RBSFormer: Enhanced Transformer Network for Raw Image Super-Resolution

Siyuan Jiang*, Senyan Xu*, Xingfu Wang†
University of Science and Technology of China
{syjiang, syxu}@mail.ustc.edu.cn, wangxfu@ustc.edu.cn

Abstract

In smartphones and mobile camera devices, the Image Signal Processor(ISP) is applied to reconstruct the RAW image into a sRGB image for human reading by a series of signal modules. Due to the non-linear ISP transformation, it is complicated to model the degradation in the sRGB domain. Most existing super-resolution methods directly handle the sRGB image processed by the ISP, introducing more difficult degradation patterns. To address this challenge, we propose an enhanced transformer network named RBSFormer. Unlike other methods that operate on sRGB images, RBSFormer takes RAW images as input, thus avoiding the complex degradation introduced by ISP processing. We design two enhanced core components, i.e., ***Enhanced Cross-Covairance Attention(EXCA)*** *and* ***Enhanced Gated Feed-forward Network(EGFN)****, in the RBSFormer, and we further introduce data augmentation in the RAW domain and hybrid ensemble strategies to enhance our results. Experimental results demonstrate superior performance against the majority of methods both qualitatively and quantitatively. Our RBSFormer achieves* ***3rd place*** *in terms of all the evaluation metrics both on the official validation and testing set with fewer parameters in the NTIRE 2024 challenge on Raw Image Super Resolution.*

1. Introduction

In recent years, there have been significant advancements in the field of image restoration and enhancement, driven by the rapid development of deep learning and computer vision technologies. It is a hot field that focuses on improving the quality of images by restoring degraded or corrupted parts and enhancing their overall appearance. It encompasses a wide range of techniques and algorithms aimed at addressing various issues such as noise reduction, image denoising[1, 4, 27], deblurring [24], color balance [19], and super-resolution [10, 30]. Super-resolution[10, 30] is a key field in image processing and computer vision aimed at enhancing the resolution of images beyond their original resolution. It encompasses various techniques, including Single Image Super-Resolution (SISR[49]), which enhances the resolution of a single image, Multi-Frame Super-Resolution (MFSR[50]), which utilizes multiple low-resolution images of the same scene, and Generative Adversarial Networks (GANs[21]), which generate high-quality images with enhanced details and textures. These techniques play a crucial role in applications such as medical imaging, surveillance, satellite imaging, and digital photography by improving image quality and visual clarity.

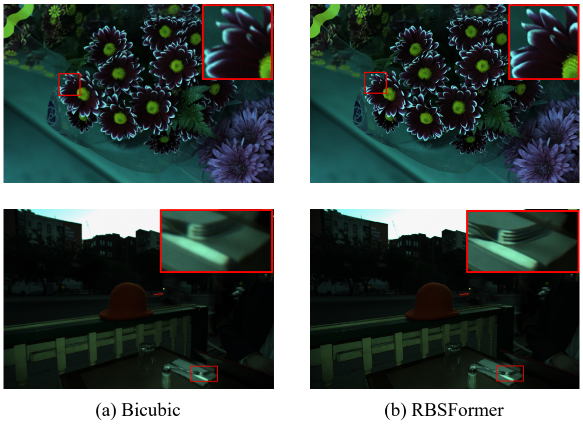

Figure 1. Visual Results of Raw Image Super-Resolution by the (a)Bicubic and (b)RBSFormer.

In the realm of smartphone and mobile camera technology, the Image Signal Processor (ISP) plays a crucial role in converting RAW images into sRGB images, which are easily interpretable by humans. However, the non-linear transformations applied by ISPs make it challenging to accurately model the degradation that occurs in the sRGB domain. Many existing super-resolution methods directly process the sRGB images, which introduces further complexities in handling the degradation patterns.

To tackle this challenge, in this work, we primarily denote the Single Image Super-Resolution problem in the RAW domain. The majority of cutting-edge Single Im-

*Co-first authors, †corresponding author.

age Super-Resolution (SISR) techniques operate on RGB imgaes[11] due to the greater abundance and accessibility of general-purpose sRGB images. However, theoretically, RAW format images are more suitable for handling super-resolution problems compared to RGB format images due to the following three advantages[11]: (i)Greater data range: RAW images carry a wider range of data compared to RGB images. (ii)the pixel values in RAW images directly represent the amount of light captured by the camera sensor, without any nonlinear transformations or color space conversions. (iii)During the process of converting RAW format to RGB through Image Signal Processing (ISP), non-linear transformations and information loss may occur.

The classical single-image super-resolution model (SISR)[16] is formulated as:

$$\mathbf{y} = (\mathbf{x} \otimes \mathbf{k}) \downarrow_{\mathbf{s}} + \mathbf{n} \tag{1}$$

Assuming the low-resolution (LR) image **y**, observed or captured, is derived from an underlying high-resolution (HR) image **x** by applying a degradation kernel (PSF) k [15], followed by downsampling operation $\downarrow_s$ with scale factor s (e.g., Bicubic [34]) and the addition of noise **n**.

