This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

NTIRE 2023 Challenge on Image Denoising: Methods and Results

Yawei Li* Yulun Zhang* Radu Timofte* Luc Van Gool* Zhijun Tu
Kunpeng Du Hailing Wang Hanting Chen Wei Li Xiaofei Wang Jie Hu
Yunhe Wang Xiangyu Kong Jinlong Wu Dafeng Zhang Jianxing Zhang
Shuai Liu Furui Bai Chaoyu Feng Hao Wang Yuqian Zhang Guangqi Shao
Xiaotao Wang Lei Lei Rongjian Xu Zhilu Zhang Yunjin Chen Dongwei Ren
Wangmeng Zuo Qi Wu Mingyan Han Shen Cheng Haipeng Li Ting Jiang
Chengzhi Jiang Xinpeng Li Jinting Luo Wenjie Lin Lei Yu Haoqiang Fan
Shuaicheng Liu Aditya Arora Syed Waqas Zamir Javier Vazquez-Corral
Konstantinos G. Derpanis Michael S. Brown Hao Li Zhihao Zhao Jinshan Pan
Jiangxin Dong Jinhui Tang Bo Yang Jingxiang Chen Chenghua Li Xi Zhang
Zhao Zhang Jiahuan Ren Zhicheng Ji Kang Miao Suiyi Zhao Huan Zheng
YanYan Wei Kangliang Liu Xiangcheng Du Sijie Liu Yingbin Zheng
Xingjiao Wu Cheng Jin Rajeev Irny Sriharsha Koundinya Vighnesh Kamath
Gaurav Khandelwal Sunder Ali Khowaja Jiseok Yoon Ik Hyun Lee Shijie Chen
Chengqiang Zhao Huabin Yang Zhongjian Zhang Junjia Huang Yanru Zhang

Abstract

This paper reviews the NTIRE 2023 challenge on image denoising ($\sigma = 50$) with a focus on the proposed solutions and results. The aim is to obtain a network design capable to produce high-quality results with the best performance measured by PSNR for image denoising. Independent additive white Gaussian noise (AWGN) is assumed and the noise level is 50. The challenge had 225 registered participants, and 16 teams made valid submissions. They gauge the state-of-the-art for image denoising.

1. Introduction

Image denoising is a classical image restoration problem that tries to recover a clean image from a noisy input image [10, 16, 52–54]. During the capturing and processing of images, there could be various noise types including Gaussian noise, Poisson noise, JPEG compression noise, and so on, making image denoising a very challenging task.

Currently, state-of-the-art image denoising methods are based on deep neural networks [26, 29, 50]. Thus, the aim of this challenge is to encourage the development of deep neural networks for image denoising. In this case, the traditional additive white Gaussian noise model can be used as a standard setting to fairly evaluate the performance of image denoising networks.

In collaboration with the 2023 New Trends in Image Restoration and Enhancement (NTIRE 2023) workshop, we organize the challenge on image denoising. The challenge's goal is to recover a clean image from a noisy input image that is corrupted by additive white Gaussian noise with noise level $\sigma = 50$. This challenge aims to discover advanced and innovative solutions for image denoising, benchmark their denoising performance, and identify general trends for the design of image denoising networks.

This challenge relates to previous challenges [1, 2, 20] and is one of the NTIRE 2023 Workshop series of challenges on: night photography rendering [40], HR depth from images of specular and transparent surfaces [49], image denoising (this challenge), video colorization [21], shadow removal [44], quality assessment of video enhancement [33], stereo super-resolution [45], light field image super-resolution [47], image super-resolution ($\times 4$) [57], 360° omnidirectional image and video super-resolution [6], lens-to-lens bokeh effect transformation [12], real-time 4K super-resolution [13], HR nonhomogenous dehazing [4], efficient super-resolution [28].

* Y. Li (yawei.li@vision.ee.ethz.ch, Computer Vision Lab, ETH Zurich), Y. Zhang, R. Timofte, and L. Van Gool were the challenge organizers, while the other authors participated in the challenge. Each team described their own method in the report. Appendix A contains the authors' teams and affiliations.
NTIRE 2023 webpage: https://cvlai.net/ntire/2023/.
Code: https://github.com/ofsoundof/NTIRE2023_Dn50.

2. NTIRE 2023 Image Denoising Challenge

The goals of this challenge include: (1) promoting research in the area of image denoising, (2) facilitating comparisons between various methods, and (3) providing a platform for academic and industrial participants to engage, discuss, and potentially establish collaborations. This section delves into the specifics of the challenge.

2.1. Dataset

The DIV2K [3, 42] dataset and LSDIR [27] dataset are utilized for this challenge. DIV2K dataset consists of 1,000 diverse 2K resolution RGB images, which are split into a training set of 800 images, a validation set of 100 images, and a test set of 100 images. LSDIR dataset contains 86,991 high-resolution high-quality images, which are split into a training set of 84,991 images, a validation set of 1,000 images, and a test set of 1,000 images. The training images from DIV2K and LSDIR are provided to the participants of the challenge. During the validation phase, the 100 images from DIV2K validation set were made available to participants. During test phase, 100 images from DIV2K test set and another 100 images from LSDIR test set are used. Throughout the entire challenge, the testing noise-free images remained hidden from participants.

2.2. Tracks and Competition

The aim is to obtain a network design capable to produce high-quality results with the best performance measured by PSNR for image denoising.

Challenge phases *(1) Development and validation phase*: Participants were given access to 800 clean training images and 100 clean/noisy validation image pairs from the DIV2K dataset. Additional 84,991 clean images from the LSDIR dataset are also provided to the participants. During training, the noisy images are generated by adding Gaussian noise with noise level $\sigma = 50$. Participants could upload their validation results to the evaluation server to calculate the PSNR of the images denoised by their models and receive immediate feedback. *(2) Testing phase*: In the final test phase, participants were granted access to 100 noisy testing images from DIV2K and 100 noisy testing images from LSDIR, while the clean ground-truth images remained hidden. Participants submitted their denoised results to the Codalab evaluation server and emailed the code and fact-sheet to the organizers. The organizers verified and ran the provided code to obtain the final results, which were then shared with participants at the end of the challenge.

Evaluation protocol Since the aim of this challenge is to foster the development of accurate image denoising networks, PSNR and SSIM on the 200 testing images are used as the quantitative evaluation metrics. A code example for calculating these metrics is available at https://github.com/ofsoundof/NTIRE2023_Dn50. The code of the submitted solutions and the pre-trained weights are also available in this repository.

Team	PSNR [dB]	SSIM	Ranking
Apply_AI	29.96	0.87	1
SRC-B	29.92	0.87	2
MiAlgo	29.87	0.86	3
HIT-IIL	29.86	0.87	4
MegNR	29.80	0.86	5
TeamYorku	29.79	0.86	6
IMAG_Denoising	29.79	0.86	7
cvmix	29.66	0.86	8
8080	29.40	0.85	9
LVGroup_HFUT	29.26	0.85	10
IMCgo	29.20	0.85	11
see you tomorrow	28.93	0.84	12
SRIB_AINR_23	28.73	0.85	13
TUK-IKLAB	28.35	0.82	14
MedI	28.13	0.82	15
yiriyou	27.94	0.82	16
Hunzy	18.96	0.46	

Table 1. NTIRE2023 image denoising ($\sigma = 50$) results. The PSNR and SSIM are calculated for on the test dataset of this challenge that contains 100 test images from DIV2K [3] dataset and 100 test images from LSDIR [27] dataset. The difference in SSIM metric between different images is very small. So the final ranking is based on PSNR.

3. Challenge Results

The final results of the image denoising challenge are shown in Tab. 1. In this table, the performance of 16 teams is ranked. The evaluation metric of one team is far below the others. Thus, this team is not ranked. The evaluation is based on two performance metrics including PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index). Since the validation set is already released to the participants, it is only used for reference in the validation phase of the challenge. The evaluation metrics are computed on the test set which contains 100 test images from DIV2K dataset [3] and 100 test images from LSDIR dataset [27]. This practice helps to keep the fairness of this challenge and discourages overfitting on the validation set. Since the difference between the SSIM of different teams is quite small, PSNR is used as the major ranking metric.

The top-ranked teams have higher PSNR and SSIM values, indicating better performance, while the lower-ranked teams have lower values, reflecting poorer performance. Comprehensively, the following is the breakdown analysis of the results:

- Among all the 16 teams, Team Apply_AI is the top-

performing team with the highest PSNR of 29.96 dB and an SSIM of 0.87.

- Team SRC-B closely follows as the second-ranked team, with a PSNR of 29.92 dB and an SSIM of 0.87.
- Team MiAlgo comes in third place, with a PSNR of 29.87 dB and an SSIM of 0.86.
- Team HIT-IIL is ranked fourth, with a PSNR of 29.86 dB and an SSIM of 0.87.
- Team MegNR holds the fifth position, with a PSNR of 29.80 dB and an SSIM of 0.86.
- Team TeamYorku ranks sixth, with a PSNR of 29.79 dB and an SSIM of 0.86.
- Team IMAG Denoising is in the seventh spot, with a PSNR of 29.79 dB and an SSIM of 0.86.
- Team cvmix ranks eighth, with a PSNR of 29.66 dB and an SSIM of 0.86.
- Team 8080 comes in ninth place, with a PSNR of 29.40 dB and an SSIM of 0.85.
- Team LVGroup HFUT is ranked tenth, with a PSNR of 29.26 dB and an SSIM of 0.85.
- Team IMCgo takes the eleventh position, with a PSNR of 29.20 dB and an SSIM of 0.85.
- Team see you tomorrow ranks twelfth, with a PSNR of 28.93 dB and an SSIM of 0.84.
- Team SRIB AINR 23 is in the thirteenth spot, with a PSNR of 28.73 dB and an SSIM of 0.85.
- Team TUK-IKLAB ranks fourteenth, with a PSNR of 28.35 dB and an SSIM of 0.82.
- Team MedI takes the fifteenth position, with a PSNR of 28.13 dB and an SSIM of 0.82.
- Team yiriyou ranks sixteenth, with a PSNR of 27.94 dB and an SSIM of 0.82.

