Deep RAW Image Super-Resolution. A NTIRE 2024 Challenge Survey

Marcos V. Conde * Florin-Alexandru Vasluianu * Radu Timofte * Jianxing Zhang
Jia Li Fan Wang Xiaopeng Li Zikun Liu Hyunhee Park Sejun Song
Changho Kim Zhijuan Huang Hongyuan Yu Cheng Wan Wending Xiang
Jiamin Lin Hang Zhong Qiaosong Zhang Yue Sun Xuanwu Yin Kunlong Zuo
Senyan Xu Siyuan Jiang Zhijing Sun Jiaying Zhu Liangyan Li Ke Chen
Yunzhe Li Yimo Ning Guanhua Zhao Jun Chen Jinyang Yu Kele Xu
Qisheng Xu Yong Dou

Abstract

This paper reviews the NTIRE 2024 RAW Image Super-Resolution Challenge, highlighting the proposed solutions and results. New methods for RAW Super-Resolution could be essential in modern Image Signal Processing (ISP) pipelines, however, this problem is not as explored as in the RGB domain. Th goal of this challenge is to upscale RAW Bayer images by 2x, considering unknown degradations such as noise and blur. In the challenge, a total of 230 participants registered, and 45 submitted results during thee challenge period. The performance of the top-5 submissions is reviewed and provided here as a gauge for the current state-of-the-art in RAW Image Super-Resolution.

1. Introduction

RAW Image Super-Resolution represents an active research direction, aiming at upscaling hardware-specific RAW image representations, while dealing with hardware characteristic properties, often depending on the technical implementation of the camera product. The lack of standardization in camera Image Processing Signal (ISP) implementation induces plenty variety into the camera market segment, with plenty corrections made through image processing algorithms [17, 18, 24, 25, 53], overcoming the hardware limitations of various devices.

The RAW image represents the discretized and quantized representation of the image signal. Of course, both the aforementioned operation depend on the sensor nature, with different sensor types, different spatial resolutions or color resolutions (the number of bits used for quantization). Naturally, portable camera devices, that are usually subject to extreme limitations in terms of size, power supply, and used optics are constricted in terms of achieved image quality, usually employing low color resolutions and spatial resolutions. However, these specific hardware implementations are currently leading the market in terms of new acquired devices, with cameras becoming ubiquitous and universal accessible. Therefore, the image information space is also characterized by a high availability of images corresponding to these devices, with complex ISP systems [15, 23, 43, 59] mapping the RAW image representation to a perceptually meaningful RGB counterpart.

RAW image upscaling remains also relevant in the professional photography field. Given the strong development of Internet-based image transfer services, the drive for high resolution images is now higher than ever before. However, building a professional photography or videography setup is a difficult challenge, since the market is divided by various application-specific parameters, with the acquisition costs remaining the main limiting factor for a new investment. Thus, it is important to develop algorithms which are robust to the various hardware limitations affecting the low-cost amateur photography systems, matching the characteristics of the high resolution sensors used in highly professional applications.

RAW information is extremely important, since the RAW data directly correlates with scene radiance, with the discretization and quantization being the only non linear operations affecting the naturally continuous radiance signal measure in a photography. This property of the RAW data is extremely beneficial in analyzing typical image degradations as acquisition noise or image blur. Futher away from the sensor representation, ISP represents a sequence of highly non-linear operations, happening with irrevocable loss of information [5, 28]. This further complicates the image restoration task, with complex algorithms needing every

∗ Marcos V. Conde († corresponding author), Florin-Alexandru Vasluianu, amd Radu Timofte are the challenge organizers, while the other authors participated in the challenge.
∗ University of Würzburg, CAIDAS & IFI, Computer Vision Lab.
NTIRE 2024 webpage: https://cvlai.net/ntire/2024.
Code: https://github.com/mv-lab/AISP

Figure 1. Samples of the **NTIRE 2024 RAW Image Super-Resolution Challenge** testing set.

achievable bit of signal based variance [5, 53]. Considering all the aforementioned factors, RAW image processing poses significant advantages over the standard sRGB representation, with superior performance in a multitude of low-level imagery applications like image denoising [1, 5, 39], deblurring [18], exposure adjustment [22], and image super-resolution [18, 40, 53, 56, 59].

Due to the lack of standardization at the hardware implementation level, RAW images are characterized by vendor or even product specific properties, which are explainable only with private, with the photography products shielded by implementation patents or trade secrets. Coupled to the abundance of the standard sRGB representation, most of the existing high complexity image restoration algorithms [10, 17, 32, 58] are specifically designed for compressed or uncompressed RGB image or video.

As a sub-task of Image Restoration, cutting-edge Single Image Super-Resolution (SISR) algorithms [2, 14, 32] follow the same data specific, even if they are relying on deep convolutional networks or Transformer [32, 58] architectures. One of the largest drawbacks characterizing these algorithms is represented by them being limited by the quality of the data used for optimization. Various image restoration applications are characterized by extreme difficulties in acquiring real domain paired data [1, 30], driving the need for realistic and relevant data synthesis [18, 53]. Accurately modeling application specific degradations in the sRGB representation proves extremely difficult, given the fact that the highly nonlinear ISP characteristic shifts also the physical characteristics of the degradation appearance. This represents the main factor limiting the performance of these algorithms in real applications deployment, with the gap observable at the data level being difficult to overcome with model-free algorithms.

Therefore, studying RAW data becomes a crucial step in driving the general image restoration performance improvement trend. Consequently, RAW Image Super-Resolution applications can benefit from the increased variance signal, with algorithms robust to fine architectural properties of the involved devices. Developing highly effective algorithms like the ones well-established in the sRGB domain [16, 17, 44, 48] can prove a step ahead in the development of high performance imagery applications, with cost effective devices.

Thus, in this work we are presenting the solutions submitted for the NTIRE 2024 RAW Image Super Resolution Challenge. We are providing information regarding the challenge setup, with the task description and the challenge data properties characterizing the challenge dataset splits. We are also listing information regarding the challenge participants, with their teams and affiliations.

In Sec. 2 we describe the challenge dataset and evaluation, and we discuss the overall results. In Sec. 3 we provide detailed descriptions of the best solutions.

Team	Method	Validation 1MP	Test 1MP	Test 12MP	# Params. (M)	Train E2E	Train Res.
Samsung	2-Stage w/ FPL	43.40 / 0.99	43.443 / 0.986	43.858 / 0.988	53.7	Yes	384
XiaomiMMAI	EffectiveSR	43.38 / 0.99	43.249 / 0.986	NA	20.9	No	64
USTCX	RBSFormer [26]	43.21 / 0.99	42.493 / 0.984	43.649 / 0.987	3.3	Yes	112
McMaster	SwinFSR Raw	42.48 / 0.98	42.366 / 0.984	NA	6.64	Yes	256
	BSRAW [18]	42.25 / 0.98	42.106 / 0.984	42.853 / 0.986	1.5	Yes	248
NUDT RSR	SAFMN FFT	41.81 / 0.98	41.621 / 0.982	NA	0.27	No	128 - 448
	Interpolation [18]	35.95 / 0.95	36.038 / 0.952	36.926 / 0.956			

Table 1. We provide **PSNR/SSIM** results on the validation set (40 images), the complete testing set (200 images), and the testing set at full-resolution (12MP) RAW images [19]. All the fidelity metrics are calculated in the RAW domain. "NA" indicates the results are not available for the method. We highlight two baseline methods. We also report the number of parameters of each method, if the method was trained end-to-end (Yes/No), and the image resolution used for training the models.