In this work, we focus on the RAW image super-resolution challenge, aiming to up-sample a 4-channel RAW image, which may contain blur and/or noise. Considering the non-linearity of Image Signal Processing (ISP) transformations, we choose to model the degradation directly in the RAW domain rather than in the sRGB domain. Inspired by XCiT[3], Restormer[55] and InceptionNeXT[51], we design an enhanced transformer network to construct a realistic RAW degradation pipeline. The main contributions of this paper are:

1. We present an enhanced transformer network named **RBSFormer** for raw image blind super-resolution while restoring images with various types of degradation.
2. We devise **Enhanced Cross-Covairance Attention(EXCA) and Enhanced Gated Feed-forward Network(EGFN)** by introducing the cross-covariance attention module and applying Inception Depth-wise convolution for better context representing learning.
3. Experimental results show that the proposed method significantly outperforms other solutions. In the NTIRE-challenge 2024 Raw Image Super Resolution track, our RBSFormer achieves **3rd** place on the official validation and testing set.

2. Related work

Vision Transformers. The Transformer model, initially developed for sequence processing in natural language tasks [36], has found widespread adoption in various vision tasks, including image recognition [35, 53], segmentation [38], and object detection [6, 26]. Vision Transformers [13, 35] decompose images into sequences of patches and learn their interrelationships, exhibiting strong capabilities in capturing long-range dependencies within image patch sequences [22].

This adaptability has extended to low-level vision tasks such as super-resolution [25, 37, 44, 48], image colorization [23], denoising [7, 42, 44], and deraining [42, 44]. However, the self-attention mechanism in Transformers can become computationally prohibitive for high-resolution images due to its quadratic complexity with the number of image patches.

Cross-covariance attention[3] has linear complexity in the number of tokens. Recent methods[37, 55] in low-level image processing employ this strategy to mitigate this complexity.

One approach is to utilize self-attention within local image regions [25, 42] using the Swin Transformer design [25]. However, this approach restricts context aggregation to local neighborhoods [55], which might not be optimal for image restoration tasks, as it undermines the primary motivation for using self-attention over convolutions.

Raw Image Super-Resolution. Existing super-resolution methods primarily focus on upscaling sRGB images, but the complexities of modeling degradation in this domain hinder their effectiveness. Zhang et al.[58] demonstrate the advantages of utilizing real RAW sensor data for machine learning-based digital zoom, highlighting the limitations of existing methods that operate on processed sRGB images. Xu et al.[46] implement Eq.1 and propose a pipeline for generating realistic training data, which results in their model's superior performance for real-world scenarios. Conde et al. [11] introduce a novel degradation pipeline and a corresponding BSRAW model, and address blind image super-resolution directly in the RAW domain. In this work, we introduce the degradation pipeline proposed in previous work into our data augmentation, and aim to identify efficient Transformer mechanisms that are genuinely suitable for processing data in the RAW domain.

3. Method

3.1. Overall Pipeline

As shown in Fig 2, RBSFormer consists of three parts, i.e., shallow feature extraction, deep feature extraction and raw reconstruction. Firstly, given a raw low-resolution image with degradation $\mathbf{I_{LR}} \in \mathbb{R}^{H \times W \times 4}$. We apply shallow feature filter to obtain feature encoding $\mathbf{F_s} \in \mathbb{R}^{H \times W \times C}$ as:

$$F_s = Conv_{3\times3}(I_{LR}) \tag{2}$$

where $Conv_{3\times3}(\cdot)$ is 3×3 convolution. Then we use K transformer blocks and one 3×3 convolution layer in a cascade manner to extract deep features. Such a process can