3.1. Fairness

To guarantee fairness in the image denoising competition, multiple regulations were put in place, primarily focusing on the dataset utilized for training the network. Firstly, incorporating supplementary external datasets, like Flickr2K, was deemed acceptable. Secondly, the use of the additional DIV2K validation set was forbidden since it served to evaluate the comprehensive performance and generalizability of the proposed network. Thirdly, employing DIV2K test noisy images for training in any manner (supervised, unsupervised, image retrieval) was strictly forbidden. Finally, the implementation of sophisticated data augmentation techniques during the training process was considered an equitable approach.

3.2. Main ideas and architectures

To improve the performance of networks for image denoising, the participants investigated different techniques. In the following, the main ideas and architectures are summarized.

1. Due to the good performance of transformers for image restoration, transformer architectures are adopted by most of the teams. Team Apply_AI proposes an image processing transformer architecture, *i.e.* IPTV2 for image restoration. The solution of Team Samsung Research China - Beijing (SRC-B), TeamYorku, Team cvmix, and Team LVGroup_HFUT based their solutions on Restormer.

2. UNet architecture is adopted by most of the teams. For image denoising, UNet achieves a good balance between accuracy and efficiency. So it is used by most of the teams.

3. Progressive training helps to improve the performance of the network. It has been well known that increasing the patch size during training could lead to better image restoration results. The core of progressive training is to increase the patch size progressively during the training of the network, which could improve training efficiency compared with training with a large patch all the time.

4. Large-scale dataset helps to improve the accuracy of image denoising networks. In particular, LSDIR [27] is used as an additional dataset in this challenge. The additional dataset helps to boost the performance of the top-ranking teams.

5. Self-ensemble [43] or model ensembling are used to squire extra accuracy at test time.

4. Challenge Methods and Teams

4.1. Apply_AI

Inspired by [48, 50], we proposed an image processing transformer architecture for image restoration, namely IPTV2. As shown in the Fig. 1, IPTV2 is a U-shape encoder-decoder network as [36] with 3 times downsampling and upsampling. The basic module used in the IPTV2 is the spatial-channel transformer block, which helps fully capture both spatial interactions and channel interactions of feature maps. For spatial transformer, we split the feature map into small patches and get the self-attention map in the fixed window size for efficient computing as [34]. For the

channel transformer, we calculate the feature similarity of different channels with cosine distance. The spatial transformer and channel transformer are serially connected in the same stage, and feature maps of the post-downsampling layers are concatenated with those of the post-upsampling layers. With the input of $128 \times 128 \times 3$, the FLOPs and parameters number of IPTV2 is 41.16 GB and 26.03 M.

During the training phase, we use the flipping, rotating, RGB channel shuffling and mix-up strategies to enhance the original input image, and progressively train the model with resolutions of $[128, 192, 256, 320, 384]$. The model is jointly trained with L1, MSE, and SOBEL loss. And we only use the DIV2K and LSDIR [27] datasets in the training stage. Following Restormer [50], the optimizer and the scheduler of the learning rate in the training stage are AdamW and 'CosineAnnealingRestartCyclicLR'. During the inference phase, the original high-resolution image are split into patches of 384×384. For higher performance, model ensemble is also used in the inference phase.

4.2. Samsung Research China - Beijing (SRC-B)

Architecture. Transformer-based architecture has achieved great success in image restoration and related tasks, such as image denoising, image deblurring, and super-resolution. Our proposed denoising method is based on Restormer [50] which is an efficient Transformer model by making several key designs in the building blocks (multi-head attention and feed-forward network) such that it can capture long-range pixel interactions.

Progressive Learning. As mentioned in Restormer [50], training a transformer model on small cropped patches may not encode the global image statistics, thereby providing sub-optimal performance on full-resolution images at test time. To this end, we also perform progressive learning where the network is trained on different image patch sizes gradually enlarged from 256 to 320 and 448. As the patch size increases, the performance can gradually improve.

Feature Ensemble. As mentioned in Swinfir [51], Feature Ensemble is a novel ensemble strategy without lengthening the training and testing periods. We select multiple models that performed well on the validation dataset and combine them using the weighted average method. Feature Ensemble strategy can steadily improve the performance of the model.

Self-ensemble. In order to maximize the potential performance of our model, we adopt the self-ensemble strategy similarly to [31, 43]. In addition, we not only generate the outputs of flips and rotation but also generate the outputs of different patch sizes. We average the outputs together to make the final result. Self-ensemble can significantly improve performance.

4.3. MiAlgo

Recently, the development of deep learning-based image enhancement techniques has been advancing rapidly. Many state-of-the-art methods based on CNN and Transformer have achieved great success in tasks such as super-resolution reconstruction, image denoising, and image deblurring. In order to further explore the potential of deep learning-based methods in image denoising tasks, we propose a denoising network with 4 concatenations (D4C). As shown in Fig. 2, our pipeline can be divided into two stages. We employed the idea of ensemble learning to design our pipeline. In the first stage, we have selected four network architectures [7,8,29,50] that have achieved outstanding results in the field of image enhancement as the backbone to process the noisy images separately, and then fusion the results. In the second stage, we used another highly effective structure [46] as the refine module to optimize the previous results and further improve the image quality, including removing residual noise and enhancing image clarity. The design of the entire pipeline aims to maximize image quality although it increases a lot of computational complexity.

During the training phase, we first train the four basic networks of the first stage, each of which will be trained with sufficient data and fully trained to a converged state. After that, we fix the parameters of these networks and then begin training the refined network of the second stage. The training data used in these two stages are completely consistent.

4.4. HIT-IIL

Recently, some research in image generation has shown that the number of model parameters plays a critical role in model performance. Thus, instead of designing a new architecture, we directly scale up the existing network as our denoising model. We adopt NAFNet [7] as our basic network. And we find the results are better improved by increasing the number of channels than depth. Finally, limited by GPU resources, we only double the channel number of NAFNet. Please see NAFNet [7] for the method pipeline.

We use the provided DIV2K [3] and LSDIR [27] datasets as training images. The model is trained with PSNR loss. We utilize AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.9$) for 125K iterations on 8 NVIDIA A100 GPUs. The learning rate is initially 3×10^{-4} and gradually reduces to 1×10^{-7} with the cosine annealing. The training batch size is set to 64 and the patch size is 256×256.

In the inference phase, we use a self-ensemble strategy and selectively adopt the TLC method [11] based on the size of input images.

4.5. MegNR

For the Gaussian denoising task with a sigma of 50, we define it as an image restoration task that urgently requires

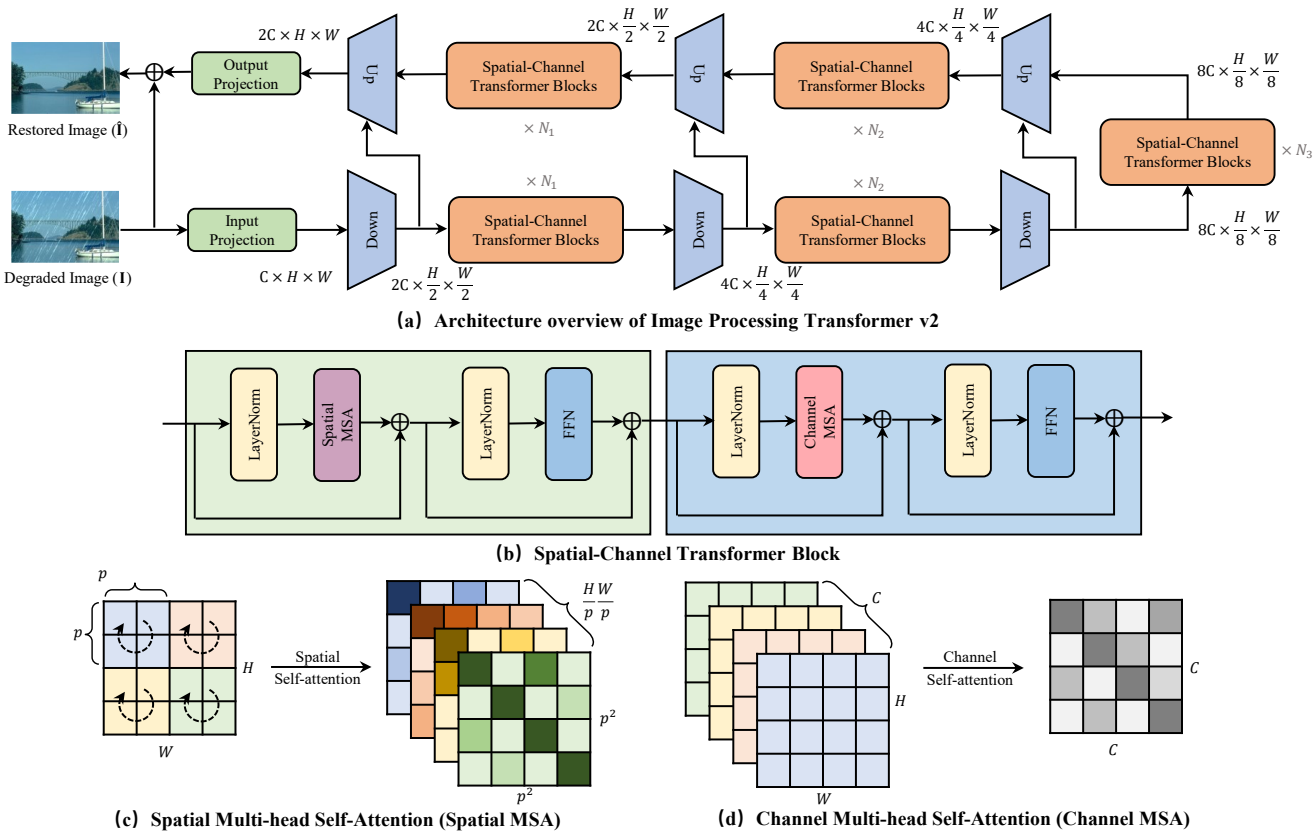

Figure 1. *Apply_AI team*. The architecture overview of the proposed Image Processing Transformer V2.

