevent further reduction in memory requirements for devices. Subsequently, (Ahn et al., 2020) propose cascading residual network (CARN). The experimental results indicate that the cascaded mechanism of residual learning introduced by CARN is beneficial in reducing parameters. Lightweight feature fusion network (LFFN) (Chu et al., 2020) leverages global features by using Softmax function and channel attention mechanisms. NAS (Zoph et al., 2020) is a novel approach for automatically refining network structure parameters and has found application in super-resolution (SR) tasks (Lin et al., 2022). NAS comprises three key components: search space, search strategy, and performance evaluation strategy. The information multi-distillation network (IMDN) (Zheng et al., 2019) introduces a channel attention mechanism grounded in contrast and an adaptive cropping strategy, thereby boosting the model's effectiveness.

Furthermore, the re-parameterization is a method for designing efficient neural networks. (Ding et al., 2019) introduce an asymmetric convolution block (ACB) to enhance the standard convolution by combining 3 different convolutions. RepVGG (Ding et al., 2021a) is the first to incorporate identity mapping and 1×1 convolution into the re-parameterizable structure community, while the diverse branch block (DBB) (Ding et al., 2021b) expands the community further with sequential convolutions such as expanding-and-squeezing convolution. Building on this, (Zhang et al., 2021) develop an edge-enhanced convolution block (ECB) to enhance the performance of real-time super-resolution networks on mobile devices.

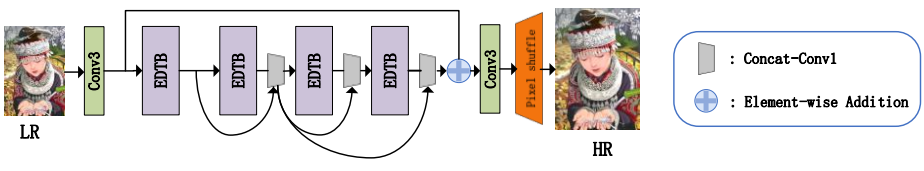


Fig. 1. Overall network architecture of the proposed EDTN.

### *2.3 Vision Transformer*

Because of the advancements achieved by Transformer in the field of NLP, it has generated considerable interest within the computer vision field, which has been effectively utilized in segmentation (Wang et al., 2021), object detection (Carion et al., 2020; Dai et al., 2020), and image recognition (Dosovitskiy et al., 2020). Presently, most approach in Vision Transformers segment the image into patch sequences, which are subsequently transformed into vectors to enable the exploration of their relationships using self-attention mechanisms. Vision Transformers exhibit strong capabilities in learning extended relationships between image pixels. However, the self-attention's computational complexity in Ttransformers grow dramatically with the quantity of image segments, thereby limiting its application to high-resolution images. The recent methods generally employ alternative strategies to reduce complexity. Consequently, the development of efficient Vision Transformers has emerged as a prominent research focus in recent years.

## 3. PROPOSED METHOD

### *3.1 Overview of Network Framework*

The edge-enhanced information distillation transformer network (EDTN) has outstanding super-resolution (SR) performance despite resource constraints. Notably, the EDTN excels in reconstructing high-quality SR images characterized by precise edges and well-defined structures. The architecture of the EDTN, illustrated in Fig. 1, comprises a shallow feature extraction module, multiple edge-enhanced information distillation transformer blocks (EDTBs), and an upscaling module. Initially, a 3×3 simple convolution is utilized to produce the initial feature maps, followed by the extraction of information from the input low-resolution image ($I^{LR}$). This intricate process can be succinctly depicted as:

$$F_0 = E(I^{LR}) \tag{1}$$

where the E expresses the feature extraction function by a 3×3 convolution, and $F_0$ is the extracted feature maps. These initial features are then passed to stacked EDTBs for further information extraction. In detail, as shown in Fig. 3a, this network embraces network-level NAS strategy suggested in DLSR (Huang et al., 2021) to determine the feature connection paths and denotes the proposed EFDB as $H_n$. The information distillation process can be expressed as:

$$F_1 = H_1(F_0) \tag{2}$$

$$F_2 = H_2(F_1) \tag{3}$$

$$F_3 = H_3(C_1(F_1, F_2)) \tag{4}$$

$$F_4 = C_3(F_2, H_4(C_2(F_2, F_3))) + F_0 \tag{5}$$

where $C_n$ is n-th fusion operator comprising of concatenation and 1×1 convolution. $F_n$ is the n-th output feature. And the upscaling module generate SR images:

$$I^{SR} = G(F_4) \tag{6}$$

where G comprises of a 3×3 vanilla convolution and sub-pixel convolution to map feature graphs into images.