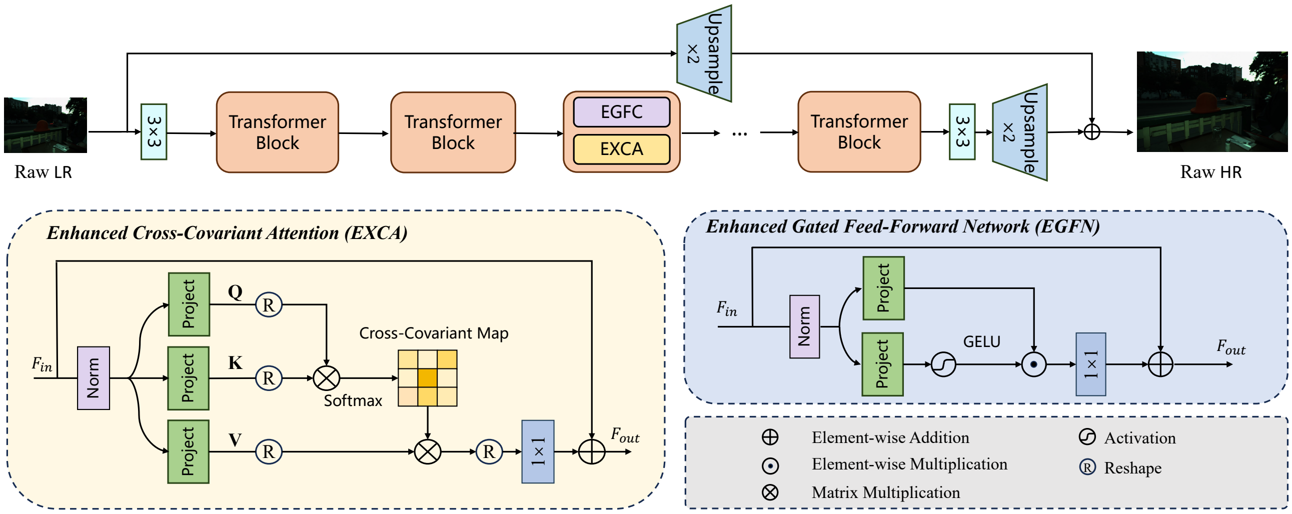

Figure 2. The architecture of the proposed RBSFormer for RAW image super-resolution. Our RBSFormer consists of enhanced transformer blocks. The core modules are Enhanced Cross-Covairance Attention(EXCA) and Enhanced Gated Feed-forward Network(EGFN).

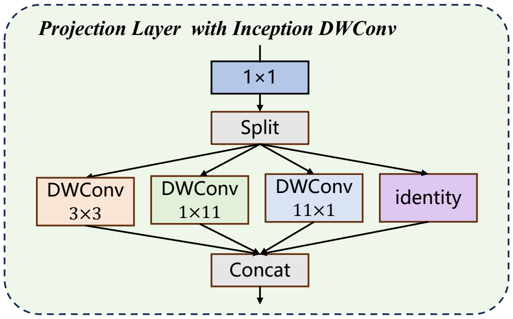

Figure 3. The architecture of Project Layer with Inception Depthwise convolution.

be expressed as:

$$F_i = \mathcal{H}_{tb_i}(F_{i-1}), i = 1, 2, \dots, K \tag{3}$$

$$F_d = Conv_{3\times3}(F_K) \tag{4}$$

where $\mathcal{H}_{tb_i}$ denotes the i-th transformer block and $\{F_i | i = 1, 2, \dots, K\}$ represent intermediate features. In many studies [25], a 3×3 convolution layer is additionally employed at the end of deep feature extraction to enhance feature aggregation. Sections 3.2 and 3.3 will provide a detailed explanation of the specific components comprising the transformer block.

Finally, PixelShuffle[33] is applied to upsample the deep feature, then RBSFormer reconstructs the HR image I_{HR} by aggregating initial input and deep features as

$$\begin{aligned} I_{HR} &= \mathcal{H}_{rec}(I_{LR}, F_d) \\ &= Up(F_s + F_d) \end{aligned} \tag{5}$$

where $\mathcal{H}_{rec}$ is the reconstruction module and $Up(\cdot)$ denotes the PixelShuffle operation. With a long residual connection, shallow features which mostly contain low-frequency information can be directly applied to reconstruct the HR image. It can help the deep feature module focus on its specific ability to extract high-frequency information[25].

3.2. Enhanced Cross-Covariance Attention

The computational cost of the self-attention layer accounts for most of the Transformers. In traditional self-attention [14, 36], it can be regarded as a specific spatial attention that calculates attention via the key-query dot-product across the spatial dimension, i.e., for an image with a resolution of $H \times W$, its computational cost is $O(H^2W^2)$. Some studies[29, 43] have proved spatial self-attention is beneficial for SR tasks in the sRGB domain. However, the spatial discontinuity inherent in the RAW domain undermines the efficacy of employing spatial self-attention mechanisms on raw data, resulting in diminished performance compared to their application on the sRGB images. Additionally, as discussed above, the introduction of such mechanisms incurs substantial memory and computational overheads. Hence, it is not a suitable method for the RAW image super-resolution task.

Inspired by [3, 55], we introduce an enhanced cross-covariance attention named EXCA across channel dimension, as shown in Fig 2. In EXCA, the computational complexity of the image is linearly related to the spatial resolution. Specifically, to emphasize local contextual information, we introduce depth-wise convolutional operations which are widely applied in transformer blocks. Subsequently, it computes the cross-covariance across channels to generate implicit attention feature maps about global context, thereby achieving simultaneous channel re-weighting and information transmission. Different from [55], we introduce Inception Depth-wise convolution(I-DWConv)[51] to project the spatial context, enhancing local feature representation. I-DWConv decomposes large-kernel depth-wise convolution into four parallel branches along channel dimension, i.e., small square kernels, two orthogonal band kernels, and an identity mapping, as shown in Fig 3. It can maintain a similar throughput to a 3 × 3 depth-wise convolution but achieves better competitive performance due to its ability to capture richer contextual information within larger local receptive fields.