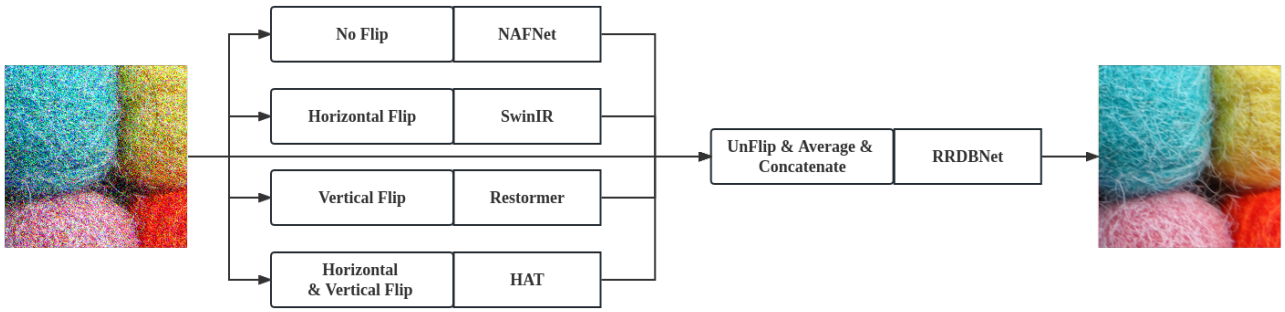

Figure 2. *MiAlgo team*. The architecture overview of the proposed D4C.

a model with a long receptive field. Inspired by the ideas of Uformer [48] and HAT [8], we combined the strengths of both networks to enable the network to see more pixels and have the ability to perform frequency division noise reduction. As shown in Fig. 3, we construct two modules including Hybrid Attention Local-Enhanced Block(**HALEB**) and Overlapping Cross-Attention Block(**OCAB**). The two modules replace the LeWin Blocks [48] and are used to capture more long-range dependencies information and useful local context. Specifically, to further refine the results, we incorporated a finetuned module consisting of 4 HALEB blocks. Moreover, we utilized the method of model ensembling(**ME**). Specifically, KBNET [56], restormer [50] and Mean-Invariant Denoising Diffusion Models(MIDPM) [18, 41] are integrated to fully utilize the performance of different models.

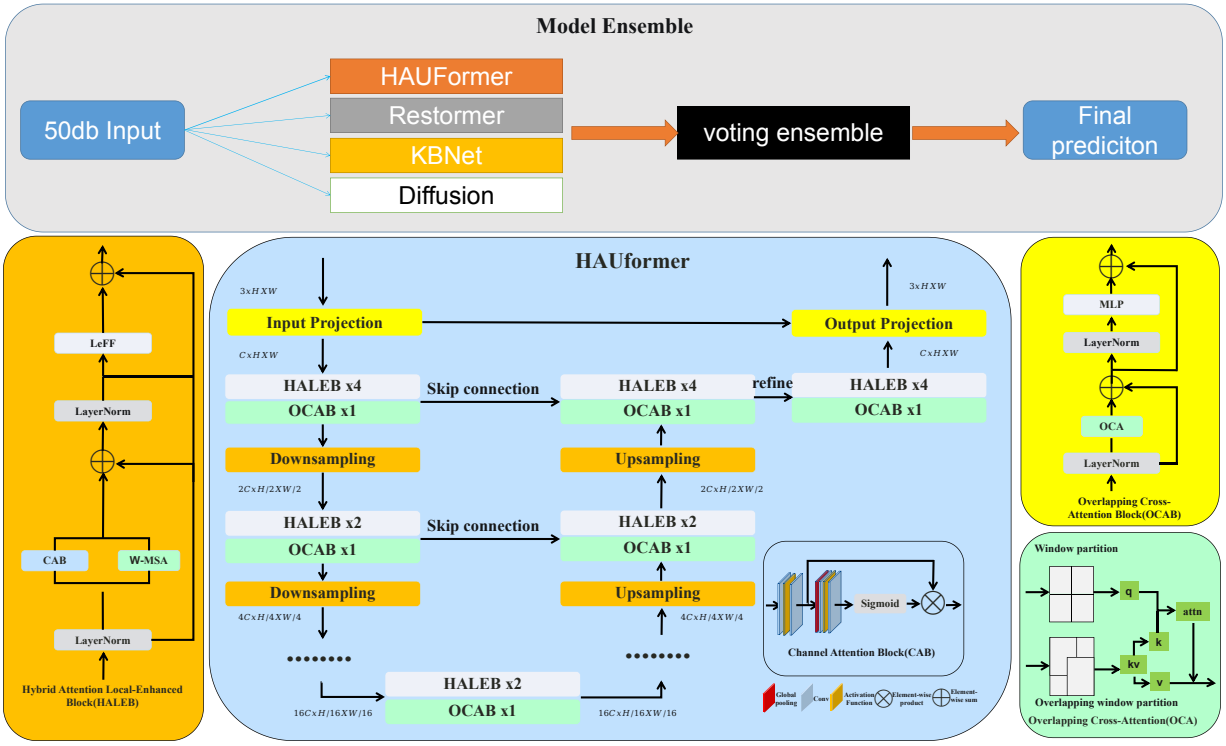

Figure 3. *MegNR team*. The architecture overview of the proposed HAUformer.

During the training phase, we train the model with MSE loss. The learning rate starts from 2×10^{-4} and is gradually reduced to 1×10^{-6} with the cosine annealing scheme. The entire training was conducted on the LSDIR dataset, using random rotation, cropping, and flipping operations. We randomly crop the training images into 128×128 sized patches with the 8-sized batches. Convergence was achieved after approximately 300,000 iterations. In the final inference stage, test-time augmentation [38] is used to get the final result.

4.6. TeamYorku

Our work is based on the Transformer-based architecture *Restormer* that is introduced in [50]. It is an efficient Transformer model that can handle high-resolution images for restoration tasks.

In Fig. 4 we present the overall pipeline of our Restormer architecture. Given a noisy image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, Restormer first applies a convolution to obtain low-level feature embeddings $\mathbf{F_0} \in \mathbb{R}^{H \times W \times C}$; where $H \times W$ denotes the spatial dimension and C is the number of channels. Next, these shallow features $\mathbf{F_0}$ pass through a 4-level symmetric encoder-decoder and are transformed into deep features $\mathbf{F_d} \in \mathbb{R}^{H \times W \times C}$. Starting from the high-resolution input, the encoder hierarchically reduces spatial size, while expanding channel capacity. For feature downsampling and upsampling, we apply pixel-unshuffle and pixel-shuffle operations, respectively. To assist the recovery process, the encoder features are concatenated with the decoder features via skip connections. Finally, a convolution layer is applied to the refined features to generate residual image $\mathbf{R} \in \mathbb{R}^{H \times W \times 3}$ to which degraded image is added to obtain the restored image: $\hat{\mathbf{I}} = \mathbf{I} + \mathbf{R}$.

In the proposed Transformer block, the core components are: **(a)** multi-Dconv head transposed attention (MDTA) and **(b)** gated-Dconv feed-forward network (GDFN).

The first one aims at reducing self-attention (SA) computational overhead. To alleviate this issue, we propose MDTA, shown in Fig. 4(a), that has linear complexity. The key ingredient is to apply SA across channels rather than the spatial dimension. Also as part of MDTA, we introduce depth-wise convolutions to emphasize on the local context before computing feature covariance to produce the global attention map.