### *3.2 Edge-enhanced distillation transformer block*

#### 3.2.1 Edge-enhanced re-parameterization branch block

Limited receptive field of the 3x3 convolution leads to visualization results with loss of edge details. The Edge-Enhanced re-parameterization branch block (ERBB) is introduced to capture and preserve the edge structure information within images effectively. As illustrated in Fig. 2, the ERBB comprises of seven branches, which is mainly divided into vanilla convolution and sequential convolutions. Except for simple convolutions such as 1×1 convolution and 3×3 convolution, ERBB adopts discrete differentiation operators such as Isotropic Sobel and Laplacian for edge detection. The Sobel operator can be divided into two types:

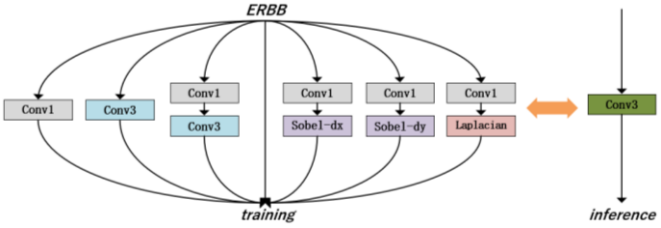


Fig. 2. Edge-enhanced re-parameterization branch block (ERBB).

horizontal and vertical, which perform convolution operations on the image separately to obtain gradient images in the horizontal and vertical directions. The final edge strength image is then calculated based on these two gradient images. Besides, the Laplacian operator is utilized for detecting edges and regional features within images. By computing the second derivative of pixel values in the image, it enhances the edge information present in the image. By integrating vanilla convolution with edge detection operators, ERBB is committed to achieving an efficient super-resolution effect.

During the training phase, the model is trained by ERBB, which can continuously update and obtain more reasonable intermediate layer features through iterative updates. The key to this process is to utilize the structural characteristics of ERBB to gradually update and optimize the intermediate layer features, thus ensuring the continuous improvement of the model in feature extraction capability. Specifically, ERBB is able to overcome the limitations of traditional convolutional layers in feature learning to a certain extent by introducing diversified re-parameterizable branches so that each branch can capture different feature information. This approach not only improves the rationality and expressiveness of the intermediate layer features, but also enhances the model's adaptability to complex image contents. In the testing phase, ERBB is converted to vanilla convolution by computing the total reparameterization parameters. The realization of this transformation process relies on the comprehensive computation of all the re-parameterization parameters, enabling the model to efficiently utilize the learned feature information in its inference. In this way, ERBB not only improves the quality of image SR but also maintains a faster speed at runtime compared to traditional simple convolution. This speed

advantage is achieved mainly due to the efficient computational mechanism and optimization strategy introduced in the ERBB framework, which makes the model more practical in real-time application scenarios while ensuring the image quality.

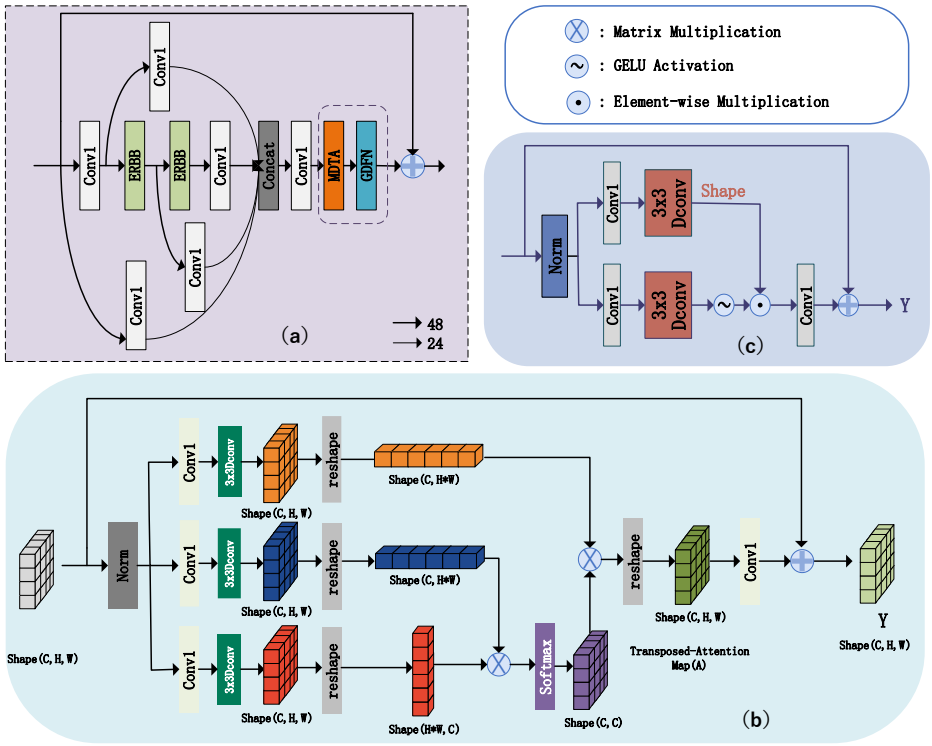


Fig 3. (a) Edge-enhanced distillation efficient transformer block (EDTB), (b) Multi-Dconv head transposed attention (MDTA), and (c) Gated-Dconv feed-forward network (GDFN).

*3.2.2 Transformer block*

To improve the network's performance, an transformer module is introduced to pay close attention to global contextual information. The module captures long-range dependencies and global contextual information within the image, enabling the model to better understand the relationships between different parts of the image. By leveraging self-attention mechanisms, transformer modules can effectively handle complex spatial relationships and patterns in the image data, making them a powerful tool for image super-resolution tasks. This transformer comprises Multi-Dconv head transposed attention (MDTA) module and Gated-Dconv feed-forward network (GDFN) module.