Given an input feature map $F_{in} \in \mathbb{R}^{H \times W \times C}$, EXCA utilizes 1×1 point-wise convolutions for aggregating cross-channel context at the pixel level, followed by I-DWConv to encode spatial context on a channel-wise basis. The projection layer can be expressed as:

$$H_{proj}(\cdot) = IDWConv(Conv_{1\times1}(\cdot)) \tag{6}$$

Next, we obtain query $(\mathbf{Q})$, key $(\mathbf{K})$ and value $(\mathbf{V})$ projections, where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{H \times W \times C}$. Next, by reshaping the Q and K values, we obtain $\mathbf{Q^r}, \mathbf{K^r} \in \mathbb{R}^{HW \times C}$. Then, the dot product of $\mathbf{Q^r}$ and $\mathbf{K^r}$ yields a channel-wise cross-covariance attention map of size $\mathbb{R}^{C \times C}$, differing in size from traditional self-attention map of size $\mathbb{R}^{HW \times HW}$.The whole EXCA process is formulated as follows:

$$\begin{aligned}
Q &= H^{Q}_{proj}(LN(F_{in})) \\
K &= H^{K}_{proj}(LN(F_{in})) \\
V &= H^{V}_{proj}(LN(F_{in})) \\
\hat{Q} &= \mathcal{R}(Q), K^r = \mathcal{R}(K), V^r = \mathcal{R}(V), \\
M_t &= \mathcal{A}(Q^r, K^r, V^r) = V^r \cdot \mathrm{SoftMax}(K^r \cdot Q^{rT}/\alpha) \\
F_{out} &= Conv_{1\times1}(R)
\end{aligned} \tag{7}$$

where F_{out} is the output feature map, LN represents layer normalization, and $\mathcal{R}(\cdot)$ denotes the reshape operation. Overall, EXCA facilitates the aggregation of both local and non-local related pixels, enabling efficient processing of high-resolution images.

3.3. Enhanced Gated Feed-Forward Network

Feed-Forward Network(FFN) is widely applied in the design of transformer block[14, 36], which facilitates the model's ability to capture and process information from local contexts within the input sequence. It usually consists of two 1×1 convolutional layers. Specifically, the first convolutional layer is utilized to increase the feature dimension to a higher dimension, while the second convolutional layer is employed to reduce it back to the original dimension. Non-linear activation functions are typically applied between these layers. In some studies [55], the gating mechanism proved effective due to its better ability in representation learning. It performs controlled feature transformation, which suppresses less informative features, allowing only useful information to propagate further through the network hierarchy.

Specifically, the gating mechanism is designed as the element-wise product of projection layers' outputs, one of which is activated by the GELU[18] non-linear activation function. It can be approximately written as:

$$GELU(x) \approx 0.5x(1 + \tan h[\sqrt{2/\pi(x + \gamma x^3)}]) \tag{8}$$

Similar to EXCA, EGFN also employs inception depth-wise convolution operations. This operation enables the encoding of information from neighboring pixels in the spatial domain, aiding in the learning of local image structures. The EGFN could be formulated as:

$$\mathbf{F_{out}} = Conv_{1\times1}(\mathrm{Gating}(\mathbf{F_{in}})) + \mathbf{F_{in}} \tag{9}$$

$$\mathrm{Gating}(\cdot) = \phi(H_{proj}(\mathrm{LN}(\cdot)) \odot H_{proj}(\mathrm{LN}(\cdot)) \tag{10}$$

where F_{in} is the input feature map, $\odot$ represents element-wise multiplication, ϕ represents the GELU non-linearity, and $LN(\cdot)$ is the layer normalization. In general, EGFN controls the information flow through EXCA layers in the network, enabling layers within the network to focus on finer image attributes, thereby producing high-quality outputs.

3.4. Loss Functions

In our work, we use the Charbonnier loss [39] to optimize our network. This loss function is particularly effective for handling outliers and robust to noise. Its formulation is as follows:"

$$\mathcal{L}_{\text{content}} = \sqrt{\left\| \hat{I}_{\text{HR}} - I_{\text{HR}} \right\|_2 + \epsilon^2}, \tag{11}$$

where $\hat{I}_{\text{HR}}$ is the predicted HR raw image, I_{HR} is the ground truth, and ϵ is set to 0.0001 as default.

In addition to the content loss, we leverage frequency domain information to introduce auxiliary loss to our network, which is defined as follows:

$$\mathcal{L}_{\text{frequency}} = \left\| \mathcal{F}\left(\hat{I}_{\text{HR}}\right) - \mathcal{F}\left(I_{\text{HR}}\right) \right\|_1, \tag{12}$$

Table 1. **PSNR/SSIM** results of NTIRE 2024 RAW Image Super Resolution Challenge on the validation set (40 images), the complete testing set (200 images), and the testing set at full-resolution (12MP) RAW images [12]. "NA" indicates the results are not available for the method.