The second component consists of two fully connected layers with a non-linearity in-between. As shown in Fig. 4(b), we reformulate the first linear transformation layer of the regular FN [14] with a gating mechanism to improve the information flow through the network. This gating layer is designed as the element-wise product of two linear projection layers, one of which is activated with the GELU non-linearity. The gating mechanism in GDFN con-

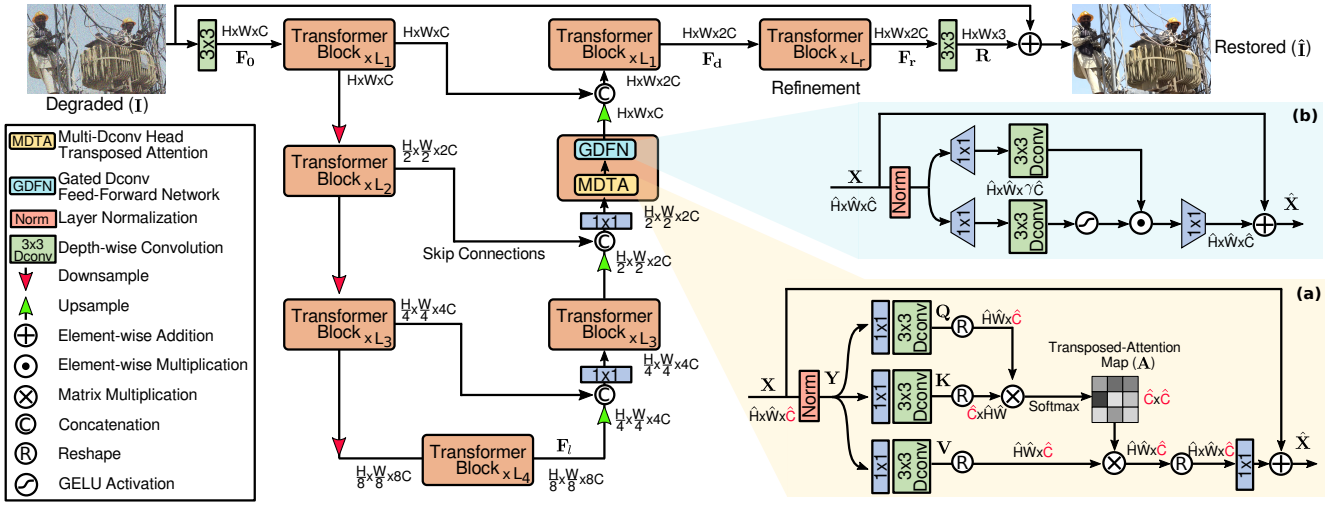

Figure 4. *TeamYorku*: Overall framework of Restormer [50].

trols which complementary features should flow forward and allows subsequent layers in the network to specifically focus on more refined image attributes, thus leading to high-quality outputs.

For training the image denoising model, we keep the sigma constant at 50 while adding the Additive White Gaussian Noise to the image. The model is trained on the provided 800 training images of the DIV2K dataset and 3000 images of the LSDIR dataset with AdamW optimizer and L_1 loss for 300K iterations with the initial learning rate 3×10^{-4} gradually reduced to 1×10^{-6} with the cosine annealing. For progressive learning, we start training with patch size 128×128 and batch size 64. The patch size and batch size pairs are updated to [(128^2,48), (160^2,32), (192^2,16), (224^2,16)] at iterations [100K, 170K, 220K, 260K]. For data augmentation, we use horizontal and vertical flips.

4.7. IMAG_Denoising

As shown in Fig. 5, our model consists of a stacking of Channle-aware Gated Feed-forward Blocks (CGFBs), while each CGFB contains two LayerNorm layers, a simplified channel attention module, and a frequency-based feed-forward network. Specifically, following [23], we employ the patch unfolding, Fast Fourier Transform (FFT) and folding operation to the feature, and introduce a learnable quantization matrix in the frequency-based feed-forward network, which determines which frequency information should be preserved.

We use DIV2K and LSDIR training dataset to train the model with Charbonnier and FFT loss for the first 200,000 iterations with the initial learning rate 1×10^{-3} gradually reduced to 1×10^{-6} for the rest 200,000 iterations with the cosine annealing scheme. After that, we finetune our model with L2 loss for 300,000 iterations with the initial learning rate 3×10^{-4} gradually reduced to 1×10^{-7} with the cosine annealing scheme.

4.8. cvmix

For this denoising task, we found that the noise level is fixed at 50. Then we decide to use the deep learning method to learn how to recover clean images from noisy images. We compare the existing methods. For the common, they are divided into two kinds of methods: CNN and Transformer. We compared the advantage and disadvantages of the methods above.

We decided to use NAFNet [7] as our baseline to do the task of denoising. NAFNet is based on the network of Restormer and turning the Transformer block to NAFNet block has the state-of-the-art performance in many image restoration tasks. The block assimilates CNN's feature, which is to use convolution to get global information. At the same time, the NAFNet block also uses the Channel Attention mechanism. Channel Attention is kind attention used to know which feature map channel is important to focus on.

Based on NAFNet, we divided the denoising task into 2 stages. The first stage is the main phase to recover images to clean one, then the images will be sent to the second stage. The second stage is focused to recover the detail. We use the default settings in NAFNet with 64 as the first-stage settings. Then we add residual in the bottleneck of the U-net framework to enhance the recovery detail ability. We reference the NAFNet in our project.

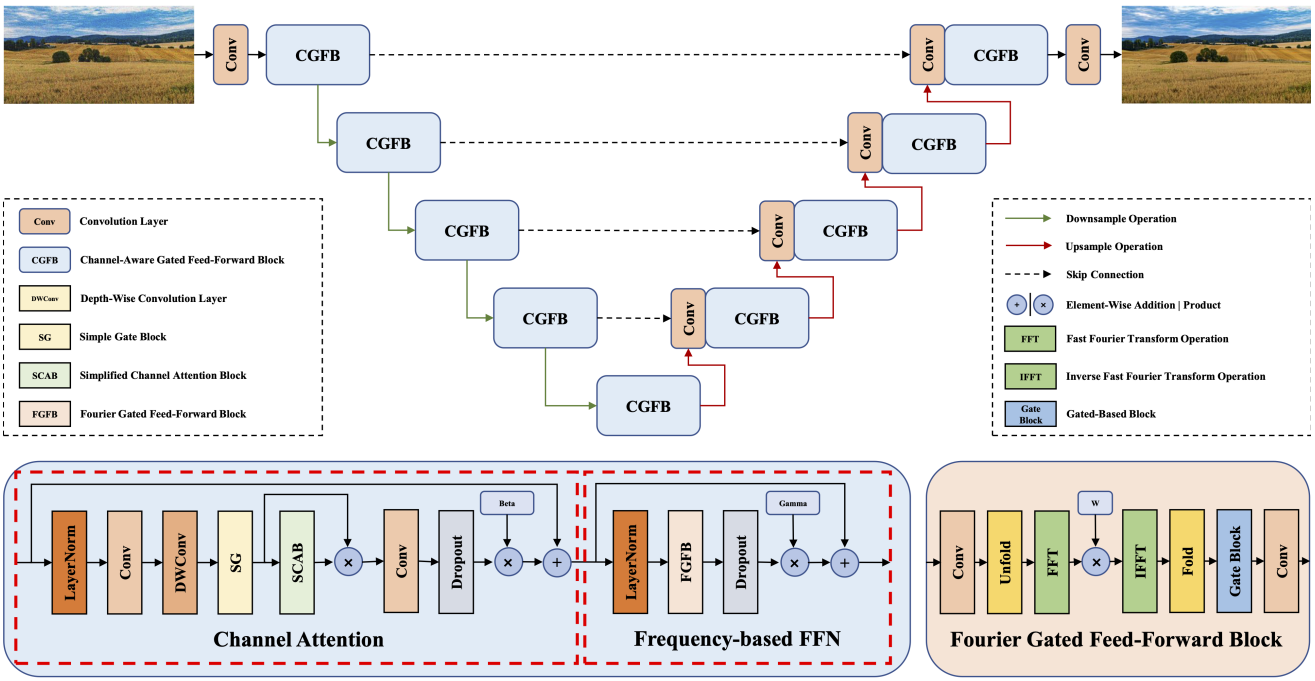

Figure 5. *Team IMAG_Denoising:* Network architectures. Our model consists of a stacking of Channle-aware Gated Feed-forward Blocks (CGFBs), while each CGFB contains a simplified channel attention and a frequency-based feed-forward network.

4.9. LVGroup_HFUT

To recover the clean images from the noisy images to, we employ a simple UNet architecture with self-attention (CATransformer), which can capture long-range pixel interactions by a multi-head self-attention layer and a multi-scale hierarchical module. Specifically, inspired by the Restormer [50], we add a skip connection between the multi-head self-attention and feed-forward network, and the skip connection is implemented by efficient channel attention. The framework of our CATransformer is shown in Fig. 6. The proposed has two following advantages: 1) the multi-scale hierarchical module can learn sufficient spatial structure information; 2) the multi-head self-attention and channel attention can exchange channel information to obtain sufficient global information. Experiments show that our model can implement the noisy image restoration task well and remain applicable to large images.

The proposed solution is implemented based on PyTorch version 1.10 and Nvidia RTX 3090 with 24G memory. During training, we perform a series of data augment operations sequentially as follows: 1)random crop to 256x256; 2)vertical flip with probability 0.5; 3)horizontal flip with probability 0.5. We train the model for 1000 epochs on provided training dataset with an initial learning rate 1×10^{-4}. Adam optimizer and multi-step learning rate scheduler are used. The learning rate is reduced by half every 500 epochs. L1 loss between the denoised image and the ground-truth images is used as the loss function.