The Transformer model expands the receptive field, but its computational complexity grows quadratically with increasing spatial resolution, making it unsuitable for most image reconstruction tasks. To address this issue, EDTN employs MDTA to implicitly encode global contextual information by computing self-attention across channels. Specifically, as shown in Figure 3 (b), EDTN achieves this by pixel-level aggregation of cross-channel contexts using 1×1 convolution and channel-level aggregation of local contexts using deep convolution. This research has employed depth-wise convolution and pixel convolution in 3 branches to generate value (V), key (K), and query (Q) from the input features $X \in R^{H\times W\times C}$. These are described as follows:

$$\mathrm{V} = H^{3}_{dconv}(H^{3}_{pconv}(LN(\mathrm{X}))) \tag{7}$$

$$\mathrm{K} = H^{2}_{dconv}(H^{2}_{pconv}(LN(\mathrm{X}))) \tag{8}$$

$$\mathrm{Q} = H^{1}_{dconv}(H^{1}_{pconv}(LN(\mathrm{X}))) \tag{9}$$

where LN, $\mathrm{H_{dconv}}$, and $\mathrm{H_{pconv}}$ denote the layer normalization, depth-wise convolution and pixel convolution, respectively. MDTA uses reshape operation to obtain $\hat{V} \in R^{HW\times C}$, $\hat{K} \in R^{C\times HW}$ and $\hat{Q} \in R^{HW\times C}$. Their interaction through dot-product computation yields a transposed-attention map A with dimensions $\mathrm{R_{CxC}}$. The formulation is given by:

$$\mathrm{A} = Softmax(\mathrm{Q}\cdot\mathrm{K}/\alpha)\cdot\mathrm{V} \tag{10}$$

$$\mathrm{Y} = \mathrm{X} + H_{pconv}(\mathrm{A}) \tag{11}$$

where Softmax represents softmax function to generate probability graph. The parameter α is a trainable proportional factor that controls the magnitude of the dot product between Q and K. In contrast to the conventional transformer model that computes self-attention on the spatial domain, MDTA significantly decreases calculated amount.

To enhance the restoration of structural feature, this work employs a Gated-Dconv feed-forward network (GDFN). In Figure 3 (c), the gating mechanism within GDFN is structured as the element-wise multiplication of two parallel pathways within the linear transformation layer, with one pathway being modulated by the GELU non-linear activation function. Similar to MDTA, GDFN incorporates depth-wise convolutions to encode information from spatially adjacent pixel positions, which is beneficial for learning local image

structures to facilitate effective image restoration. Specifically, given the input features X, GDFN can be expressed as:

$$X_G^1 = \phi(H_{dconv}(H_{pconv}(LN(X)))) \quad (12)$$

$$X_G^2 = H_{dconv}(H_{pconv}(LN(X))) \quad (13)$$

$$Y_G = X_G^1 \odot X_G^2 \quad (14)$$

$$Y = H_{pconv}(Y_G) \quad (15)$$

where Φ and LN express GELU activation function and layer normalization. GDFN effectively controls the information flow across different hierarchical levels, permitting each level to focus on intricate details that enhance the overall representation.

*3.2.3 Edge-enhanced gradient-variance loss*

In anterior models, $L_1$ and $L_2$ losses have been used to achieve higher evaluation metrics. $L_1$ loss is characterized by gradient stability and fast convergence rate. $L_2$ loss is known for its ease of training and differentiability. Since $L_2$ loss calculates the average of the squares of the difference between the predicted and true values. $L_2$ loss is more sensitive to larger errors compared to the $L_1$ loss. However, the effectiveness of using these two functions individually is not ideal. Due to the presence of 7 parallel branches in ERBB, both gradient and differentiability issues become more complex during the training process. To achieve optimal performance in the edge-enhanced re-parameterization branches (ERBBs), it is essential to maintain clear edge details, which is crucial for enhancing the visual effects of the output. Therefore, this work proposes an edge-enhanced gradient-variance (EG) loss based on the gradient variance (GV) loss (Abrahamyan et al., 2022), which take advantage of the filters of the ERBB to monitor the optimization of the model. Firstly, the SR image $I_{SR}$ and the HR image $I_{HR}$ are respectively transferred to gray-scale images $G_{SR}$ and $G_{HR}$. This work leverages the Laplacian and Sobel filters to compute the gradient graphs, which are further decomposed into $\frac{HW}{n^2} \times n^2$ patches $G_x$, $G_y$, $G_z$. In detail, $G_x$, $G_y$, and $G_z$ represent the gradient variance loss of the Sobel filter in the horizontal direction, the gradient variance loss of the Sobel filter in the vertical direction and the gradient variance loss of the Laplacian filter, respectively. The i-th variance graphs are represented as:

$$v_i = \frac{\sum_{j=1}^{n^2}(G_{i,j} - \bar{G}_i)}{n^2 - 1} \quad (16)$$

where $\bar{G}_i$ is the mean value of the i-th patch. This study separately measures the variance norms $v_z$, $v_x$, $v_y$ between SR and HR images. This study calculates the gradient variance loss of various filtering operator by:

$$L_x = E_{I_{SR}} \left\| v_x^{HR} - v_x^{SR} \right\| \quad (17)$$

$$L_y = E_{I_{SR}} \left\| v_y^{HR} - v_y^{SR} \right\| \quad (18)$$

$$L_z = E_{I_{SR}} \left\| v_z^{HR} - v_z^{SR} \right\| \quad (19)$$

Besides, $L_E$ adds $L_1$ and $L_2$ to enhance the restoration efficacy and expedite convergence. To optimize the edge-enhanced re-parameterization branches within ERBBs and conserve clear edge details for enhanced visual effects, influencing weights $\lambda_x$, $\lambda_y$, and $\lambda_z$ are defined, which correspond to the scaling parameters of the Sobel filter in the horizontal direction, the Sobel filter in the vertical direction and the Laplacian filter, respectively. Specifically, the final convolution layer is substituted with ERBB in the pre-training phase to acquire appropriate proportional parameters $s_x$, $s_y$, and $s_z$ for various branches. Besides, λ is set to a constant according to the ablation experiment of 4.4.2. $\lambda^*$ is determined by calculating the weights of $s^*$, and the ERBB is converted back to a 3x3 convolution for training. The overall loss function is formulated as:

$$L_E = L_1 + \lambda L_2 + \lambda_x L_x + \lambda_y L_y + \lambda_z L_z \quad (20)$$

The convergence properties of the proposed model arise from the complementary nature of the loss components. $L_1$ loss imposes sparse gradients, reducing oscillation risks, while $L_2$ loss provides smooth optimization trajectories. The EG loss acts as a regularizer, constraining edge-related parameter updates through variance penalties. During training, ERBB's adaptive weights $\lambda_x$, $\lambda_y$, $\lambda_z$ modulate gradient magnitudes in edge-critical regions, preventing overfitting to high-frequency noise.

## 4. EXPERIMENTS

*4.1 Metrics and datasets*

In the experiment, the DIV2K (Dong et al., 2014) and Flick2K (Zhang et al., 2018) are used as training datasets, in which total of 3450 images are contained to generate high-resolution and bicubic down-sampled image pairs. During the evaluation phase, this research employs 4 widely recognized benchmark datasets: Urban100 (Huang et al., 2012), BSD100 (Martin et al., 2001), Set5 (Bevilacqua et al., 2021), and Set14 (Zeyde et al., 2010). Evaluation metrics such as peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) (Wang et al., 2004) are calculated in the Y channel of the YCbCr color space to assess the fidelity of the generated images.

*4.2 Experimental Setup*

For training purposes, DIV2K and Flick2K datasets are utilized to train the Edge-enhanced Information Distillation Transformer Network (EDTN). Data augmentation involves unpredictable variations of 90◦, 180◦, 270◦, and horizontal reflections applied to the training set. The HR patch size is configured as 256×256, with a mini-batch size of 32. The loss function is the edge-enhanced gradient-variance loss ($L_E$). To expedite training, a cosine annealing learning schedule is favored over the multi-step scheme. The original maximum learning rate is established at 1e−3, while the minimum learning rate is established to 1e−7. The cosine period spans 250k iterations. The propose algorithm (its code can by obtained by https://github.com/lzhua217/EDTN) is executed using the PyTorch framework (Paszke et al., 2019) on a system equipped with an NVIDIA RTX 3090 GPU

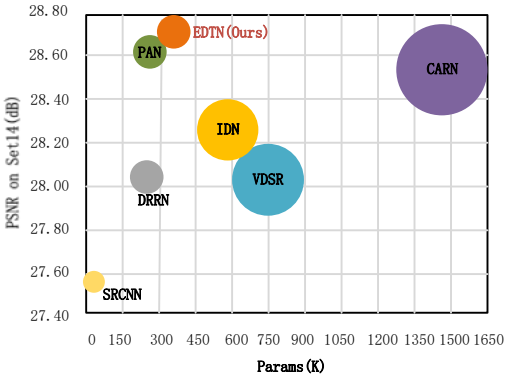


Fig. 4. Balancing act between model complexity and performance on the Set14 datasets with a magnification factor of ×4 is examined. The computational load is determined using a 320×180 input for multi-adds operation.

The training pipeline integrates $L_1$-$L_2$-EG hybrid loss and ERBB-driven adaptive regularization. Initially, $L_2$ loss coarsely aligns features, followed by joint L1+EG optimization for edge-texture fidelity. ERBBs refine spatial gradients via $\lambda^*$ weights, which are dynamically calibrated using edge-aware entropy metrics. This two-stage process ensures stable convergence while preserving high-frequency details.

**Table 1. The datasets Set5, B100, Set14, Manga109, and Urban100 were used to evaluate the Average PSNR/SSIM values for scale factors ×2, ×3, and ×4. Best and second-best results are highlighted in red and blue.**