Rank	**Team**	**Method**	**Validation 1MP**	**Test 1MP**	**Test 12MP**	**# Params. (M)**
1	Samsung	2-Stage w/ FPL	$43.40/0.99_{(1)}$	$43.443/0.986_{(1)}$	$43.858/0.988_{(1)}$	$53.7_{(6)}$
2	XiaomiMMAI	EffectiveSR	$43.38/0.99_{(2)}$	$43.249/0.986_{(2)}$	NA	$20.9_{(5)}$
3	**USTCX(Ours)**	RBSFormer	$43.21/0.99_{(3)}$	$42.493/0.984_{(3)}$	$43.649/0.987_{(2)}$	$3.19_{(3)}$
4	McMaster	SwinFSR Raw	$42.48/0.98_{(4)}$	$42.366/0.984_{(4)}$	NA	$6.64_{(4)}$
5	-	BSRAW [11]	$42.25/0.98_{(5)}$	$42.106/0.984_{(5)}$	$42.853/0.986_{(3)}$	$1.50_{(2)}$
6	NUDT RSR	SAFMN FFT	$41.81/0.98_{(6)}$	$41.621/0.982_{(6)}$	NA	$0.27_{(1)}$
7	-	Interpolation [11]	$35.95/0.95_{(7)}$	$36.038/0.952_{(7)}$	$36.926/0.956_{(4)}$	-

where $\mathcal{F}(\cdot)$ indicates the Fast Fourier Transform (FFT). Finally, the total loss could be defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{content}} + \lambda \mathcal{L}_{\text{frequency}} \tag{13}$$

where λ denotes the balanced weight, and we empirically set λ to 0.5 as default.

4. Experiment

4.1. Dataset

We conduct the experiments strictly following the setting of the NTIRE-Challenge 2024 Raw Image Super Resolution track[12]. The training data contains about 1000 RAW images from DSLR cameras. The images have been filtered, normalized, white-black level corrected, and in 4-channel format (RGGB). The training data only provides clear RAW images. The input format is a ".npz" file including keys "raw" and "max_val", indicating the raw image array and the max value of the raw image, respectively. The validation and testing sets consist of 40 and 200 low-quality(1MP) RAW images each. Besides, the 12MP testing set consists of 200 full-resolution images. Each set contains RAW images of unknown degradation.

4.2. Data Augmentation

Since the competition's official dataset doesn't provide LR-HR pairs, it is challenging to synthesize realistic degraded RAW images. Inspired by [11], we introduce its degradation pipeline on RAW images. Firstly, the input in the test/validation split contains degradation. To enhance the robustness and generalization of the model, we need to introduce degradation to the original HR images. In this paper, we refer to [11] and introduce noise and blur degradation.

Noise. We utilize a more pragmatic shot-read noise model[4, 59]. In Equation 14, the intensity y is depicted as a sample from a Gaussian distribution with the input signal x as its mean and variance determined by the parameters λ_s (shot) and λ_r (read) [4]. This model is derived from a Poisson-Gaussian noise model [40] and can be formulated as:

$$y \sim \mathcal{N}\left(\mu = x, \sigma^2 = \lambda_r + \lambda_s x\right) \tag{14}$$

Blur. It is a prevalent degradation observed during image acquisition, such as camera shake in mobile photography, motion blur, and defocus blur [2, 20, 52]. Obtaining aligned blurry-clean pairs in real-world scenarios is challenging, leading to the preference for synthetic datasets in deblurring tasks. Many existing methods employ a uniform blur by convolving images with iso/anisotropic Gaussian kernels [57]. For instance, some studies[46, 47] utilized a disk kernel for defocus blur and introduced modest motion blur [32]. In our work, followed by [11], we model blur degradation by convolving images with a diverse range of kernels, including classical isotropic and anisotropic Gaussian blur kernels [57], as well as real estimated motion blur kernels [28, 31]and PSFs (point-spread-functions) from real data [45, 56].

Finally, for SR tasks in the RAW domain, we follow [11, 46, 57] to synthesize LR RAW images from the assumed HR real captures. We downsample the high-quality RAW images to obtain the low-resolution raw images so that each pixel could have its ground truth red, green and blue values.

4.3. Implementation Details

We implement our proposed approach via the PyTorch 1.8 platform. Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$ is adopted to optimize our network.The initial learning rate is 3×10^{-4} and changes with Cosine Annealing scheme to 1×10^{-7}, including 120K iterations in total. We start training with patch size 156×156 and batch size 8. For data augmentation, we use the data augmentation tech-

Table 2. Quantitative comparisons of methods on the Validation Set at NTIRE 2024 RAW Image Super Resolution Challenge. The MACs is computed using a 224 × 224 image as input. The best and the second results are boldfaced and underlined, respectively.

Models	**Basic Block**	**# Param.(M)**	**MACs(G)**	**Metrics**	
				PSNR↑	SSIM↑
Model-A	Dual Attention Block(Simple SA&CA)[54]	3.22	183.5	42.30	0.98
Model-B-S	Simple Gate&Simplified CA[8]	0.98	41.9	40.79	0.98
Model-B-L(default)	Simple Gate&Simplified CA[8]	4.31	184.3	42.16	0.98
Model-C	MDTA&GDFA[55]	3.31	166.2	42.49	0.99
Ours	EXCA&EGFA	3.19	158.4	**42.67**	**0.99**

Table 3. Ablation study of loss functions. The best and the second results are boldfaced and underlined, respectively.