4.10. IMCgo

The proposed method by team IMCgo comes from DDT [32]. The overall architecture is shown in Fig. 7. We use a 4-stage Unet-like encoder-decoder architecture. Inspired by [50], we introduce a refinement stage after the decoder, which aims to enhance feature representation for more details of images, and each stage consists of multiple Dual-branch Deformable Transformer Blocks (DDTB). Dual-Branch Deformable Attention (DDA), as the core module in DDTB, uses a dual-branch structure to do the local and global modeling in parallel. Specifically, in the local branch, we divide the feature into non-overlapping patches with pre-defined patch sizes and apply the spatial attention mechanism inside the patches. In the global brunch, we use a pre-defined number of patches to do the patch partitioning and perform calculations among the corresponding positions of each patch. To focus on more important regions, the deformable attention mechanism (Fig. 8) is applied in both branches for efficient spatial operation, which reduces the number of keys and values to reduce redundant calculations.

We set the number of DDTBs from the 1st stage to Bottleneck as (4, 6, 6, 8) with the number of attention heads (1, 2, 4, 8) and 4 extra blocks for the Refinement stage. We use AdamW and L_1 loss with 300K iterations for optimization. The learning rate is initialized as 3×10^{-4} and reduced to 1×10^{-6} with the cosine annealing scheduler. Progressive learning strategy [50] from 128×128 to 256×256 is

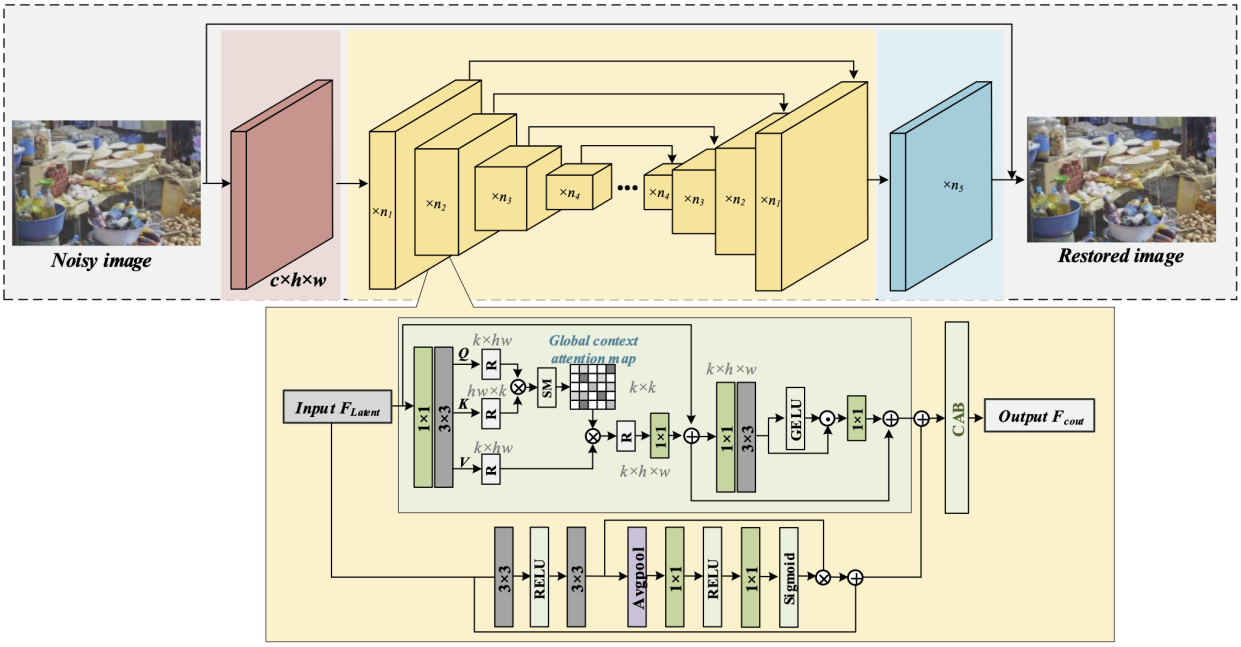

Figure 6. *Team LVGroup_HFUT:* The framework of CATransformer.

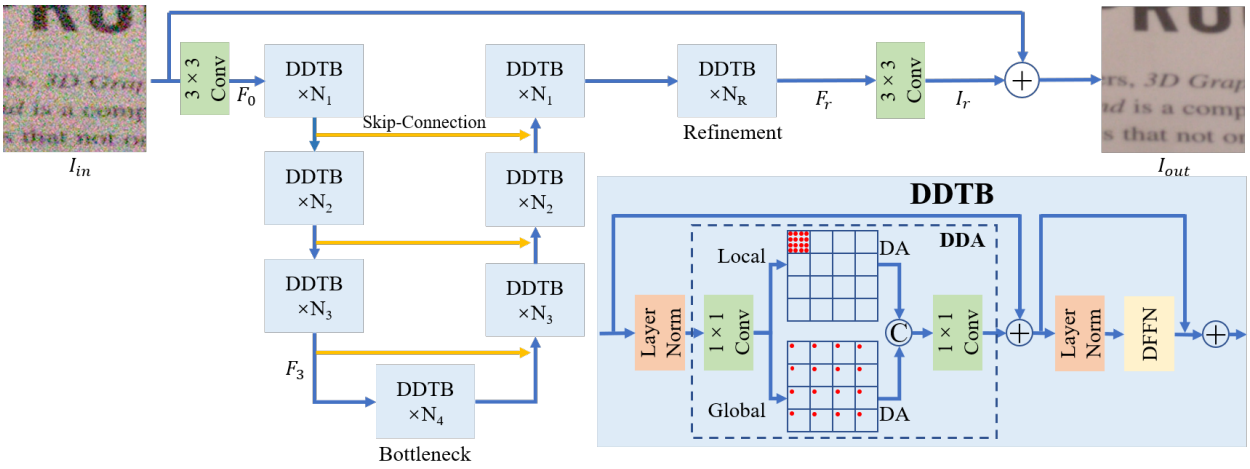

Figure 7. *Team IMCgo:* The overall architecture of Dual-branch Deformable Transformer (DDT).

also used, and rotation and flips are used for data augmentations.

4.11. SRIB_AINR_23

Deep learning-based solutions for image enhancement have been popular for quite some time now. However, not much attention has been given to efficient denoising for compute-limited and power-sensitive devices, such as smartphones. To this end, we propose a technique to optimize existing image enhancement networks that take RGB frames as input. Our solution builds on multiscale Del-Net [17] network as the backbone network which uses channel and spatial attention blocks and Enhanced Attention Module (EAM) [5]. We optimize this network by introducing a novel, hand-crafted feature re-arrangement block designed for RGB input frames. As shown in Fig. 9, given an RGB image of dimension h x w x 3, we convert this to feature representation using a re-arrangement block like PixelShuffle and pass it through Del-Net. We recover a 3 channel RGB frame by inverting the re-arrangement performed

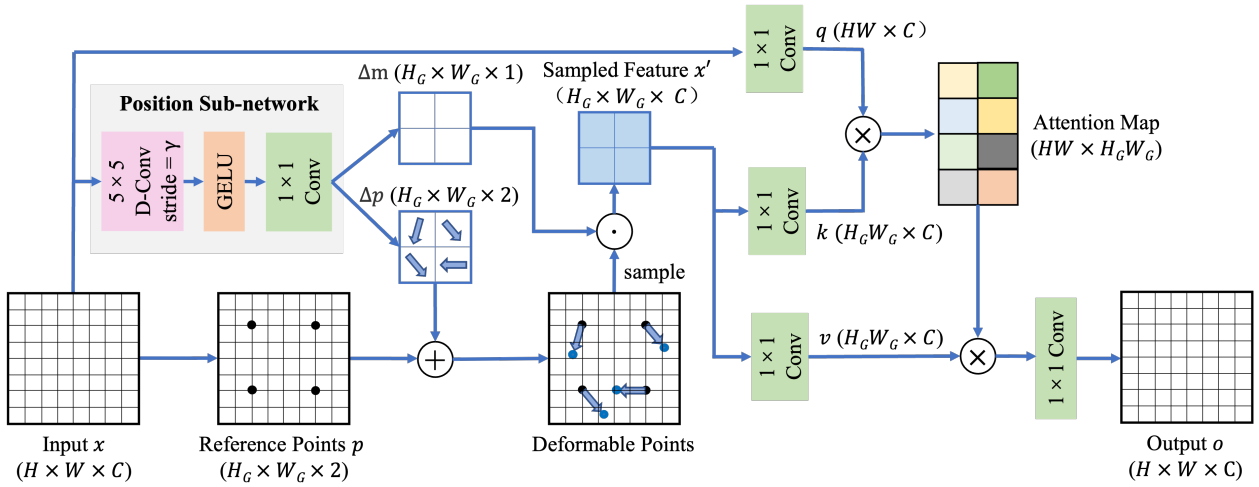

Figure 8. *Team IMCgo:* The structure of deformable attention (DA) in DDT.