| Method | Scale | Params | Set5 PSNR/SSIM | Set14 PSNR/SSIM | B100 PSNR/SSIM | Manga109 PSNR/SSIM | Urban100 PSNR/SSIM |
|---|---|---|---|---|---|---|---|
| Bicubic | | - | 33.66/0.9299 | 30.24/0.8688 | 29.56/0.8431 | 30.80/0.9339 | 26.88/0.8403 |
| SRCNN | | 8K | 36.66/0.9542 | 32.45/0.9067 | 31.36/0.8879 | 35.60/0.9663 | 29.50/0.8946 |
| VDSR | | 666K | 37.53/0.9587 | 33.03/0.9124 | 31.90/0.8960 | 37.22/0.9750 | 30.76/0.9140 |
| DRRN | | 298K | 37.74/0.9591 | 33.23/0.9136 | 32.05/0.8973 | 37.88/0.9749 | 31.23/0.9188 |
| DRCN | | 1774K | 37.63/0.9588 | 33.04/0.9118 | 31.85/0.8942 | 37.55/0.9732 | 30.75/0.9133 |
| IDN | ×2 | 553K | 37.83/0.9600 | 33.30/0.9148 | 32.08/0.8985 | 38.01/0.9749 | 31.27/0.9196 |
| CARN | | 1592K | 37.76/0.9590 | 33.52/0.9166 | 32.09/0.8978 | 38.36/0.9765 | 31.92/0.9256 |
| IMDN | | 694K | 38.00/0.9605 | 33.63/0.9177 | 32.19/0.8996 | 38.88/0.9774 | 32.17/0.9283 |
| PAN | | 261K | 38.00/0.9605 | 33.59/0.9181 | 32.18/0.8997 | 38.70/0.9773 | 32.01/0.9273 |
| RFDN | | 534K | 38.05/0.9606 | 33.68/0.9184 | 32.16/0.8994 | 38.88/0.9773 | 32.12/0.9278 |
| EDTN(Ours) | | 370K | 38.12/0.9611 | 33.72/0.9187 | 32.22/0.8999 | 38.98/0.9776 | 32.19/0.9281 |
| Bicubic | | - | 30.39/0.8682 | 27.55/0.7742 | 27.21/0.7385 | 26.95/0.8556 | 24.46/0.7349 |
| SRCNN | | 8K | 32.75/0.9090 | 29.30/0.8215 | 28.41/0.7863 | 30.48/0.9117 | 26.24/0.7989 |
| VDSR | | 666K | 33.66/0.9213 | 29.77/0.8314 | 28.82/0.7976 | 32.01/0.9340 | 27.14/0.8279 |
| DRRN | | 298K | 33.82/0.9226 | 29.76/0.8311 | 28.80/0.7963 | 32.24/0.9343 | 27.15/0.8276 |
| DRCN | | 1774K | 34.03/0.9244 | 29.96/0.8349 | 28.95/0.8004 | 32.71/0.9379 | 27.53/0.8378 |
| IDN | ×3 | 553K | 34.11/0.9253 | 29.99/0.8354 | 28.95/0.8013 | 32.71/0.9381 | 27.42/0.8359 |
| CARN | | 1592K | 34.29/0.9255 | 30.29/0.8407 | 29.06/0.8034 | 33.50/0.9440 | 28.06/0.8493 |
| IMDN | | 703K | 34.36/0.9270 | 30.32/0.8417 | 29.09/0.8046 | 33.61/0.9445 | 28.17/0.8519 |
| PAN | | 261K | 34.40/0.9271 | 30.36/0.8423 | 29.11/0.8050 | 33.61/0.9448 | 28.11/0.8511 |
| RFDN | | 541K | 34.41/0.9273 | 30.34/0.8420 | 29.09/0.8050 | 33.67/0.9449 | 28.21/0.8525 |
| EDTN(Ours) | | 370K | 34.45/0.9277 | 30.40/0.8426 | 29.16/0.8057 | 33.70/0.9452 | 28.28/0.8530 |
| Bicubic | | - | 28.42/0.8104 | 26.00/0.7027 | 25.96/0.6675 | 24.89/0.7866 | 23.14/0.6577 |
| SRCNN | | 8K | 30.48/0.8626 | 27.50/0.7513 | 26.90/0.7101 | 27.58/0.8555 | 24.52/0.7221 |
| VDSR | | 666K | 31.35/0.8838 | 28.01/0.7674 | 27.29/0.7251 | 28.83/0.8870 | 25.18/0.7524 |
| DRRN | | 298K | 31.53/0.8854 | 28.02/0.7670 | 27.23/0.7233 | 28.93/0.8854 | 25.14/0.7510 |
| DRCN | | 1774K | 31.68/0.8888 | 28.21/0.7720 | 27.38/0.7284 | 29.45/0.8946 | 25.44/0.7638 |
| IDN | ×4 | 553K | 31.82/0.8903 | 28.25/0.7730 | 27.41/0.7297 | 29.41/0.8942 | 25.41/0.7632 |
| CARN | | 1592K | 32.13/0.8937 | 28.60/0.7806 | 27.58/0.7349 | 30.47/0.9084 | 26.07/0.7837 |
| IMDN | | 715K | 32.21/0.8948 | 28.58/0.7811 | 27.56/0.7353 | 30.45/0.9075 | 26.04/0.7838 |
| PAN | | 272K | 32.13/0.8948 | 28.61/0.7822 | 27.59/0.7363 | 30.51/0.9095 | 26.11/0.7854 |
| RFDN | | 550K | 32.24/0.8952 | 28.61/0.7819 | 27.57/0.7360 | 30.58/0.9089 | 26.11/0.7858 |
| EDTN(Ours) | | 374K | 32.33/0.8962 | 28.66/0.7831 | 27.62/0.7377 | 30.77/0.9114 | 26.19/0.7880 |

*4.3 Comparisons with State-of-the-art Methods*

4.3.1 Comparison of model complexity with state-of-the-art experiments

To further demonstrate the superiority of EDTN in terms of model complexity, experiments compare EDTN with 6 advanced super-resolution models. The models were evaluated from the perspectives of parameters and performance. As shown in Figure 4, EDTN achieved excellent performance with fewer parameters. In detail, only 374K Params are spent during the reconstruction, higher than SRCNN (Dong et al., 2016b) and PAN (Zhao et al., 2020) but much less than VDSR (Kim et al., 2016b) and CARN (Ahn et al., 2020). On Set14 ×4 datasets, the proposed method is superior to the other methods.