Losses			**Metrics**	
L1	Charbonnier L1	Frequency	PSNR↑	SSIM↑
✓			42.46	0.98
	✓		42.52	0.98
✓		✓	42.61	0.98
	✓	✓	**42.67**	**0.99**

Table 4. Ablation study of ensemble strategies. The best and the second results are boldfaced and underlined, respectively.

Ensemble Strategies		**Metrics**	
Multi-Scale	Multi-Config	PSNR↑	SSIM↑
		42.66	0.99
✓		42.92	0.99
	✓	42.96	0.99
✓	✓	**43.21**	**0.99**

nology mentioned above. Our training is performed on the NVIDIA 4090 device.

4.4. Evaluation Metrics

In this work, we utilize two established reference-based metrics, widely employed in similar tasks[11, 46, 47, 55], namely Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) [41], to evaluate the effectiveness of our approach in raw image restoration. Higher PSNR and SSIM values indicate superior performance in raw image super-resolution tasks.

4.5. Comparisons

Table 1 presents a comprehensive comparison of various solutions for the NTIRE 2024 RAW Image Super Resolution Challenge. The proposed RBSFormer achieves **3rd** place in terms of all the evaluation metrics both on the official validation and testing set. In terms of performance on full-resolution predictions at 12 megapixels, RBSFormer achieves **2nd** place with significantly fewer parameters. Note that the top-ranking team exceeded our parameter count by a factor of 16. However, it's worth highlighting that despite this considerable difference, the proposed approach is only 0.21dB lower than the first-place team in terms of PSNR. Evidently, our method surpasses the fourth-ranked team 0.73dB and 0.172dB in terms of PSNR on the validation set and testing set, respectively.

Besides, in Table 2, we demonstrate comparable performance methods on the official validation datasets when compared to some ISP methods and general image restoration methods. Note that due to the methods that can directly apply for raw image super-resolution are rare, we reconstruct these models using their basic block, i.e., Simple Gate and Simplified CA in NAFNet[8, 9], Dual Attention Block in CycleISP[54] and DWConv-based Transformer Block in Restormer[55]. Our method consistently demonstrates outstanding performance. Compared to the methods Model-C and Model-A, we obtain 0.18dB and 0.37dB gain in PSNR. Besides, in Fig. 4, to more intuitively show our excellent performance, we compare the visual quality between our predicted output and other models. The comparison clearly demonstrates that our technology produces superior visual results and outperforms others in terms of visual quality, demonstrating its efficiency in image restoration.

4.6. Ensemble Strategies

Ensemble learning is a potent and adaptable technique capable of enhancing the performance and dependability of predictive models across various applications. As discussed above, the proposed approach demonstrates promising performance. With the aim of producing more robust and diverse predicted results, we introduce ensemble learning.

Motivated by [5, 17], we adopt two different ensemble strategies. Our approach consists of two main ensemble