Related Computer Vision Challenges Our challenge is one of the NTIRE 2024 Workshop associated challenges on: dense and non-homogeneous dehazing [3], night photography rendering [4], blind compressed image enhancement [55], shadow removal [49], efficient super resolution [41], image super resolution ($\times$4) [13], light field image super-resolution [52], stereo image super-resolution [50], HR depth from images of specular and transparent surfaces [57], bracketing image restoration and enhancement [60], portrait quality assessment [7], quality assessment for AI-generated content [36], restore any image model (RAIM) in the wild [33], RAW image super-resolution [19], short-form UGC video quality assessment [31], low light enhancement [37].

2. NTIRE 2024 RAWSR Challenge

2.1. Dataset

The challenge dataset is based on BSRAW [18]. Following previous work [18, 53, 54], we use images from the Adobe MIT5K dataset [6], which includes images from multple Canon and Nikon DSLR cameras.

The DSLR images are manually filtered to ensure diversity and natural properties (*i.e.* remove extremely dark or overexposed images), we also remove the blurry images (*i.e.* we only consider all-in-focus images).

The **pre-processing** is as follows: (i) we normalize all RAW images depending on their black level and bit-depth. (ii) we convert ("pack") the images into the well-known RGGB Bayer pattern (4-channels), which allows to apply the transformations and degradations without damaging the original color pattern information [35].

Training: We provide the participants 1064 $1024\times1024\times4$ clean high-resolution (HR) RAW images. The LR degraded images can be generated on-line during training using the degradation pipeline proposed in BSRAW [18].

Such degradation pipeline considers different noise profiles, multiple blur kernels (PSFs) and a simple downsampling strategy to synthesize low-resolution (LR) RAW images. The participants can apply other augmentation techniques or expand the degradation pipeline to generate more realistic training data.

2.2. Baselines

We use BSRAW [18] as the main baseline. The top performing challenge solutions improve the baseline performance, however, the neural networks are notably more complex in terms of design and computation.

2.3. Results

We use three testing splits: (i) Validation, 40 1024px images using during the model development phase. (ii) Test 1MP, 200 images of 1024px resolution. (iii) The same 200 test images at full-resolution $\approx$ 12MP. The participants process the corresponding LR RAW images (*e.g.* $512 \times 512 \times 4$), and submit their results. Thus, the participants do not have access to the ground-truth images.

We provide samples of the testing set in Fig. 1.

In Tab. 1 we provide the challenge benchmark. Besides fidelity metrics such as PSNR and SSIM, we also provide relevant implementation details of each method. The methods can greatly improve the RAW images quality and resolution, even in the case of full-resolution 12MP images as output. We provide detailed visual comparisons in Fig. 8, Fig. 9 and Fig. 10. All the proposed methods are able to increase the resolution and details of the RAW images while reducing blurriness and noise. Moreover, there are not detectable color artifacts.

We can conclude that (synthetic) RAW image super-resolution can be solved similarity to RAW denoising. However, more realistic downsampling remains an open challenge.

Acknowledgements This work was partially supported by the Humboldt Foundation. We thank the NTIRE 2024 sponsors: Meta Reality Labs, OPPO, KuaiShou, Huawei and University of Würzburg (Computer Vision Lab).

https://cvlai.net/ntire/2024/

3. Challenge Methods and Teams

3.1. Dual Stage RAW SR with Focal Pixel Loss

Team Samsung MX,SRC-B

Jianxing Zhang [1], *Jia Li* [1], *Fan Wang* [1], *Xiaopeng Li* [1], *Zikun Liu* [1], *Hyunhee Park* [2], *Sejun Song* [2], *Changho Kim* [2]

[1] *Samsung Research China - Beijing (SRC-B)*
[2] *Samsung MX(Mobile eXperience) Business*

Team Samsung MX,SRC-B is introducing a two-stage network for RAW Image Super Resolution. The solution is using the divide-and-conquer strategy, with the first stage tasked with recovering the image structure from the low resolution degraded RAW image, and the second stage aiming at recovering the maximum amount of details, offering a refined reconstruction. Moreover, the team is further extending existing methods for synthetic data generation, studying further into hardware specific RAW image degradations, proposing new definitions for the relevant device-specific noise profiles and new blur kernels aligning with typical real-world scenarios. They propose a randomized degradation model, simulating different interactions between the observed simulated defects.

Finally, Team Samsung MX,SRC-B is proposing a novel Focal Pixel Loss, which was proven through the performance improvements during the model fine-tuning stage.

As shown in Fig. 2, the network structure mainly includes two stages. The first stage mainly draws on Restormer [58], whose main role is to restore the main content of raw images. The second stage mainly uses a NAFNet [10] based design, whose main role is to restore more details on the basis of the first stage of recovery.

The training procedure accounts for the dual stage design of the proposed method. The first stage of the training procedure follows the optimization of the parameters corresponding to the first stage of the model. In the second stage of the training procedure, the optimized parameters are frozen, with the parameters of the second stage starting being refined, for optimum specialization. The final estimation is then performed using both sets of optimized parameters (Restormer and NAFNet models).

one of the main details of the solution is the proposal of a novel training objective, based on the characteristics of the input data. The proposition of the Focal Loss (see Eq. (1)) is a solution for the observed imbalance in terms of highly affected pixels ratio, given the non-unifor effect of the signal degradation function . Therefore, the Focal Pixel Loss (FPL) is introducing exponential penalties to those pixels characterized by a large signal shift.

$$FPL(\hat{I}, I) = -D(\hat{I}, I)^{\gamma} \log_{10} D(\hat{I}, I) \qquad (1)$$

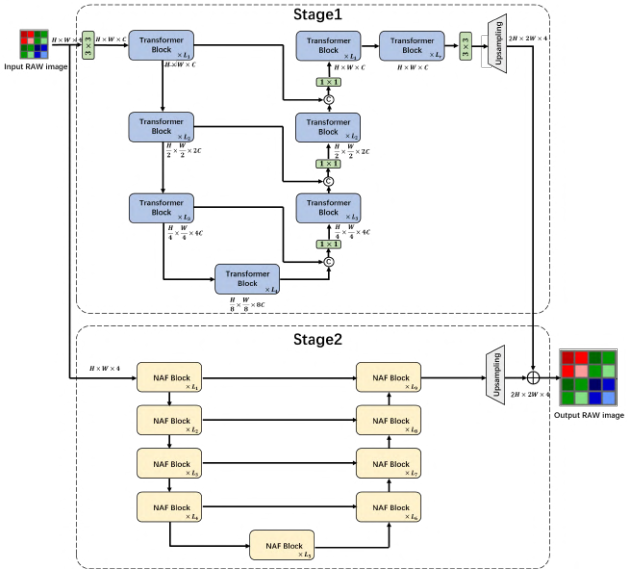

Figure 2. Dual-Stage RAWSR Framework proposed by Team Samsung MX,SRC. Best viewed in electronic version.

In Eq. (1), $D(.,.)$ is a standard L-norm distance between the restored image $\hat{I}$ and the reference image I, and γ is an adjustable factor, controlling the strength of the penalty.

Implementation details The model is trained solely on the data provided by the challenge organizers. It only contains more than 1000 RAW images from various DSLR camera sensors. The dataset was augmented using standard image augmentation techniques and simulations of the image degradation pipeline. The training procedure is a dual stage operation, optimizing the model stages sequentially. The optimization technique used is the AdamW [38] optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.999$, weight decay 0.0001) with the cosine annealing strategy, where the learning rate gradually decreases from the initial learning rate 5×10^{-5} to 1×10^{-7} for 5×10^5 iterations. The model goes through a pre-train optimization phase, base don the L_1 loss. During the fune-tune phase, the objective is set to the Focal Pixel Loss, with the initial learning rate being set to 5×10^{-6}. The training batch size is set to 4 and patch size is 384. Horizontal/vertical flipping and rotation are used for data augmentation. All experiments are conducted on A100 GPUs.