4.3.2 Comparison results with advanced model experiments

Bicubic degradation is widely used to simulate the LR images obtained from image SR degradation. Therefore, bicubic interpolation is often used as the baseline for image

super-resolution experiments. To evaluate the effectiveness of EDTN, the researcher compares it with 10 state-of-the-art image SR methods, including SRCNN (Dong et al., 2016b), VDSR (Kim et al., 2016b), DRCN (Kim et al., 2016a), DRRN (Tai et al., 2017b), IDN (Chu et al., 2020), CARN (Ahn et al., 2020), IMDN (Zheng et al., 2019), PAN (Zhao et al., 2020), RFDN (Liu et al., 2020). For quantitative results on various scaling factors on datasets Set5, B100, Set14, Manga109, and Urban100 are detailed in Table 1. EDTN has slightly more parameters compared to PAN, indicating room for improvement. The experimental results show that EDTN achieves good experimental results on various datasets for each scaling factor. Overall, EDTN having fewer parameters and lower computational complexity.

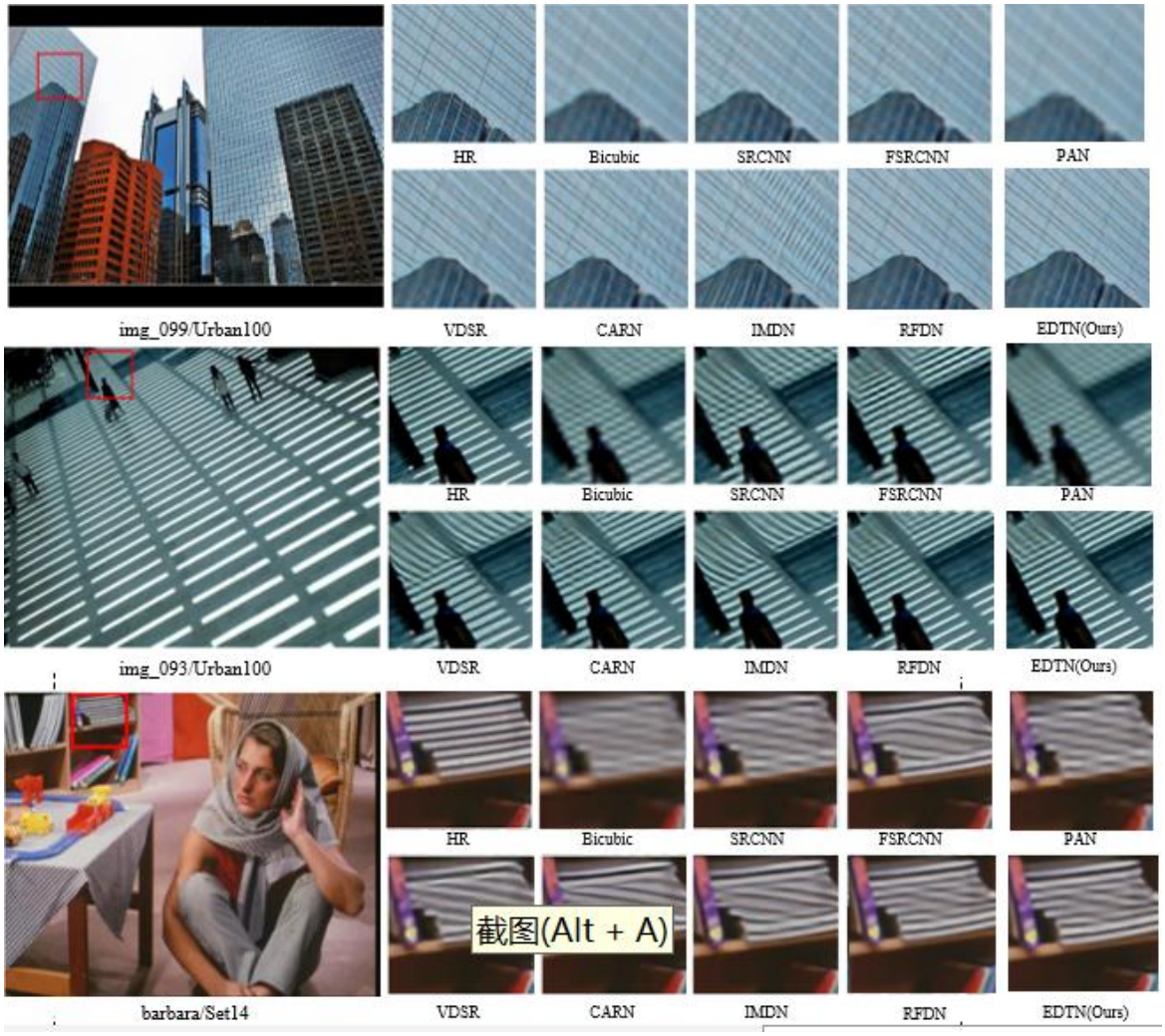


Fig. 5. Visual outcomes from compact super-resolution models at a magnification factor of ×4.