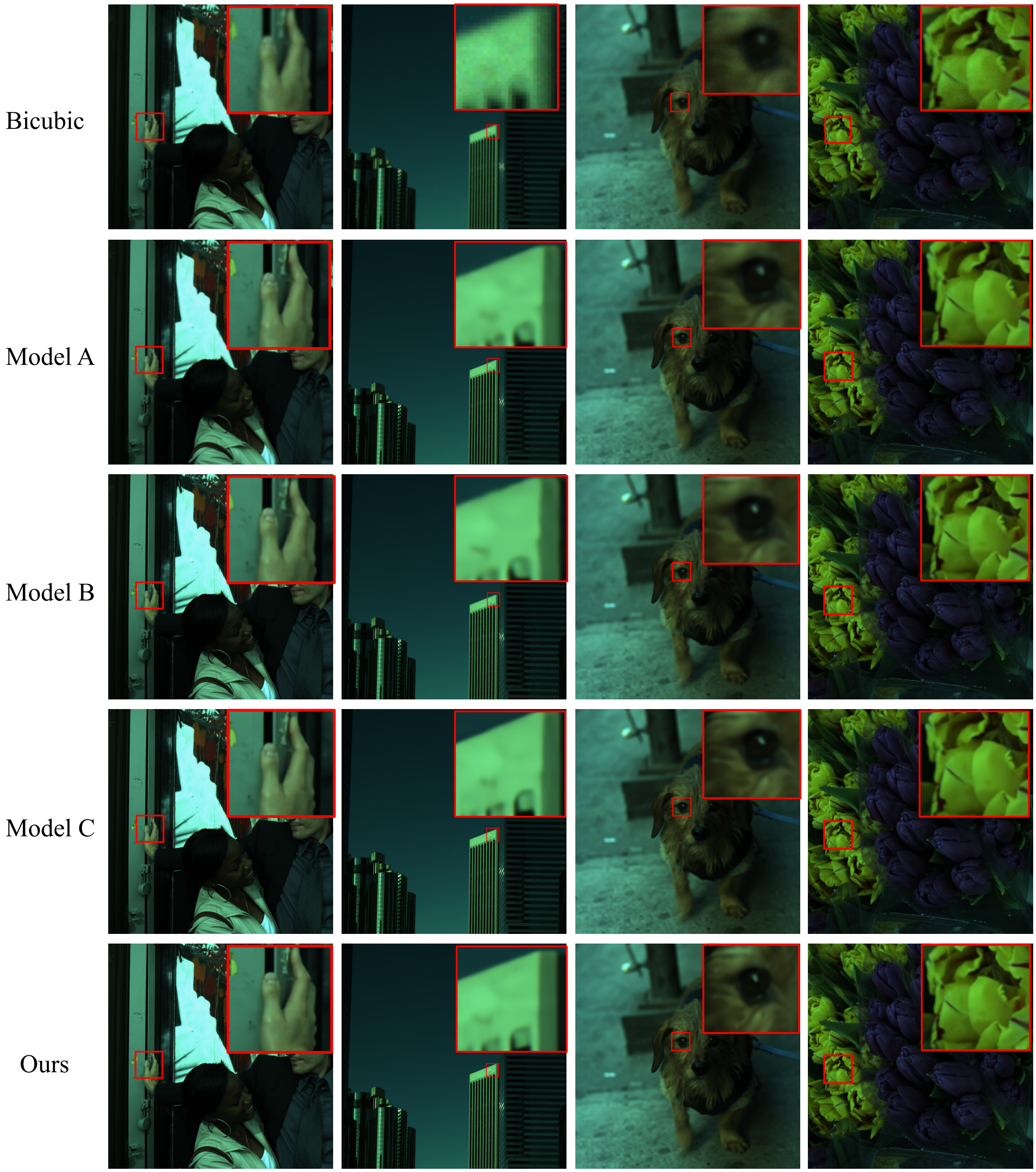

Figure 4. Visual Results of Bicubic, Model A, Model B, Model C and the proposed RBSFormer.

strategies. First, we utilize a multi-scale ensemble, which involves training the same configuration of the model with different patch sizes and then averaging the outputs to enhance the restoration quality.

Secondly, we implement a top-k multi-configuration ensemble approach, where we vary the number of modules and channel configurations within the same model architecture and then aggregate their outputs. These ensemble techniques offer several benefits. The multi-configuration ensemble allows for improved robustness by integrating diverse configurations, leveraging their strengths. Moreover, the top-k multi-model ensemble enhances model diversity and generalizability by combining different configurations and providing a more comprehensive representation of underlying data patterns.

4.7. Ablation Studies

We conduct extensive experiments to verify the effects of each component of our method, e.g., modules, strategies and losses.

Impact of Loss Functions. The network performance with different losses is reported in Table 3. We observe that the combination of Charbonnier L1 loss and frequency loss yields the best performance. Compared to the second-best combination of L1 loss and Frequency loss, there is an increase of 0.06dB in PSNR, along with an improvement in SSIM. Experimental results also indicate that integrating Frequency loss indeed enhances the network's performance.

Impact of Ensemble Strategies. From Table 4, we can see that employing the Multi-Scale strategy alone leads to a PSNR increase of 0.27dB, while using the Multi-Config strategy alone results in a PSNR increase of 0.31dB. When we simultaneously utilize both the Multi-Scale and Multi-Config strategies, a gain of 0.6dB in PSNR was achieved. Additionally, we also observe that these ensemble strategies impact the SSIM metric values.

5. Conclusion

In this paper, we present RBSFormer, an effective transformer network for RAW image super-resolution. We developed the Enhanced Cross-Covariance Attention (EXCA) and Enhanced Gated Feed-forward Network (EGFN) by introducing the cross-covariance attention module and applying Inception Depth-wise convolution to improve context representation learning. Additionally, the adoption of data augmentation strategies and ensemble strategies further improves the model's robustness and effectiveness. Finally, RBSFormer achieves 3rd place in terms of all the evaluation metrics both on the official validation and testing set with fewer parameters in the NTIRE 2024 challenge on Raw Image Super Resolution.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1692–1700, 2018. 1

[2] Abdullah Abuolaim, Mahmoud Afifi, and Michael S Brown. Improving single-image defocus deblurring: How dual-pixel images help through multi-task learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1231–1239, 2022. 5

[3] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, et al. Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027, 2021. 2, 3

[4] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11036–11045, 2019. 1, 5

[5] Yuanhao Cai, Jing Lin, Xiaowan Hu, Haoqian Wang, Xin Yuan, Yulun Zhang, Radu Timofte, and Luc Van Gool. Mask-guided spectral-wise transformer for efficient hyperspectral image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17502–17511, 2022. 6

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 2

[8] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 6

[9] Xiaojie Chu, Liangyu Chen, and Wenqing Yu. Nafssr: Stereo image super-resolution using nafnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1239–1248, 2022. 6

[10] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration. In *European Conference on Computer Vision*, pages 669–687. Springer, 2022. 1

[11] Marcos V Conde, Florin Vasluianu, and Radu Timofte. Bsraw: Improving blind raw image super-resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8500–8510, 2024. 2, 5, 6

[12] Marcos V Conde, Florin Vasluianu, and Radu Timofte. Deep raw image super-resolution. a ntire 2024 challenge survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 5