3.2. EffectiveSR for RAW Images

Team XiaomiMMAI

Zhijuan Huang[1], *Hongyuan Yu*[1], *Cheng Wan*[2], *Wending Xiang, Jiamin Lin*[1], *Hang Zhong*[1], *Qiaosong Zhang*[1], *Yue Sun*[1], *Xuanwu Yin*[1], *Kunlong Zuo*[1]

[1]*Xiaomi Inc.* [2]*Georgia Institute of Technology*

The solution proposed by Team XiaomiMMAI is a dual branch network based on HAT [12], adopting re-parameterization [20] during training, using the additional parameters to fully exploit the potential of the method. They are introducing, a task-by-task and step-by-step training method for RAW Image Super-Resolution to simultaneously address three tasks: denoising, deblurring, and 2× Super Resolution.

To address the limitation given by the low number of samples offered for training, Team XiaomiMMAI converts the RAW images into an RGB images and performs combined data enhancement on the set of produced RGB image, using random rotations, flips, color changes, brightness changes, random blur, etc.. Then, the set of enhanced RGB images are translated back to the RGGB RAW domain, with the processed data used to train [5] the proposed solution.

The model proposed by Team XiaomiMMAI is inspired by HAT [11], with the architecture being optimized for the RAW Image Super Resolution task. The optimized dual-branch network structure (DB-HAT) is shown in Fig. 3. The introduced Step-by-step and task-by-task training method for RAWISR further enhances the performance level achieved by their solution.

Step-by-step: To accelerate training and achieve good performance, Team XiaomiMMAI adopted a strategy where each sub-task, including the final joint optimization, is trained based on a pyramid image representation. Initially, the model is trained on small scale images (64×64), gradually increasing the resolution of the image patches to 128×128 and 256×256.

Task-by-task: Team XiaomiMMAI divided RAWISR into three sub-tasks: denoising, deblurring, and 2x SR. Initially, they start by training for RAW image denoising, followed by the connected tasks of deblurring and 2× Super Resolution. Finally, a joint optimization procedure is applied for entire network to produce the final estimator.

In the process of training denosing and deblurring, Team XiaomiMMAI used the RepConv re-parameterization technique on the final stage of the proposed DB-HAT, to improve the visual image quality of the task. The reparameterizable convolution block (RepConv) is shown in Fig 4.

Implementation details The dataset used for three sub-task training consists of 1000+ RAWs, and the data augmentation methods can refer to the previous section for details. In the final stage of joint optimization of the entire network, Team XiaomiMMAI used the provided 1000+ dataset instead of the augmented dataset. The learning rate is initialized at 4×10^{-4} and decays according to the CosineAnnealing strategy during the training of three sub-tasks. The network undergoes training for a total of $2x10^5$ iterations, with the L2 loss function being minimized as the trainign objective of the Adam optimizer.

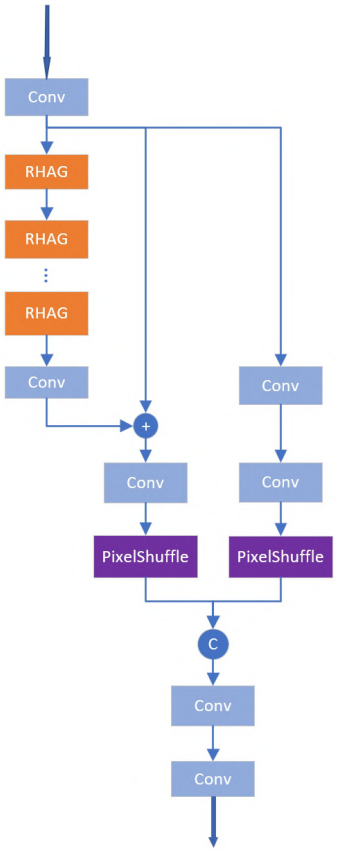

Figure 3. DB-HAT model proposed by Team XiaomiMMA

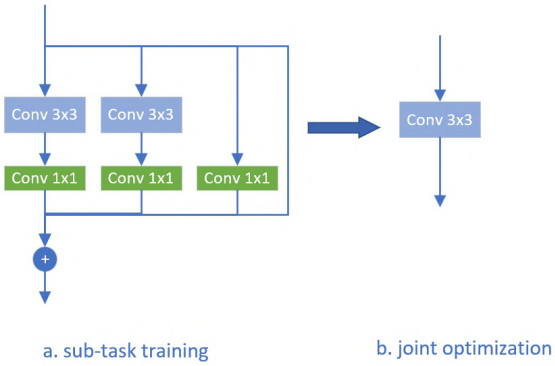

Figure 4. RepConv used by Team XiaomiMMA

Subsequently, finetuning is executed for two iterations, using the L2 loss and SSIM loss functions, with an initial learning rate of 5×10^{-5} for $2x10^5$ iterations. All experiments are conducted with the PyTorch 2.0 framework on 8 A100 GPUs.

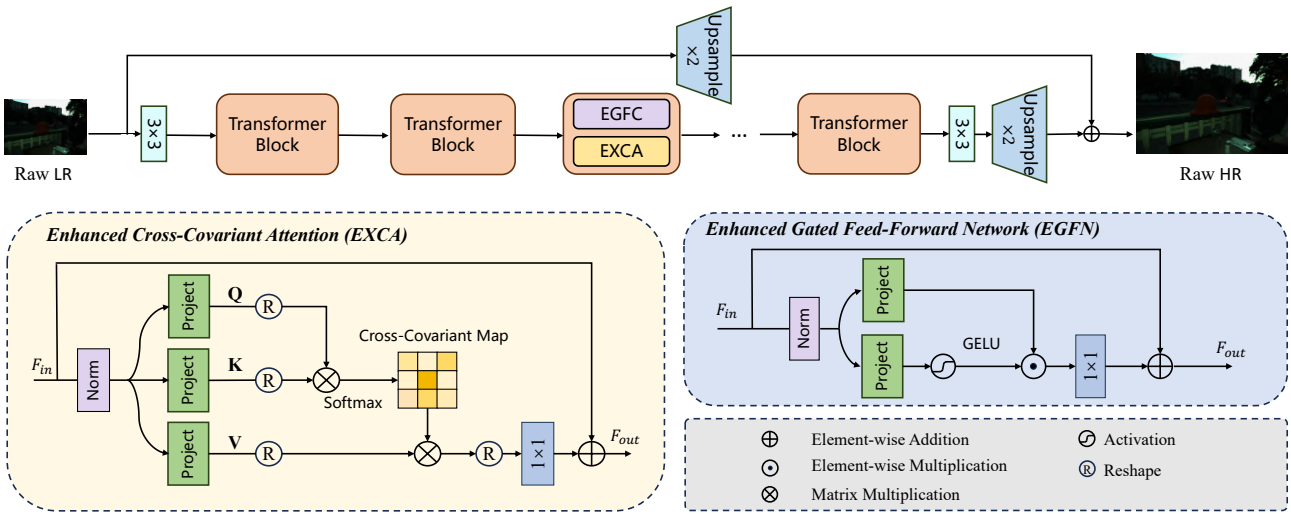

Figure 5. The RBSFormer [26] Framework proposed by Team USTC604.

3.3. RBSFormer: Enhanced Transformer Network for Raw Image Super-Resolution

Team USTC604

Senyan Xu, Siyuan Jiang, Zhijing Sun, Jiaying Zhu

University of Science and Technology of China

Contact: a804235820@gmail.com

Team USTC604 proposed a transformer framework for raw image super-resolution, with a design based on the transformer block proposed in Restormer[58] (see Fig. 5), solution that excels in capturing long-range pixel interactions by applying self-attention across channels.