*4.3.3 Visual Outcomes of Recent Methods*

To demonstrate the preponderance of EDTN in terms of visual effects, as shown in Figure 5, the researcher presents the visualization results with a variety of methods. It is evident that most baseline models struggle to accurately reconstruct models, leading to significant aliasing issues. Oppositely, EDTN produces sharper visual effects and restores more high-frequency points. For instance, when considering the image img099/Urban100, many compared methods exhibit pronounced aliasing. Early developed methods such as Bicubic upsampling, VDSR (Kim et al., 2016b), DRRN (Tai et al., 2017b)0, and DRCN (Kim et al., 2016a) lose substantial structure due to the limited ineffective features and network depth. PAN (Zhao et al., 2020) can restore the silhouettes but fall short in restoring finer details. EDTN excels in recovering intricate details and enhancing edge sharpness, leading in superior visual quality. This improvement is attributed to its enhanced feature extraction efficiency and ability to obtain global information.

*4.4 Ablation Study*

*4.4.1 Contrast of different attention modules*

To demonstrate the effectiveness of Gated-Dconv feed-forward network (GDFN) and Multi-Dconv head transposed attention (MDTA), we analyze the impact of varying self-attention strategies and diverse feedforward networks on the performance of the model. The architecture of the baseline model is an SR network consisting of multiple EDTBs containing channel attention blocks.

It can be seen from the results in Table 2, the Channel and Spatial Attention Module (CBAM) block achieves a small improvement in performance compared to the baseline

model, specifically 0.04 dB. Meanwhile, the model with self-attention block MDTA shows an improvement of 0.15 dB compared to the baseline. In addition, GDFN performs better compared to other feedforward networks and successfully achieves a performance improvement of more than 0.1 dB and significantly reduces the computational cost.

**Table 2. Ablation studies of different attention. We report the PSNR (dB) values on Set14 datasets (×4). The best results are highlighted.**

| Block | Component | Params | PSNR |
|---|---|---|---|
| Baseline | Channel attention | 255K | 28.49 |
| Attention | Channel attention+ Spatial attention | 310K | 28.53 |
| Self- | MTA+FN | 770K | 28.59 |
| Attention | MDTA+FN | 723K | 28.64 |
| Overall | MDTA+GDFN | 374K | 28.66 |

*4.4.2 Contrast of different value of λ*

To find the most efficient value of λ, which is the parameter of L2 in EG loss. Since L2 loss calculates the average of the squares of the difference between the predicted and true values. L2 loss is more sensitive to larger errors compared to the L1 loss. If the values predicted by the model differ significantly from the actual discrete values, it may lead to faster adjustment of the model weights during the optimization process. L2 loss usually leads to faster convergence due to its derived gradient being smoother and continuous.

To be specific, λ is given different value under the same network architecture. As shown in Table 3, this paper treats λ=0 as a baseline experiment. On this basis, three different values are respectively assigned to λ, which are 1, 0.1, 0.01. L2 loss is more sensitive to discrete data but it converges faster, so this research expects to find a suitable value of λ to strike a balance and make the network more efficient. The experimental results show that the network's performance is the optimal when λ=0.01, while the parameter amount is smaller.

**Table 3. Ablation of different λ values through ablation studies is conducted, with PSNR (dB) measurements reported for the Set14 datasets at a magnification factor of ×4. The greatest results are highlighted.**

| $\lambda$ | Params | PSNR | SSIM |
|---|---|---|---|
| 0 | 370K | 28.39 | 0.7826 |
| 1 | 375K | 28.45 | 0.7827 |
| 0.1 | 382K | 28.21 | 0.7824 |
| 0.01 | 374K | 28.66 | 0.7831 |

## 5. CONCLUSIONS

In this paper, a lightweight Edge-enhanced Feature Distillation Transformer network is proposed to solve the Image Super-Resolution problem. The proposed model introduces an edge-enhanced convolution block which pays more attention to the edge information in the image. To enhance the use of contextual information from edge textures, an transformer is proposed which helps network capture consistent features across long distances. In addition, a novel gradient-variance loss is proposed to improve the ability to recover edge details and retain sharp edge visuals. Experimental results show that the proposed method achieves good performance under limited resource utilization. Furthermore, multiple experiments have demonstrated that the suggested approach attains an outstanding balance between parameters amount and visual quality.

ACKNOWLEDGEMENT

This work was partly supported by NSFC of China (U1504610, 62262005, 61962010), the High-level Innovative Talents in Guizhou Province (GCC [2023]033).

## REFERENCES

Abrahamyan, L., Truong, A. M., Philips, W., and Deligiannis, N. (2022). Gradient variance loss for structure-enhanced image super-resolution. arXiv preprint arXiv: 2202.00997.

Agustsson, E., and Timofte, R. (2017). Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pp. 126–135.

Bevilacqua, M., Roumy, A., Guillemot, C., and Alberi-Morel, M. L. (2012). Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, pp. 1–10.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In *European conference on computer vision, Springer*, pp. 213–229.