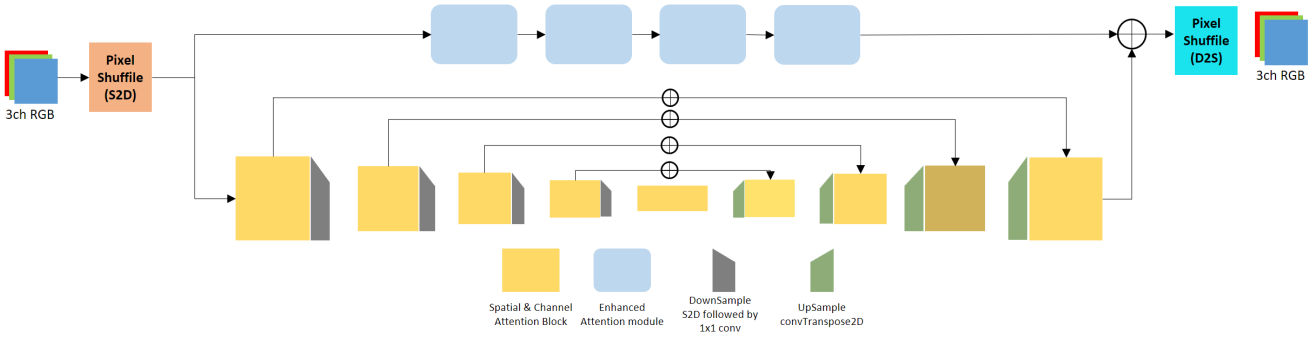

Figure 9. *Team SRIB_AINR_23:* The architecture overview of the proposed Updated DelNet for Image Denoising.

earlier from the output of Del-Net. The proposed network is trained on DIV2K dataset over random patches of size 256 and batch size 8 for over 2000 epochs. In addition to L1 and SSIM losses for training, we use Total Variational Loss to maintain and retain image quality. Total Variational Loss helps to guide the reconstruction of RGB frames from Del-Net output. Inference is performed over patches of size 256 with mirror padding. When compared against the baseline Del-Net, we observe that the inference time is reduced by a factor of 8.

4.12. IKLAB-TUK

Team IKLAB-TUK proposes Dense Residual Swin Transformers (DRSTNet) for Image Denoising. The proposed method is composed of four modules including Hierarchical Feature Extraction, Dense Residual Feature Enhancement, Fusion, and Residual Block, as shown in Fig. 10. Existing studies have revealed that using the hierarchical feature extraction module allows the network to extract meaningful representations from images at different scales in a divide-and-conquer manner [9,35]. Furthermore, it helps the network deal with complex and severe degradation in an efficient manner. The term hierarchical is used for this module as it extracts the representation from low-resolution (LR) images with three-step architecture that applies convolution operation with varying strides and kernels using three different scales. The implementation of hierarchical feature extraction modules is detailed in [35]. The first step comprises of padding, stride, number of channels, and kernel size, which are set to be 3, 1, 60, and 7, respectively. For the second step in this hierarchy, we follow the same convention for feature extraction but with the values, 2, 2, 60, and 5, followed by the third step that takes the values 1, 2, 60, and 3, respectively.

Existing works for image restoration and super-resolution consider convolutional neural networks as feature enhancers [25, 35, 55]. Recently, some of the studies considered Swin transformer block to enhance the features as well as model long-range dependencies [24,30,52]. Swin transformers have proven to be effective for such degrada-

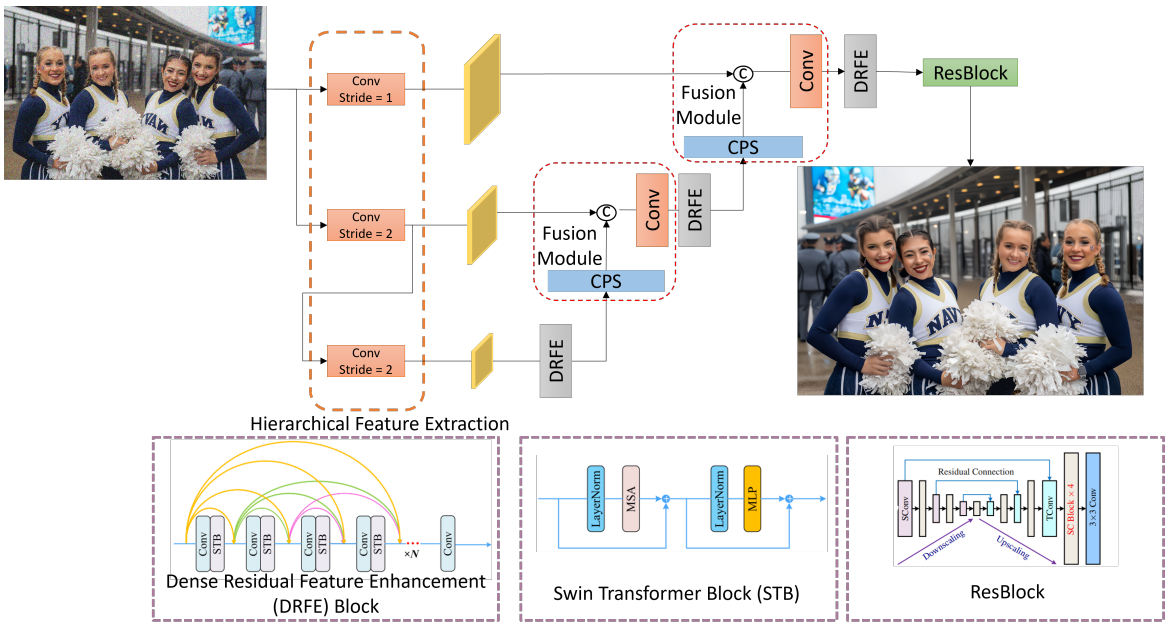

Figure 10. *Team TUK-IKLAB:* Architecture for the proposed dense residual Swin transformer network (DRSTNet) for Image Denoising.

tion tasks, including image restoration while yielding less number of parameters. We propose a dense residual feature enhancement (DRFE) block. As shown in Fig. 10, the DRFE block combines the Swin transformer layers with dense residual convolutional blocks. We use 4 layers for DRFE. The residual convolutional and Swin transformer blocks are then connected in a dense scheme. This dense connection scheme was inspired by DenseNet, which helps to deal with such a complex degradation task of image denoising. The Swin transformer blocks are further divided into Swin transformer layers, multi-head self-attention, and layer normalization, accordingly.

The feature fusion module undertakes the enhanced features from DRFE and performs the feature level fusion with the ascending hierarchical step module. Such a fusion strategy leverages contextual information while performing the fusion on features extracted and enhanced from the middle and lower branches. Within the fusion module, the features are upscaled and concatenated with the middle branch features. Similarly, the features from the middle branch are extracted and enhanced using DRFE, followed by the upscaling and concatenation with the feature maps from the first branch, respectively. The upsampling operation is performed using convolutional layers and PixelShuffle layer [39].
Finally, the last module is the ResBlock module which undertakes the enhanced features and outputs high-quality denoised RGB images. As shown in Fig. 10, the said module has strided convolution, and transposed convolutional layers, respectively. Lastly, a Swin-Conv block [52] and a convolutional layer are used to generate the denoised image.

4.13. MedI

This Fig. 11 shows the main parts of the proposed method. Inspired by the GRDB [22] and squeeze-excitation networks (SE) [19], this work proposes a channel attention-enhanced denoising network(CADN) as the generator. The detailed discriminator architecture shown in Fig. 1 is based on convolutional blocks. Here the position attention module (PAM) [15], also a space attention module is introduced to this discriminator to catch and merge deep features from max pooling and average pooling.

The important contribution of this work is to introduce grad-CAM loss l_{cam} of the generator. Considering the key to classifying an object is part of meaningful features in maps, this work introduces the classification activation map (CAM) loss function leading the generator to generate indistinguishable images. This work utilizes the Grad-CAM [37] method to calculate the I^{cam}_{normal} and $I^{cam}_{denoising}$ of I_{normal} and $I_{denoising}$ by the discriminator. l_{cam} can propel the generator networks to express more features of the target category and decrease attention to the background. A hybrid loss function of the generator for overcoming smooth and saving meaningful features is composed of image loss l_{img}, perceptual loss l_{per}, Sobel loss l_{sobel}, grad-CAM loss l_{cam} and adversarial loss l_{adv}. And the discriminator loss l_D is represented by binary cross-entropy (BCE) loss.

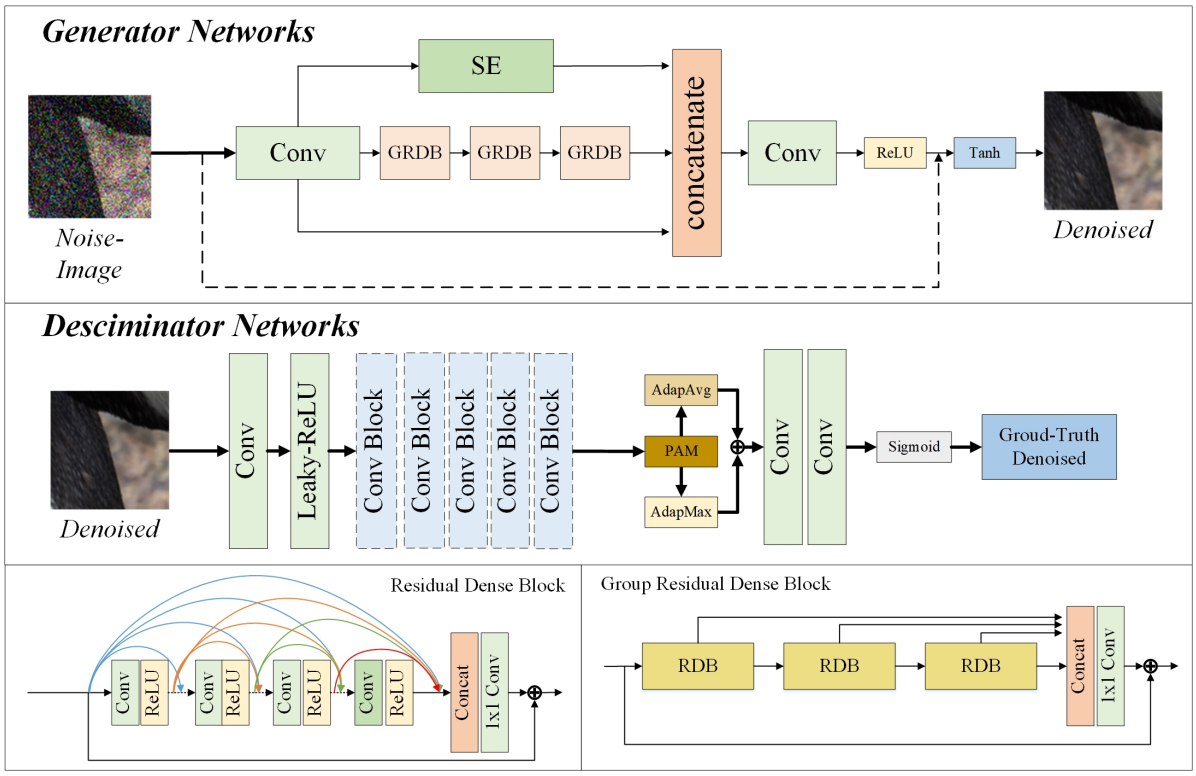

Figure 11. *Team MedI:* Channel attention-enhanced denoising network (CADN).