The solution used the data provided in the NTIRE 2024 RAW Image Super Resolution challenges, with the degradation pipeline described in [18].

For a 4-channel RGGB RAW image patch of size 224×224, the computational cost of the model proposed by Team USTC604 amounts to 14.6 GFLOPS, being characterized by a number of 3.31 trainable parameters. On a consumer-grade gaming GPU, the NVIDIA RTX4090Ti, the forward pass needed of a full resolution image estimation needs 650 ms, following the limitations of the used backbone, as one of the computationally expensive solutions proposed in the image restoration field.

The software characteristic to the performed experiments is based on the PyTorch 1.8 framework, with the experiments being performed on NVIDIA RTX4090Ti devices. The trainign procedure is based on the Adam optimizer with the decay parameters parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate is 3×10^{-4} and changes with Cosine Annealing scheme to 1×10^{-7}, with the training procedure covering 120K iterations in a time-frame of around 10 hours. We refer the reader to the author's paper RBSFormer [26] for more details.

3.4. SwinFSR Raw Image Super Resolution

Team McMaster

Liangyan Li, Ke Chen, Yunzhe Li, Yimo Ning, Guanhua Zhao, Jun Chen

McMaster University

Contact: lil61@mcmaster.ca

Team McMaster proposed an algorithm that considers multiple acquisition sensors, accounting for various image signal degradations induced by hardware limitations. The model is trained and learned directly from the 4-channel RAW data with an enhanced degradation pipeline. With a broader variance of noise during the degradation process, the solution demonstrates increased robustness, efficiently producing high-quality images from degraded inputs, thereby enhancing overall performance on the official datasets. Their approach works directly with 4-channel RGGB RAW images after a designed degradation process. The architecture is a hybrid model which integrates SwinFSR [8] with simple CNN layers. The model was trained and validated only on the official datasets [19].

The team utilized the noise model, blur kernel, and degradation model as demonstrated in BSRAW [18].

Their approach of adding noise is inspired by the strategy deployed in DiT [42] that gradually adds Gaussian noise

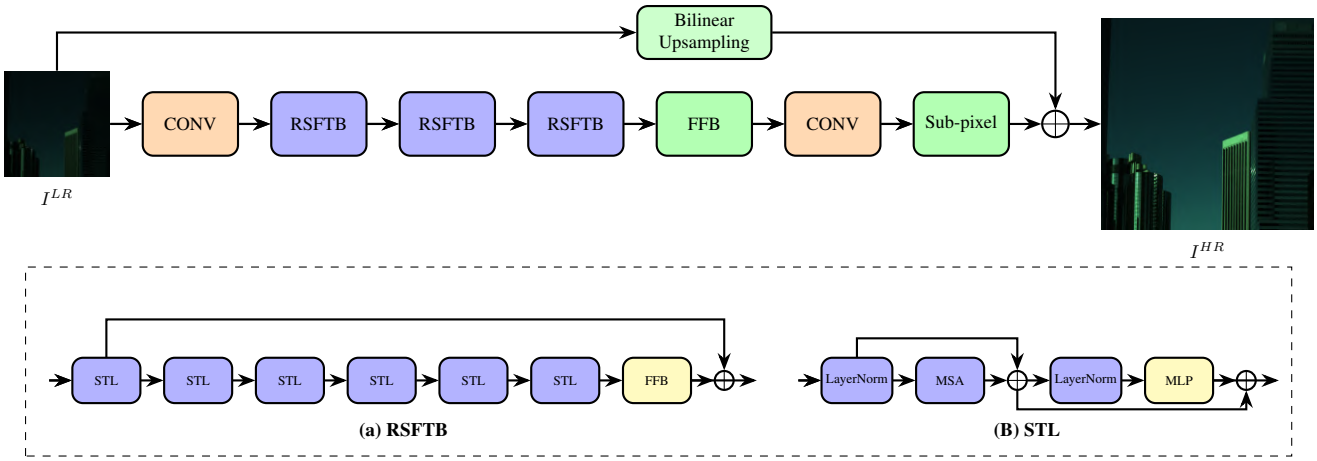

Figure 6. Main branch of the framework proposed by Team McMaster. (a) Residual Swin Fourier Transformer Block (RSFTB) (b) Swin Transformer Layer (STL).

as a form of degradation in the forward process. Diffusion models[27, 42] have been proposed for various imaging tasks, including image super-resolution, demonstrating their capability for end-to-end training to transform pure Gaussian noise into meaningful data representations. They investigate the applicability of gradual magnitude Gaussian noise, as utilized in diffusion models, for addressing the degradation process inherent in Raw Image Super-Resolution tasks. Team McMaster adopted the additive noise model from [42], exposing the input RAW images to noise using the forward diffusion definition described in [42]. In the forward diffusion process, the input data undergoes a degradation process the iterative addition of Gaussian noise across 1000 discrete steps. In the backward denoising process of DiT, the model is optimized for the forward diffusion process, maximizing the data likelihood via the variational lower bound.

The architectural configuration of the Team McMaster proposed model is depicted in Fig. 6. In the proposed method, a SwinFSR-based design [9] performs RAW domain image feature extraction, combined with feature upsampling, matching the size of high-resolution images via a complex convolution operator. SwinFSR builds on the success of SwinIR [32], with an additional data modality given by the Frequency Domain Knowledge, through the FFT image representation [8]. This proves to be a superior strategy, combining spatial and spectral features as a way to balance the local information of the spatial domain, and the global information accessed through the spectral representation. It introduces a novel cross-attention module for efficient information exchange between the two modalities and adapts to rectangular input patches for flexibility.

For the proposed model, only the feature extraction branch of SwinFSR is deployed.

Implementation details The solution was optimized solely on the NTIRE 2024 official challenge data [19], using the proposed Development Phase submission set for validation. This dataset contains 1064 4-channel DSLR-specific RGGB RAW images for training, and an additional 40 raw images set for validation. The images were pre-processed, applying white-black level correction, then being normalized to the unit interval. Since the degraded low-resolution raw images are characterized by low quality, with a considerable level of details being lost, a data augmentation technique is applied, to improve the training procedure in terms of stability, convergence, and the achieved performance level. The strategy combines simple horizontal or vertical flips with channel shifts and mixup augmentations. The training objective is based on the L_1 loss.

3.5. Spatially-Adaptive Feature Modulation for RAW Super-Resolution

Team NUDT RSR

Jinyang Yu, Kele Xu, Qisheng Xu, Yong Dou

National University of Defense Technology, Computer Dept., Changsha, China

Contact: a804235820@gmail.com

The solution proposed by Team NUDT RSR addresses three major components of the proposed challenge, extending on the degradation pipeline, model design, and model supervision, achieving a significant performance level in terms of restoration fidelity.

For the image signal degradation pipeline, the considered degradations include diverse blur kernels, exposure defects, image downsampling, finally coupled with a noise model characteristic to real-world raw image data.