Chu, X., Zhang, B., and Xu, R. (2020). Multi-objective reinforced evolution in mobile neural architecture search. In *European Conference on Computer Vision*, pp. 99–113.

Chu, X., Zhang, B., Ma, H., Xu, R., and Li, Q. (2021). Fast, accurate and lightweight super-resolution with neural architecture search. In *ICPR*, pp. 59–64.

Dai, T., Cai, J., Zhang, Y., Xia, S. T., and Zhang, L. (2019). Second-order attention network for single image super-resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11057–11066.

Dai, X., Chen, Y., Yang, J., Zhang, P., and Yuan, L. (2021). Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2988–2997.

Ding, X., Guo, Y., Ding, G., and Han, J. (2019). Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *ICCV*, pp. 1911–1920.

Ding, X., Zhang, X., Han, J., and Ding, G. (2021a). Diverse branch block: Building a convolution as an inception-like unit. In *CVPR*, pp. 10886–10895.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., and Sun, J. (2021b). Repvgg: Making vgg-style convnets great again. In *CVPR*, pp. 13733–13742.

Dong, C., Loy, C. C., and Tang, X. (2016a). Accelerating the super-resolution convolutional neural network. In *ECCV*, pp. 391–407.

Dong, C., Loy, C. C., He, K., and Tang, X. (2014). Learning a deep convolutional network for image super-resolution. In: *ECCV (4). Lecture Notes in Computer Science*, Springer, 8692: 184–199.

Dong, C., Loy, C. C., He, K., and Tang, X. (2016b). Image super-resolution using deep convolutional networks. *IEEE TPAMI*, 38(2): 295–307.

Dosovitskiy, A., Beyer, L., and Kolesnikov, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Freeman, W. T., Pasztor, E. C., and Carmichael, O. T. (2000). Learning low-level vision. *International journal of computer vision*, 40(1): 25–47.

Gao, D., and Zhou, D. (2023). A very lightweight and efficient image super-resolution network. *Expert Systems with Applications*, 213: 118898.

Gao, Q., Zhao, Y., Li, G., and Tong, T. (2018). Image super-resolution using knowledge distillation. In *Asian Conference on Computer Vision*, pp. 527–541.

Huang, H., Shen, L., He, C., Dong, W., Huang, H., and Shi, G. (2021). Lightweight image super-resolution with hierarchical and differentiable neural architecture search. arXiv preprint arXiv: 2105.03939.

Huang, J. B., Singh, A., and Ahuja, N. (2015). Single image super-resolution from transformed self-exemplars. In *CVPR*, pp. 5197–5206.

Hui, Z., Gao, X., Yang, Y., and Wang, X. (2019). Lightweight image super-resolution with information multi-distillation network. In *Proceedings ofthe 27th acm international conference on multimedia*, pp. 2024–2032.

Hui, Z., Wang, X., and Gao, X. (2018). Fast and accurate single image super-resolution via information distillation network. In *Proceedings ofthe IEEE conference on computer vision and pattern recognition*, pp. 723–731.

Kim, J., Lee, J. K., and Lee, K. M. (2016a). Deeply-recursive convolutional network for image super-resolution. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1637–1645.

Kim, J., Lee, J. K., and Lee, K. M. (2016b). Accurate image super-resolution using very deep convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1646–1654.

Li, Y., Gu, S., Zhang, K., and Van Gool, L. (2020). Differentiable meta pruning via hyper networks. In *ECCV*, Springer, pp. 608–624.

Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K. (2017) Enhanced deep residual networks for single image super-resolution. In *CVPRW*, pp. 136–144.

Lin, Z., Garg, P., Banerjee, A., and Magid, S. (2022). Revisiting rcan: Improved training for image super-resolution. arXiv preprint arXiv:2201.11279.

Liu, J., Tang, J., and Wu, G. (2020). Residual feature distillation network for lightweight image super-resolution. In *European Conference on Computer Vision*, Springer, pp. 41–55.

Martin, D., Fowlkes, C., Tal, D., and Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, volume 2: 416–423.

Paszke, A., Gross, S., and Massa, F. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Shafiq, M., and Gu, Z. (2022). Deep residual learning for image recognition: A survey. *Applied Sciences*, 12(18): 8972.

Tai, Y., Yang, J., and Liu, X. (2017a). Image super-resolution via deep recursive residual network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2790–2798.

Tai, Y., Yang, J., and Liu, X. (2017b). A persistent memory network for image restoration. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4549–4557.

Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., and Liang, D. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 568–578.

Wang, Y. (2022). Edge-enhanced feature distillation network for efficient super-resolution. In P*roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 777–785.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M. H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728-5739.

Zeyde, R., Elad, M., and Protter, M. (2010). On single image scale-up using sparse-representations. In *7th International Conference on Curves and Surfaces*, volume 6920: 711–730.

Zhang, X., Zeng, H., and Zhang, L. (2021). Edge-oriented convolution block for real-time super resolution on mobile devices. In *ACMMM*, pp. 4034–4043.

Zhao, H., Kong, X., He, J., Qiao, Y., and Dong, C. (2020). Efficient image super-resolution using pixel attention. In *European Conference on Computer Vision*, Springer, pp. 56–72.

Zoph, B., and Le, Q. V. (2016). Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578.