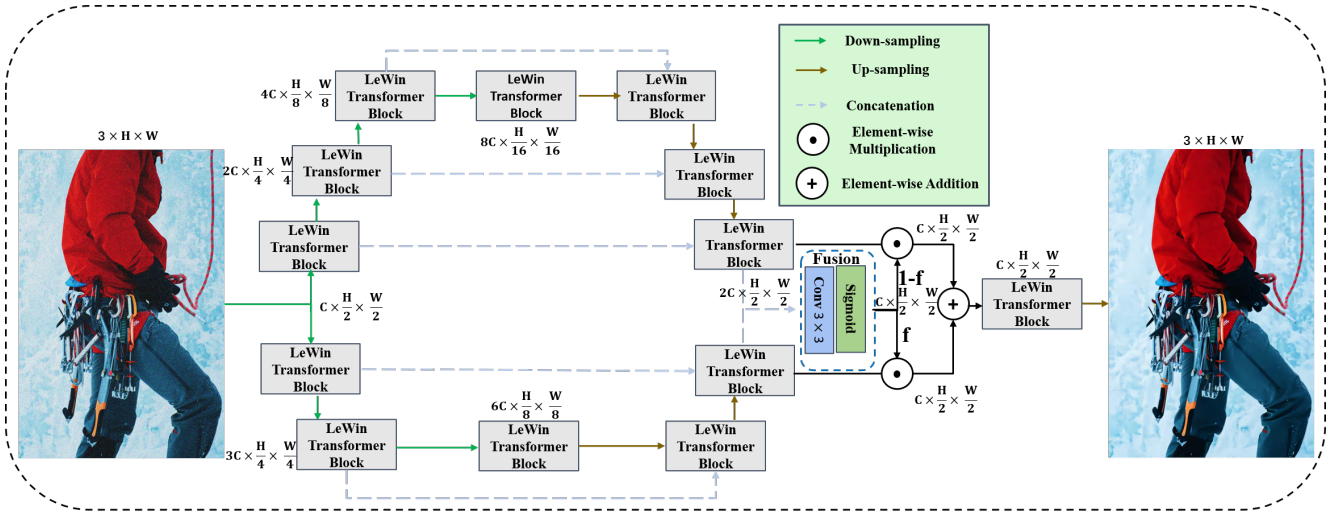

Figure 12. *Team yiriyou:* The architecture of the proposed framework.

4.14. yiriyou

We propose a dual-view U-Net based Transformer for high-resolution image restoration. As shown in Figure 12, our model consists of two sub-modules. Each of them models information from channel and pixel dimensions at different scales, respectively. Each up-sampling or down-sampling operation is followed by a **L**ocally-**e**nhanced **Win**dow (LeWin) Transformer block [48]. Benefiting from the hierarchical architecture of U-Net and the window mechanism, the LeWin Transformer block is capable of capturing long-range dependencies at low-resolution feature maps while reducing the computational cost. Specifically, we use 4×4 convolution block with a stride of 2 and a padding of 1 for down-sampling, and 2×2 trans-

posed convolution with stride of 2 for up-sampling. The top sub-module has three sub-layers, we triple the channels at each sub-layer along the down-sampling path and do the opposite along the up-sampling path. However, the bottom sub-module has only two sub-layers, we expand the channels hierarchically by a factor of 2 along the down-sampling path and reduce the channel capacity by half along the up-sampling path. After that, the output feature maps of the two sub-modules are concatenated and fed into the fusion block, which consists of a convolution and a sigmoid activation. Finally, an optimal combination between two feature maps can be attained and used to recover the clean image,

$$f = sigmoid(W\,[x, y] + b), \tag{1}$$

$$z = f \odot x + (1 - f) \odot y \tag{2}$$

where $[\,,\,]$ denotes tensor concatenation, $\odot$ denotes element-wise multiplication, W, b are trainable parameters, f denotes the learned weights, x and y denote the input feature maps and z denotes the weighted sum of feature maps.

In our implementation, the objective function is the L1-Loss, the batch size is 1, the window size is 8, the optimizer is Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.99$, the learning rate is 2e-3 with a decay factor of 0.5 for every 20 epochs, the total training epochs is 200.

Acknowledgments

This work was partially supported by the Humboldt Foundation. We thank the NTIRE 2023 sponsors: Sony Interactive Entertainment, Meta Reality Labs, ModelScope, ETH Zürich (Computer Vision Lab) and University of Würzburg (Computer Vision Lab).

A. Teams and affiliations

NTIRE 2023 team

Title: NTIRE 2023 Efficient Super-Resolution Challenge
Members:
Yawei Li[1] (yawei.li@vision.ee.ethz.ch),
Yulun Zhang[1] (yulzhang@ethz.ch),
Luc Van Gool[1] (vangool@vision.ee.ethz.ch),
Radu Timofte[1,2] (radu.timofte@uni-wuerzburg.de)
Affiliations:
[1] Computer Vision Lab, D-ITET, ETH Zürich, Switzerland
[2] Computer Vision Lab, IFI & CAIDAS, University of Würzburg, Germany

Apply_AI

Title: Image Processing Transformer V2
Members: Zhijun Tu[1] (zhijun.tu@huawei.com), Kunpeng Du[1], Hailing Wang[1], Hanting Chen[1], Wei Li[1], Xiaofei Wang[2], Jie Hu[1], Yunhe Wang[1]
Affiliations:
[1] Huawei Noah'Ark Lab
[2] Huawei Consumer Business Group

Samsung Research China - Beijing (SRC-B)

Title: Feature ensemble and self-ensemble apply on Restormer for denoising
Members:
Xiangyu Kong[1] (xiangyu.kong@samsung.com), Jinlong Wu[1], Dafeng Zhang[1], Jianxing Zhang[1]
Affiliations:
[1] Samsung Research China - Beijing (SRC-B)

MiAlgo

Title: Denoising network with 4 Concatenations (D4C)
Members:
Shuai Liu (liushuai21@xiaomi.com), Furui Bai, Chaoyu Feng, Hao Wang, Yuqian Zhang, Guangqi Shao, Xiaotao Wang, Lei Lei
Affiliations:
Xiaomi Inc., China

HIT-IIL

Title: Scaling Up NAFNet for Denoising
Members:
Rongjian Xu (ronjon.xu@gmail.com), Zhilu Zhang, Yunjin Chen, Dongwei Ren, Wangmeng Zuo
Affiliations:
Harbin Institute of Technology

MegNR

Title: HAUformer: Hybrid Attention-guided U-shaped Transformer for Gaussian image denoising
Members:
Qi Wu[1] (wuqi02@megvii.com), Mingyan Han[1], Shen Cheng[1], Haipeng Li[1], Ting Jiang[1], Chengzhi Jiang[1], Xinpeng Li[1], Jinting Luo[1], Wenjie Lin[1], Lei Yu[1], Haoqiang Fan[1] and Shuaicheng Liu[2,1*]
Affiliations:
[1] Megvii Technology
[2] University of Electronic Science and Technology of China (UESTC)

TeamYorku

Title: Restormer: Efficient Transformer for High-Resolution Image Restoration
Members:
Aditya Arora[1] (adityadvlp@gmail.com), Syed Waqas

Zamir[2], Javier Vazquez-Corral[3], Konstantinos G. Derpanis[1], Michael S. Brown[1]
Affiliations:
[1] York University, Toronto, Canada
[2] Inception Institute of Aritificial Intelligence (IIAI), Abu Dhabi, UAE
[3] Computer Vision Center, Universitat Autonòma de Barcelona, Spain

IMAG_Denoising

Title: Gated Asymmetric UNet with Channel Attention for Image Denoising
Members:
Hao Li [1] (lihao9605@gmail.com), Zhihao Zhao [1], Jinshan Pan [1], Jiangxin Dong [1], Jinhui Tang [1]
Affiliations:
[1] Nanjing University of Science and Technology

cvmix

Title: Dual-Unet Denoising
Members:
Bo Yang (yangboatsuzhou@gmail.com), Jingxiang Chen, Chenghua Li, Xi Zhang
Affiliations:
Nanjing University of Information Science and Technology