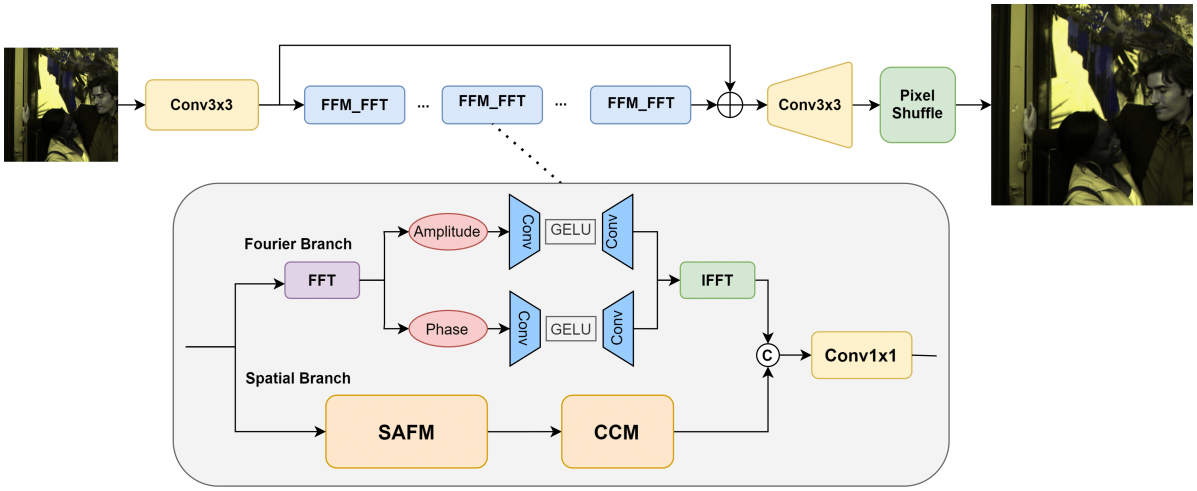

Figure 7. The overall network architecture proposed by Team NUDT RSR. The proposed FFM_FFT block is a two-branched dolution, with the Fourier branch extracting the amplitude and phase, guiding the global-local feature mixing performed in the spatial branch, through SAFM and CCM blocks[47].

Following [29], the proposed solution is benefitting from the amplitude and phase components, added to the SAFMN [47] backbone. the model is fusing frequency domain and spatial domain information, for global-local level feature mixing.

Moreover, the knowledge distillation is deployed to the described solution, using a NAFNet [10] teacher, with multiple complexity level feature supervision. Finally, we apply a progressive training strategy, gradually increasing the patch size at each stage to accommodate larger test inputs.

Degradation pipeline: Inspired by [51] and [18], in order to enable the model to learn real degradation information, the Bayer pattern RAW images are cropped, following by degrading the RAW signal in a sequence of operations composed of multiple blurring operations, exposure compensations, downsampling, and hardware specific added noise.

The first step is represented by a blurring operator, based on randomly generated Gaussian blur, generalized Gaussian blur, with a plateau-shaped distribution, and their anisotropic version. The augmented PSF kernels provided by starter kit are also considered. All kernel sizes are ranging from 7×7 to 25×25.

Then, the pipeline continues with linear adjustment for image exposure. As [18] discussed, to simulate the artifacts caused by underexposure and overexposure, the pipeline implements exposure adjustment by linearly scaling the image. The adjustment factor is tuned to the [-0.25, 0.25] interval, applied in the unit interval normalized images.

Next, the image suffers further downsampling, considering different downsampling kernels, including bicubic interpolation, bilinear interpolation, and an average-pooling operator. To build multi-scale training pairs, the input image is either upsampled or downsampled to a random size first, and rescaled back to half of the original size as for $2\times$ super-resolution tasks.

On the downscaled image, heteroscedastic Gaussian noise [18] is then applied, followed by the practical shot-read noise [18] for different exposure levels. An image with higher exposure factor in step 2 is more likely to get noised by heteroscedastic noise, and the shot-read noise for low-light images.

The last step of the degradation pipeline is a second blurring operator. To expand the degradation space like the high-order degradation model [51], a random operation based on a set of second blurring kernels (same kernels considered in the first step), characterized by smaller standard deviations, is applied in the final stage.

Network According to [29], the Fourier spectrogram of an image has similar amplitude to its downsampled one's, while the phase is related to the noise observed in the acquired image signal. Although it is designed for low light image enhancement tasks, Team NUDT RS started with the observation that blind RAW Image super-resolution can also benefit from refining these two components of an low resolution image.

Thus, the Team NUDT RS proposed model is represented in Fig. 7, where the input image is encoded by a 3×3 convolution layer for shallow feature extraction, and the FFM_FFT blocks are used for deep feature extraction. Following [29], the main blocks are divided into spatial branches and Fourier branches. In each block, the input is simultaneously sent to both branches, then the processed

Method	Params. [M]	FLOPS [G]	PSNR [dB]
NAFNet (Large)	116	255	41.76
NAFNet (Small)	0.290	14.08	40.78
SAFMN	0.229	59.06	41.20
SAFMN_FFT	0.272	67.29	41.81

Table 2. A comparison between the NAFNet, vanilla SAFMN and the method proposed by Team NUDT RS. The large version of NAFNet uses [2, 2, 4, 8] encoding blocks, [2, 2, 2, 2] decoding blocks for each stage, and 12 middle blocks, the width is set to 64. While the smaller version use width 32, with both [2, 1] configuration for encoding and decoding blocks and 1 middle block. The vanilla SAFMN has dim=36, ffn_scale=2 configuration, and 8 main blocks, which are the same to teh poposed SAFMN_FFT. All flops are calculated with input size $1 \times 4 \times 512 \times 512$.

features of the branches are fused. After a residual connection, a final pixel-reshuffle operator is used to upscale the feqature set to the resolution of the reference image [45].

In the spatial branch, the cross domain communication is performed through efficient Feature Mixing Module (FMM) blocks of SAFMN [47]. A FMM block consists of a SAFM block and a CCM block. SAFM splits the channels into different parts, fuses their features at different scale levels and obtain attention map after GELU activation. Then, the original input multiplies the attention map. The CCM block consists of a 3×3 convolution layer and a 1×1 convolution layer, which works as a channel mixer to capture local context information.

In the Fourier branch, an image is transformed to frequency map by Fast Fourier Transform (FFT) operator to get amplitude and phase component, which are later processed by two 1×1 convolution layers with GELU activations respectively. Next, these refined components are combined to a new frequency map, and the inverse FFT its used to transfer back to the spatial domain. Features from different branches are concatenated and fused by another 1×1 convolution. The proposed model uses 8 such blocks with width 36. To avoid expensive computations on high dimensions, the SFT layers in the final reconstruction stage are removed [29]. This results in an efficient estimator, with the total parameter count being 272.068 K for the proposed solution. The team presents an ablation study in Tab. 2.

Training strategy A two-stage optimizing strategy is applied. All the training is based on the development dataset and there are no external datasets used.

Firstly, a large version of NAFNet [10] model is trained as a teacher network, with increased complexity degradations on patch size 128. Then this NAFNET model is used to apply knowledge distillation, as part of the optimization technique used for the proposed solution. To reduce the performed computations, the knowledge does not rely on distances computed between multi-level feature sets, but rather on statistics defining the feature see [21].

Secondly, to accommodate the high resolution test images, the student model is progressively finetuned on increased resolution patches. The training patch size increases from 128, 256, 352 to 448.

The optimization objective is based on the L! distance simultaneously applied in the spatial and Fourier domains [46]. The total loss is defined in Eq. (5).

$$\mathcal{L}_p = \|I_1 - I_2\|_1 \tag{2}$$

$$\mathcal{L}_f = \|\mathcal{F}(I_1) - \mathcal{F}(I_2)\|_1 \tag{3}$$

$$\mathcal{L}_{kd} = \frac{1}{|D|} \sum_{i \in D} \|\mathcal{G}(\mathcal{N}_i(I_1)) - \mathcal{G}(\mathcal{N}_i(I_2))\|_c \tag{4}$$

$$\mathcal{L}_{total} = \mathcal{L}_p + \lambda \mathcal{L}_f + \mu \mathcal{L}_{kd} \tag{5}$$

I_1, I_2 denotes the restored image prediction and the corresponding reference image. $\mathcal{F}(\cdot)$ is the FFT operator, $\mathcal{N}_i(\cdot)$ is the feature extracting operator determined by the network, where $i \in D$ is a set of the intermediate layers and $\mathcal{G}(\cdot)$ is defined to the square of the channel-wise mean of intermediate features. The $\|\cdot\|_c$ represents Charbonnier loss. λ and μ are used to control the weights of different components. The finetuning stage only uses the $\mathcal{L}_p$ and $\mathcal{L}_f$ loss terms.