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Trans-

formers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4

[15] Netalee Efrat, Daniel Glasner, Alexander Apartsin, Boaz Nadler, and Anat Levin. Accurate blur models vs. image priors in single image super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2832–2839, 2013. 2

[16] Michael Elad and Arie Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IEEE transactions on image processing*, 6(12):1646–1658, 1997. 2

[17] Zhihao Fan, Xun Wu, Fanqing Meng, Yaqi Wu, and Feng Zhang. Otst: A two-phase framework for joint denoising and remosaicing in rgbw cfa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2832–2841, 2023. 6

[18] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 4

[19] Daniel Hernandez-Juarez, Sarah Parisot, Benjamin Busam, Ales Leonardis, Gregory Slabaugh, and Steven McDonagh. A multi-hypothesis approach to color constancy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2270–2280, 2020. 1

[20] Mahdi S Hosseini and Konstantinos N Plataniotis. Convolutional deblurring for natural imaging. *IEEE Transactions on Image Processing*, 29:250–264, 2019. 5

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[22] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 2

[23] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. *arXiv preprint arXiv:2102.04432*, 2021. 2

[24] Chih-Hung Liang, Yu-An Chen, Yueh-Cheng Liu, and Winston H Hsu. Raw image deblurring. *IEEE Transactions on Multimedia*, 24:61–72, 2020. 1

[25] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 2, 3

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[27] Jiaqi Ma, Shengyuan Yan, Lefei Zhang, Guoli Wang, and Qian Zhang. Elmformer: Efficient raw image restoration with a locally multiplicative transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5842–5852, 2022. 1

[28] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1628–1636, 2016. 5

[29] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018. 3

[30] Guocheng Qian, Yuanhao Wang, Chao Dong, Jimmy S Ren, Wolfgang Heidrich, Bernard Ghanem, and Jinjin Gu. Rethinking the pipeline of demosaicing, denoising and superresolution. *arXiv e-prints*, pages arXiv–1905, 2019. 1

[31] Dongwei Ren, Kai Zhang, Qilong Wang, Qinghua Hu, and Wangmeng Zuo. Neural blind deconvolution using deep priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3341–3350, 2020. 5

[32] Christian J Schuler, Harold Christopher Burger, Stefan Harmeling, and Bernhard Scholkopf. A machine learning approach for non-blind image deconvolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1067–1074, 2013. 5

[33] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 3

[34] Radu Timofte, Vincent De Smet, and Luc Van Gool. A+: Adjusted anchored neighborhood regression for fast superresolution. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part IV 12*, pages 111–126. Springer, 2015. 2

[35] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3, 4

[37] Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, and Jinfan Liu. Omni aggregation networks for lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22387, 2023. 2

[38] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 568–578, 2021. 2

[39] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced

deformable convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 4

[40] Yuzhi Wang, Haibin Huang, Qin Xu, Jiaming Liu, Yiqun Liu, and Jue Wang. Practical deep raw image denoising on mobile devices. In *European Conference on Computer Vision*, pages 1–16. Springer, 2020. 5

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[42] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 2

[43] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 3

[44] Jie Xiao, Xueyang Fu, Man Zhou, Hongjian Liu, and Zheng-Jun Zha. Random shuffle transformer for image restoration. In *International Conference on Machine Learning*, pages 38039–38058. PMLR, 2023. 2

[45] Li Xu and Jiaya Jia. Two-phase kernel estimation for robust motion deblurring. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11*, pages 157–170. Springer, 2010. 5

[46] Xiangyu Xu, Yongrui Ma, and Wenxiu Sun. Towards real scene super-resolution with raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1723–1731, 2019. 2, 5, 6

[47] Xiangyu Xu, Yongrui Ma, Wenxiu Sun, and Ming-Hsuan Yang. Exploiting raw images for real-scene super-resolution. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):1905–1921, 2020. 5, 6

[48] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5791–5800, 2020. 2

[49] Wenming Yang, Xuechen Zhang, Yapeng Tian, Wei Wang, Jing-Hao Xue, and Qingmin Liao. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12):3106–3121, 2019. 1

[50] Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Learning a deep convolutional network for light-field image super-resolution. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 24–32, 2015. 1

[51] Weihao Yu, Pan Zhou, Shuicheng Yan, and Xinchao Wang. Inceptionnext: when inception meets convnext. *arXiv preprint arXiv:2303.16900*, 2023. 2, 4

[52] Lu Yuan, Jian Sun, Long Quan, and Heung-Yeung Shum. Image deblurring with blurred/noisy image pairs. In *ACM SIGGRAPH 2007 papers*, pages 1–es. 2007. 5

[53] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567, 2021. 2

[54] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2696–2705, 2020. 6

[55] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 2, 3, 4, 6

[56] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3217–3226, 2020. 5

[57] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 5

[58] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2019. 2

[59] Yi Zhang, Hongwei Qin, Xiaogang Wang, and Hongsheng Li. Rethinking noise synthesis and modeling in raw denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4593–4601, 2021. 5