LVGroup_HFUT

Title: Channel Attention Guided Transformer for Image Denoising (CATransformer)
Members:
Zhao Zhang[1] (cszzhang@gmail.com), Jiahuan Ren[1], Zhicheng Ji[1], Kang Miao[1], Suiyi Zhao[1], Huan Zheng[1], YanYan Wei[1]
Affiliations:
[1] Hefei University of Technology (HFUT)

IMCgo

Title: DDT: Dual-Branch Deformable Transformer for image denoising
Members:
Kangliang Liu[1] (klliu21@m.fudan.edu.cn), Xiangcheng Du[1,2], Sijie Liu[1], Yingbin Zheng[2], Xingjiao Wu[1], Cheng Jin [1]
Affiliations:
[1] Fudan University
[2] Videt Technology

see you tomorrow

Title: see you tomorrow
Members:
Yan Wang (fywyan@mail.scut.edu.cn), Jiayin Chen, Xiaoxuan Wu, Huiming Chen, Xing Zheng, Yejia Chen
Affiliations:
South China University of Technology

SRIB_AINR_23

Title: Optimizing RGB image denoising using 12-Channel re-arrangement
Members:
Rajeev Irny[1] (rajeev.i@samsung.com), Sriharsha Koundinya[1], Vighnesh Kamath[1], , Gaurav Khandelwal[1]
Affiliations:
[1] Samsung R&D Institute India Bangalore (SRI-B)

IKLAB-TUK

Title: Dense Residual Swin Transformers for Image Denoising
Members:
Sunder Ali Khowaja[1] (sandar.ali@usindh.edu.pk), Jiseok Yoon[2], Ik Hyun Lee[3]
Affiliations:
[1] University of Sindh, Pakistan
[2] IKLAB Inc.
[3] IKLAB Inc. and Tech University of Korea, Republic of Korea

MedI

Title: CADN: Classification Activation Map Guided and Channel Attention-Enhanced Denoising Networks
Members:
Shijie Chen[1](shijie.chen.cn@gmail.com), Chengqiang Zhao[2]
Affiliations:
[1] Xuzhou Medical University
[2] Southwest Jiaotong University

yiriyou

Title: Dual-View U-Net Based Transformer for Super-Resolution Image Restoration
Members:
Huabin Yang [1] (huabinyang12@gmail.com), Zhongjian Zhang[1], Junjia Huang[1], Yanru Zhang [1]
Affiliations:
[1] University of Electronic Science and Technology of China

References

[1] Abdelrahman Abdelhamed, Mahmoud Afifi, Radu Timofte, and Michael S Brown. Ntire 2020 challenge on real image denoising: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 496–497, 2020. 1

[2] Abdelrahman Abdelhamed, Radu Timofte, and Michael S Brown. Ntire 2019 challenge on real image denoising: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[3] 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, 4

[4] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, Radu Timofte, et al. NTIRE 2023 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1

[5] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 9

[6] Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Ying Shan, Gen Li, Radu Timofte, et al. NTIRE 2023 challenge on 360° omnidirectional image and video super-resolution: Datasets, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1

[7] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple Baselines for Image Restoration. *arXiv e-prints*, page arXiv:2204.04676, Apr. 2022. 4, 7

[8] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*, 2022. 4, 5

[9] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Yan Shuicheng, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3434–3443, 2019. 10

[10] Yunjin Chen and Thomas Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2016. 1

[11] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Revisiting global statistics aggregation for improving image restoration. *arXiv preprint arXiv:2112.04491*, 2021. 4

[12] Marcos V Conde, Manuel Kolmet, Tim Seizinger, Thomas E. Bishop, Radu Timofte, et al. Lens-to-lens bokeh effect transformation. NTIRE 2023 challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1

[13] Marcos V Conde, Eduard Zamfir, Radu Timofte, et al. Efficient deep models for real-time 4k image super-resolution. NTIRE 2023 benchmark and report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1

[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 16×16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 6

[15] J Fu, J Liu, H Tian, Y Li, Y Bao, Z Fang, and H Lu. Dual Attention Network for Scene Segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3141–3149, 2019. 11

[16] Shuhang Gu and Radu Timofte. A brief review of image denoising algorithms and beyond. *Inpainting and Denoising Challenges*, pages 1–21, 2019. 1

[17] Saumya Gupta, Diplav Srivastava, Umang Chaturvedi, Anurag Jain, and Gaurav Khandelwal. Del-net: A single-stage network for mobile camera isp, 2021. 9

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 5

[19] J Hu, L Shen, S Albanie, G Sun, and E Wu. Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023, 2020. 11

[20] Andrey Ignatov, Kim Byeoung-Su, Radu Timofte, and Angeline Pouget. Fast camera image denoising on mobile gpus with deep learning, mobile ai 2021 challenge: Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2515–2524, 2021. 1

[21] Xiaoyang Kang, Xianhui Lin, Kai Zhang, Zheng Hui, Wangmeng Xiang, Jun-Yan He, Xiaoming Li, Peiran Ren, Xuansong Xie, Radu Timofte, et al. NTIRE 2023 video colorization challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1

[22] Dong Wook Kim, Jae Ryun Chung, and Seung Won Jung. GRDN:Grouped residual dense network for real image denoising and GAN-based real-world noise modeling. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2019-June:2086–2094, 2019. 11

[23] Lingshun Kong, Jiangxin Dong, Mingqiang Li, Jianjun Ge, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. *arXiv preprint arXiv:2211.12250*, 2022. 7

[24] Bingchen Li, Xin Li, Yiting Lu, Sen Liu, Ruoyu Feng, and Zhibo Chen. Hst: Hierarchical swin transformer for compressed image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 651–668, 2022. 10

[25] Xin Li, Simeng Sun, Zhizheng Zhang, and Zhibo Chen. Multi-scale grouped dense network for vvc intra coding. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 615–618, 2020. 10

[26] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. *arXiv preprint arXiv:2303.00748*, 2023. 1

[27] Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 2, 3, 4

[28] Yawei Li, Yulun Zhang, Luc Van Gool, Radu Timofte, et al. NTIRE 2023 challenge on efficient super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1

[29] 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. 1, 4

[30] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844, 2021. 10

[31] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1132–1140, 2017. 4

[32] Kangliang Liu, Xiangcheng Du, Sijie Liu, Yingbin Zheng, Xingjiao Wu, and Cheng Jin. DDT: Dual-branch deformable transformer for image denoising. In *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023. 8

[33] Xiaohong Liu, Xiongkuo Min, Wei Sun, Yulun Zhang, Kai Zhang, Radu Timofte, Guangtao Zhai, Yixuan Gao, Yuqin Cao, Tengchuan Kou, Yunlong Dong, Ziheng Jia, et al. NTIRE 2023 quality assessment of video enhancement challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1

[34] 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. 3

[35] Yingxue Pang, Xin Li, Xin Jin, Yaojun Wu, Jianzhao Liu, Sen Liu, and Zhibo Chen. Fan: Frequency aggregation network for real image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 468–483, 2020. 10

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[37] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. 11

[38] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1214–1223, 2021. 6

[39] 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. 11

[40] Alina Shutova, Egor Ershov, Georgy Perevozchikov, Ivan A Ermakov, Nikola Banic, Radu Timofte, Richard Collins, Maria Efimova, Arseniy Terekhin, et al. NTIRE 2023 challenge on night photography rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 5

[42] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 2

[43] Radu Timofte, Rasmus Rothe, and Luc Van Gool. Seven ways to improve example-based single image super resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1865–1873, 2016. 3, 4

[44] Florin-Alexandru Vasluianu, Tim Seizinger, Radu Timofte, et al. NTIRE 2023 image shadow removal challenge report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1

[45] Longguang Wang, Yulan Guo, Yingqian Wang, Juncheng Li, Shuhang Gu, Radu Timofte, et al. NTIRE 2023 challenge on stereo image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1

[46] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 4

[47] Yingqian Wang, Longguang Wang, Zhengyu Liang, Jungang Yang, Radu Timofte, Yulan Guo, et al. NTIRE 2023 challenge on light field image super-resolution: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1

[48] 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. 3, 5, 12

[49] Pierluigi Zama Ramirez, Fabio Tosi, Luigi Di Stefano, Radu Timofte, et al. NTIRE 2023 challenge on hr depth from im-

ages of specular and transparent surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1

[50] 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. 1, 3, 4, 5, 6, 7, 8

[51] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv preprint arXiv:2208.11247*, 2022. 4

[52] Kai Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Radu Timofte, and Luc Van Gool. Practical blind denoising via swin-conv-unet and data synthesis. *arXiv preprint arXiv:2203.13278*, 2022. 1, 10, 11

[53] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[54] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017. 1

[55] Xiaoshuai Zhang, Wenhan Yang, Yueyu Hu, and Jiaying Liu. Dmcnn: Dual-domain multi-scale convolutional neural network for compression artifacts removal. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 390–394, 2018. 10

[56] Yi Zhang, Dasong Li, Xiaoyu Shi, Dailan He, Kangning Song, Xiaogang Wang, Honwei Qin, and Hongsheng Li. Kbnet: Kernel basis network for image restoration. *arXiv preprint arXiv:2303.02881*, 2023. 5

[57] Yulun Zhang, Kai Zhang, Zheng Chen, Yawei Li, Radu Timofte, et al. NTIRE 2023 challenge on image super-resolution (x4): Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023. 1