Implementation details The experiments are based on the Pytorch framework for implementation. The NAFNet teacher model uses a width 64, with 2, 2, 4, 8 encoder blocks and 2, 2, 2, 2 decoder blocks at each stage. The middle block num is set to 12. The teacher module is trained for 600000 iterations. The learning rate is set to 2e-4, and halves for every 100000 steps after 300000 steps. The training procedure applied for the teacher network uses only the $\mathcal{L}_p$ loss. For knowledge distillation, the procedure starts with an initial learning rate warm-up stage, lasting for 30000 iterations. Then, a cosine annealing stage is applied, decreasing the learning rate to 1e-6.

The FFT L1 loss is balanced with a weight $\lambda = 0.05$, and the distillation loss with $\mu = 0.1$. For the progressive finetuning, each training stage for the proposed model lasts for 200,000 iterations with learning rate 5e-5 that is reduced to 1e-6 with a cosine scheduler. Alongside the augmentation involved in the degradation pipeline, some traditional data augmentation methods, such as random cropping, flipping, and rotation are deployes to further diversify the training dataset. The testing phase deploys the geometric self-ensemble strategy [34], which averages 8 outputs corresponding to 8 augmented versions of the input image. All the experiments are conducted on a NVIDIA GeForce RTX 2080Ti GPU, using the AdamW optimizer with $\beta_1 = 0.99$, $\beta_2 = 0.9$.

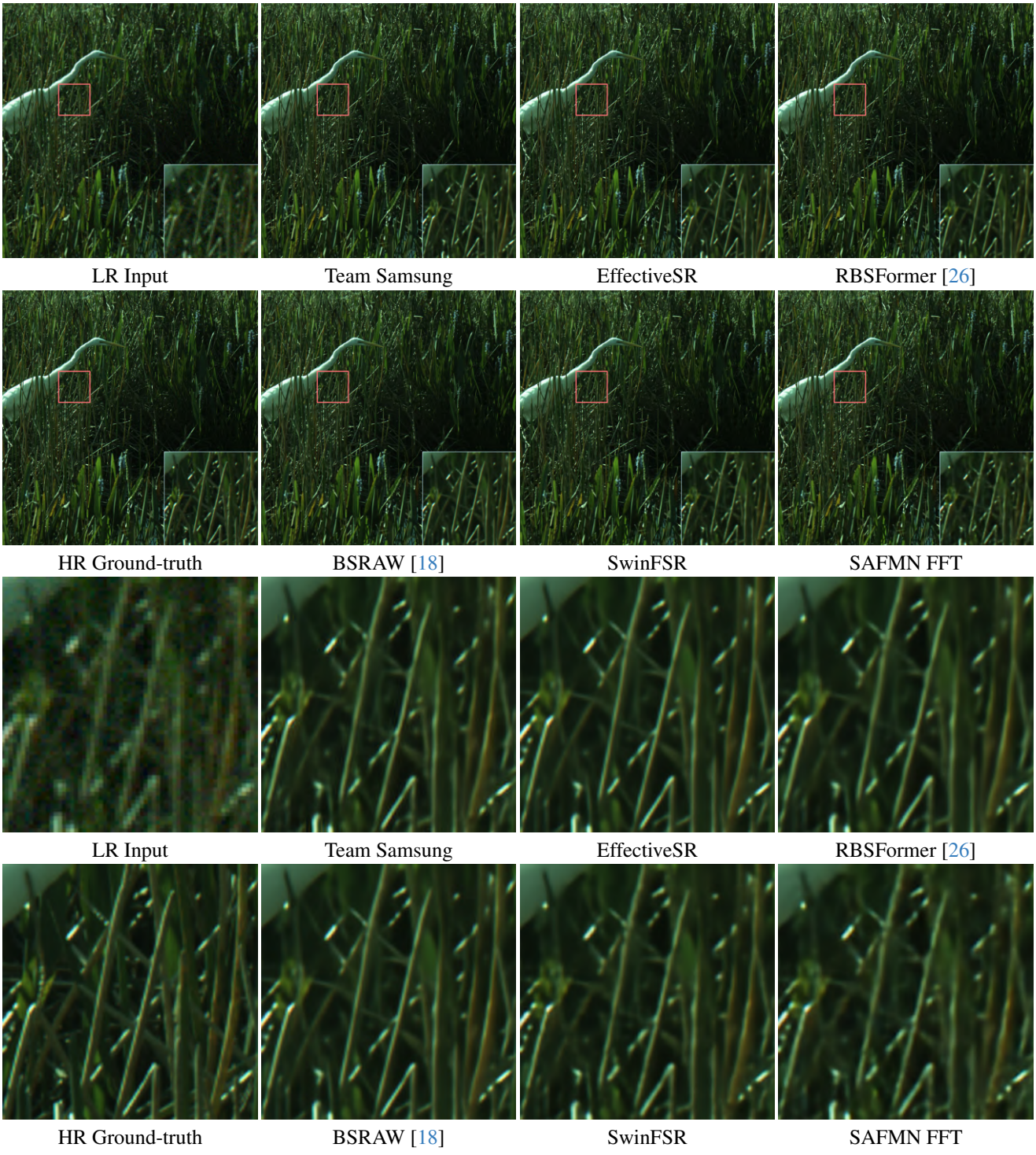

Figure 8. Visual comparison using the **NTIRE 2024 RAW Image Super-Resolution Challenge** testing set (`190.npz`). The HR resolution RAW images have 1024×1024 resolution and 4-channels (RGGB Bayer pattern). RAW images are visualized using bilinear demosaicing, gamma correction and tone mapping.

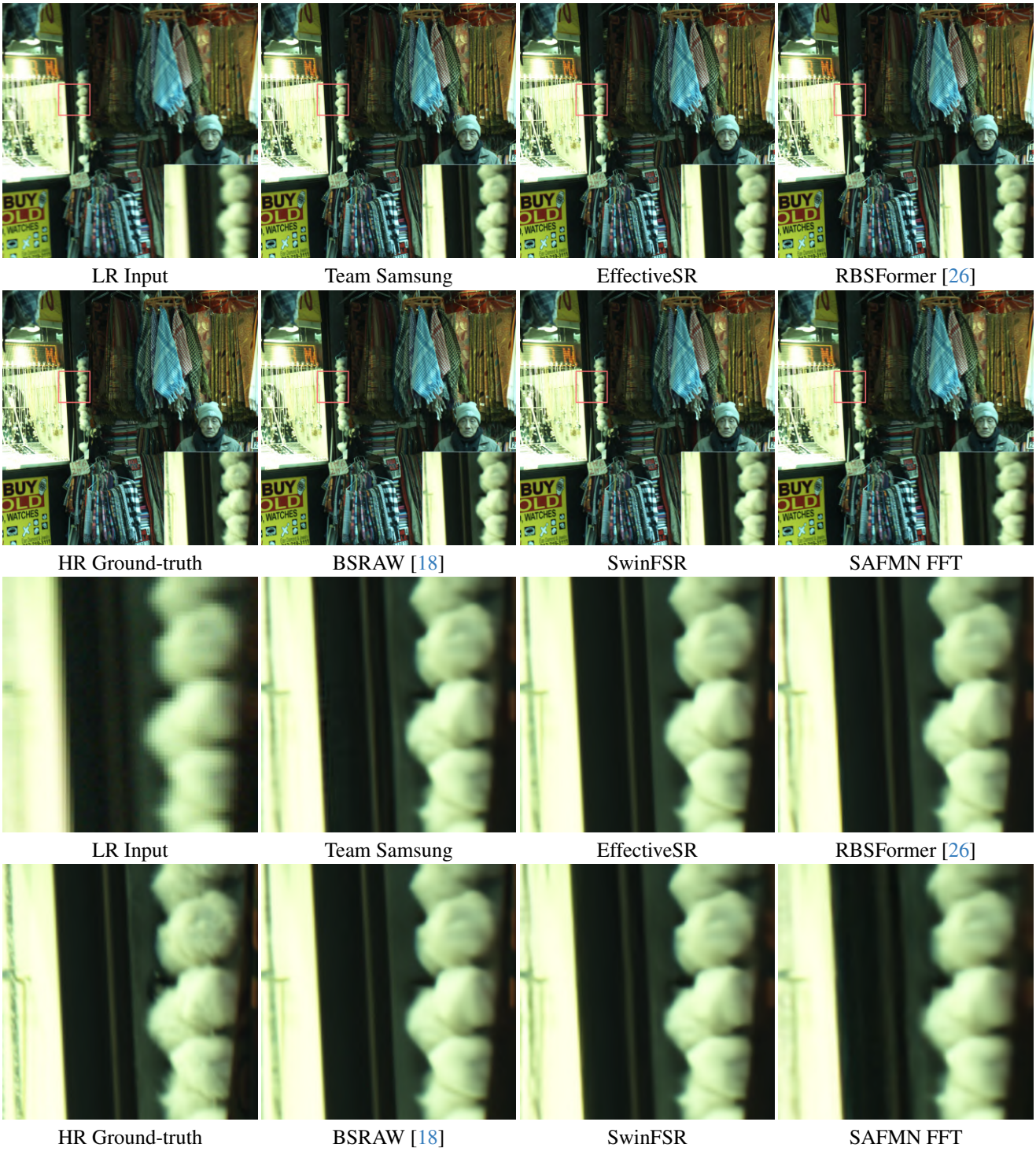

Figure 9. Visual comparison using the **NTIRE 2024 RAW Image Super-Resolution Challenge** testing set (`155.npz`). The HR resolution RAW images have 1024×1024 resolution and 4-channels (RGGB Bayer pattern). RAW images are visualized using bilinear demosaicing, gamma correction and tone mapping.

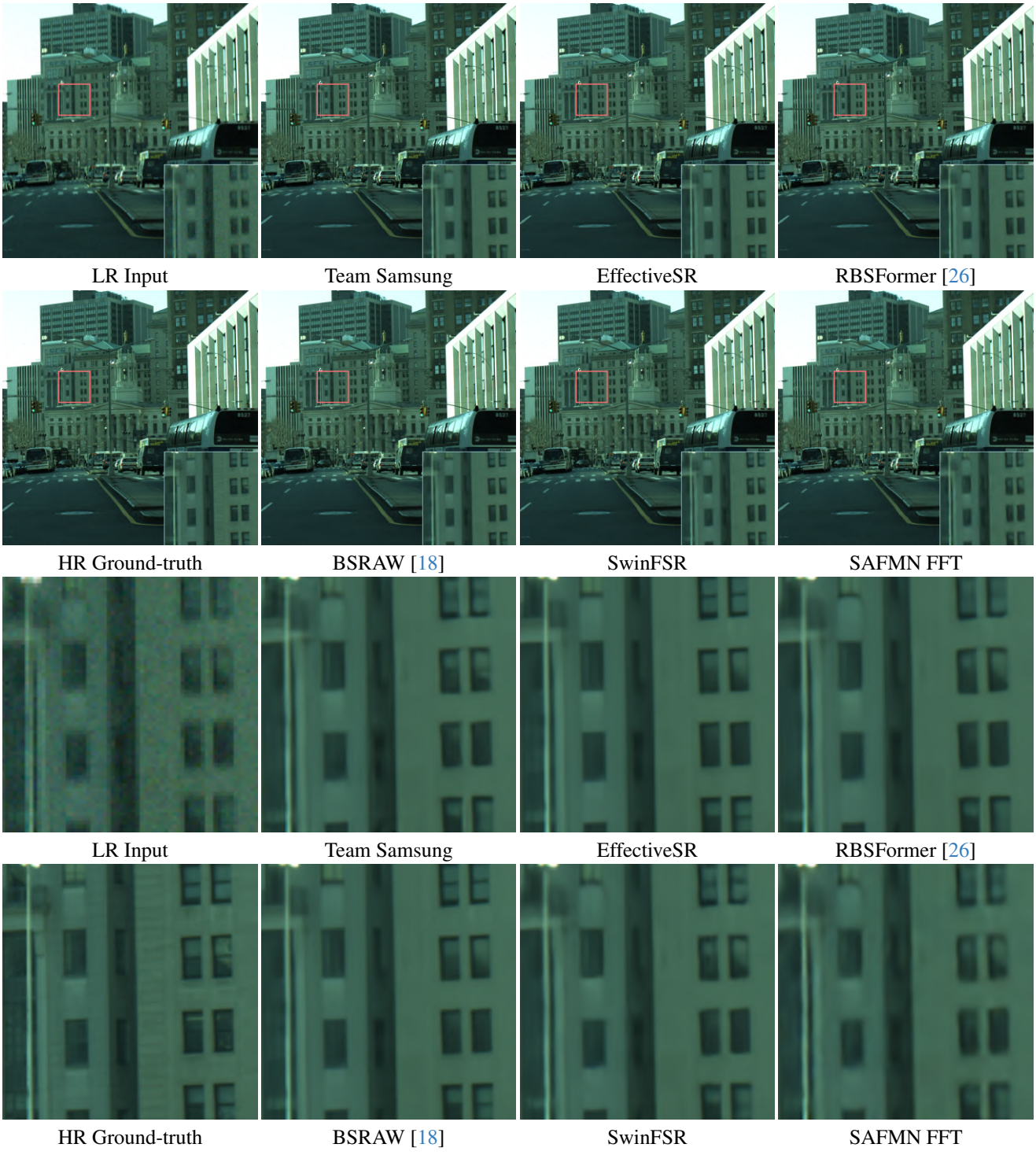

Figure 10. Visual comparison using the **NTIRE 2024 RAW Image Super-Resolution Challenge** testing set (`177.npz`). The HR resolution RAW images have 1024×1024 resolution and 4-channels (RGGB Bayer pattern). RAW images are visualized using bilinear demosaicing, gamma correction and tone mapping.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, pages 1692–1700, 2018. 2

[2] Eirikur Agustsson and Radu 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*, pages 126–135, 2017. 2

[3] Cosmin Ancuti, Codruta O Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2024 dense and non-homogeneous dehazing challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[4] Nikola Banić, Egor Ershov, Artyom Panshin, Oleg Karasev, Sergey Korchagin, Shepelev Lev, Alexandr Startsev, Daniil Vladimirov, Ekaterina Zaychenkova, Dmitrii R Iarchuk, Maria Efimova, Radu Timofte, Arseniy Terekhin, et al. NTIRE 2024 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[5] 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, 2, 5

[6] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *CVPR*, 2011. 3

[7] Nicolas Chahine, Marcos V. Conde, Gabriel Pacianotto, Daniela Carfora, Benoit Pochon, Sira Ferradans, Radu Timofte, et al. Deep portrait quality assessment. A NTIRE 2024 challenge survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 3

[8] Ke Chen, Liangyan Li, Huan Liu, Yunzhe Li, Congling Tang, and Jun Chen. Swinfsr: Stereo image super-resolution using swinir and frequency domain knowledge, 2023. 6, 7

[9] Ke Chen, Liangyan Li, Huan Liu, Yunzhe Li, Congling Tang, and Jun Chen. Swinfsr: Stereo image super-resolution using swinir and frequency domain knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1764–1774, 2023. 7

[10] 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. 2, 4, 8, 9

[11] Xiangyu Chen, Xintao Wang, Wenlong Zhang, Xiangtao Kong, Yu Qiao, Jiantao Zhou, and Chao Dong. Hat: Hybrid attention transformer for image restoration. *arXiv preprint arXiv:2309.05239*, 2023. 5

[12] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22367–22377, 2023. 5

[13] Zheng Chen, Zongwei WU, Eduard Sebastian Zamfir, Kai Zhang, Yulun Zhang, Radu Timofte, Xiaokang Yang, et al. NTIRE 2024 challenge on image super-resolution (×4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[14] Marcos V Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2SR: Swinv2 transformer for compressed image super-resolution and restoration. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2022. 2

[15] Marcos V Conde, Steven McDonagh, Matteo Maggioni, Ales Leonardis, and Eduardo Pérez-Pellitero. Model-based image signal processors via learnable dictionaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 481–489, 2022. 1

[16] Marcos V. Conde, Florin Vasluianu, Sabari Nathan, and Radu Timofte. Real-time under-display cameras image restoration and hdr on mobile devices. In *Computer Vision – ECCV 2022 Workshops*, pages 747–762, Cham, 2023. Springer Nature Switzerland. 2

[17] Marcos V Conde, Florin Vasluianu, Javier Vazquez-Corral, and Radu Timofte. Perceptual image enhancement for smartphone real-time applications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1848–1858, 2023. 1, 2

[18] 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. 1, 2, 3, 6, 8, 10, 11, 12

[19] 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. 3, 6, 7

[20] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742, 2021. 5

[21] Qinquan Gao, Yan Zhao, Gen Li, and Tong Tong. Image super-resolution using knowledge distillation. In *Computer Vision – ACCV 2018*, pages 527–541, Cham, 2019. Springer International Publishing. 9

[22] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 35(6), 2016. 2

[23] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics (ToG)*, 33(6):1–13, 2014. 1

[24] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 536–537, 2020. 1

[25] Andrey Ignatov, Cheng-Ming Chiang, Hsien-Kai Kuo, Anastasia Sycheva, and Radu Timofte. Learned smartphone isp on mobile npus with deep learning, mobile ai 2021 challenge: Report. In *CVPR Workshops*, pages 2503–2514, 2021. 1

[26] Siyuan Jiang, Senyan Xu, and Xingfu Wang. Rbsformer: Enhanced transformer network for raw image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 3, 6, 10, 11, 12

[27] Pieter Abbeel Jonathan Ho, Ajay Jain. Denoising diffusion probabilistic models. *arXiv:2006.11239*, 2020. 7

[28] Hakki Can Karaimer and Michael S Brown. A software platform for manipulating the camera imaging pipeline. In *ECCV*, pages 429–444, 2016. 1

[29] Chongyi Li, Chun-Le Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement, 2023. 8, 9

[30] Siyuan Li, Iago Breno Araujo, Wenqi Ren, Zhangyang Wang, Eric K Tokuda, Roberto Hirata Junior, Roberto Cesar-Junior, Jiawan Zhang, Xiaojie Guo, and Xiaochun Cao. Single image deraining: A comprehensive benchmark analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3838–3847, 2019. 2

[31] Xin Li, Kun Yuan, Yajing Pei, Yiting Lu, Ming Sun, Chao Zhou, Zhibo Chen, Radu Timofte, et al. NTIRE 2024 challenge on short-form UGC video quality assessment: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[32] 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, 7

[33] Jie Liang, Qiaosi Yi, Shuaizheng Liu, Lingchen Sun, Rongyuan Wu, Xindong Zhang, Hui Zeng, Radu Timofte, Lei Zhang, et al. NTIRE 2024 restore any image model (RAIM) in the wild challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[34] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017. 9

[35] Jiaming Liu, Chi-Hao Wu, Yuzhi Wang, Qin Xu, Yuqian Zhou, Haibin Huang, Chuan Wang, Shaofan Cai, Yifan Ding, Haoqiang Fan, et al. Learning raw image denoising with bayer pattern unification and bayer preserving augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[36] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, Xiele Wu, Radu Timofte, et al. NTIRE 2024 quality assessment of AI-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[37] Xiaoning Liu, Zongwei WU, Ao Li, Florin-Alexandru Vasluianu, Yulun Zhang, Shuhang Gu, Le Zhang, Ce Zhu, Radu Timofte, et al. NTIRE 2024 challenge on low light image enhancement: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[38] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 4

[39] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2502–2510, 2018. 2

[40] Guocheng Qian, Yuanhao Wang, Chao Dong, Jimmy S Ren, Wolfgang Heidrich, Bernard Ghanem, and Jinjin Gu. Rethinking the pipeline of demosaicing, denoising and super-resolution. *arXiv preprint arXiv:1905.02538*, 2019. 2

[41] Bin Ren, Yawei Li, Nancy Mehta, Radu Timofte, et al. The ninth NTIRE 2024 efficient super-resolution challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[42] William Peebles Saining Xie. Scalable diffusion models with transformers. *arXiv:2212.09748*, 2022. 6, 7

[43] Eli Schwartz, Raja Giryes, and Alex M Bronstein. Deepisp: Toward learning an end-to-end image processing pipeline. *IEEE Transactions on Image Processing*, 28(2):912–923, 2018. 1

[44] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016. 2

[45] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016. 9

[46] Long Sun, Jinshan Pan, and Jinhui Tang. Shufflemixer: An efficient convnet for image super-resolution. In *Advances in Neural Information Processing Systems*, pages 17314–17326. Curran Associates, Inc., 2022. 9

[47] Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Spatially-adaptive feature modulation for efficient image super-resolution. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13144–13153, 2023. 8, 9

[48] Florin Vasluianu and Radu Timofte. Efficient video enhancement transformer. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 4068–4072, 2022. 2

[49] Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei WU, Cailian Chen, Radu Timofte, et al. NTIRE

2024 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[50] Longguang Wang, Yulan Guo, Juncheng Li, Hongda Liu, Yang Zhao, Yingqian Wang, Zhi Jin, Shuhang Gu, Radu Timofte, et al. NTIRE 2024 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[51] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1905–1914, 2021. 8

[52] Yingqian Wang, Zhengyu Liang, Qianyu Chen, Longguang Wang, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2024 challenge on light field image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[53] 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. 1, 2, 3

[54] 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. 3

[55] Ren Yang, Radu Timofte, et al. NTIRE 2024 challenge on blind enhancement of compressed image: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[56] Huanjing Yue, Zhiming Zhang, and Jingyu Yang. Real-rawvsr: Real-world raw video super-resolution with a benchmark dataset. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 608–624. Springer, 2022. 2

[57] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, Alex Costanzino, Matteo Poggi, et al. NTIRE 2024 challenge on HR depth from images of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3

[58] 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, 4, 6

[59] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. Zoom to learn, learn to zoom. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2019. 1, 2

[60] Zhilu Zhang, Shuohao Zhang, Renlong Wu, Wangmeng Zuo, Radu Timofte, et al. NTIRE 2024 challenge on bracketing image restoration and enhancement: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 3