# MAXIM: Multi-Axis MLP for Image Processing

Zhengzhong Tu[1,2*] Hossein Talebi[1] Han Zhang[1] Feng Yang[1]
Peyman Milanfar[1] Alan Bovik[2] Yinxiao Li[1]
[1] Google Research [2] University of Texas at Austin

## Abstract

*Recent progress on Transformers and multi-layer perceptron (MLP) models provide new network architectural designs for computer vision tasks. Although these models proved to be effective in many vision tasks such as image recognition, there remain challenges in adapting them for low-level vision. The inflexibility to support high-resolution images and limitations of local attention are perhaps the main bottlenecks. In this work, we present a multi-axis MLP based architecture called MAXIM, that can serve as an efficient and flexible general-purpose vision backbone for image processing tasks. MAXIM uses a UNet-shaped hierarchical structure and supports long-range interactions enabled by spatially-gated MLPs. Specifically, MAXIM contains two MLP-based building blocks: a multi-axis gated MLP that allows for efficient and scalable spatial mixing of local and global visual cues, and a cross-gating block, an alternative to cross-attention, which accounts for cross-feature conditioning. Both these modules are exclusively based on MLPs, but also benefit from being both global and 'fully-convolutional', two properties that are desirable for image processing. Our extensive experimental results show that the proposed MAXIM model achieves state-of-the-art performance on more than ten benchmarks across a range of image processing tasks, including denoising, deblurring, deraining, dehazing, and enhancement while requiring fewer or comparable numbers of parameters and FLOPs than competitive models. The source code and trained models will be available at* `https://github.com/google-research/maxim`.

## 1. Introduction

Image processing tasks, such as restoration and enhancement, are important computer vision problems, which aim to produce a desired output from a degraded input. Various types of degradations may require different image enhancement treatments, such as denoising, deblurring, super-resolution, dehazing, low-light enhancement, and so on. Given the increased availability of curated large-scale training datasets, recent high-performing approaches [13, 15, 18, 20, 47, 48, 56, 100, 101, 115] based on highly designed convolutional neural network (CNN) have demonstrated state-of-the-art (SOTA) performance on many tasks.

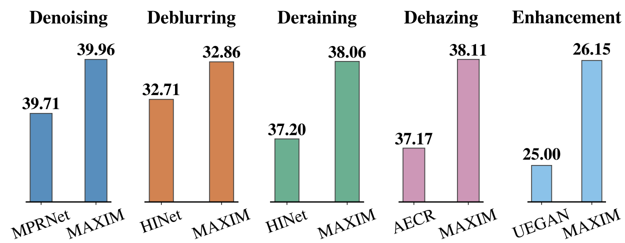



Figure 1. Our proposed MAXIM model significantly advances state-of-the-art performance on five image processing tasks in terms of PSNR: 1) Denoising (+**0.24** dB on SIDD [1]), 2) Deblurring (+**0.15** dB on GoPro [57]) 3) Deraining (+**0.86** dB on Rain100L [95]), 4) Dehazing (+**0.94** dB on RESIDE [43]), and 5) Retouching (Enhancement) (+**1.15** dB on FiveK [6]).

Improving the architectural design of the underlying model is one of the keys to improving the performance of most computer vision tasks, including image restoration. Numerous researchers have invented or borrowed individual modules or building blocks and implemented them into low-level vision tasks, including residual learning [40, 86, 110], dense connections [86, 111], hierarchical structures [34, 38, 39], multi-stage frameworks [14, 32, 101, 103], and attention mechanisms [60, 83, 100, 101].

Recent research explorations on Vision Transformers (ViT) [9, 22, 53] have exemplified their great potential as alternatives to the go-to CNN models. The elegance of ViT [22] has also motivated similar model designs with simpler global operators such as MLP-Mixer [79], gMLP [50], GFNet [69], and FNet [41], to name a few. Despite successful applications to many high-level tasks [3, 22, 53, 77, 81, 93], the efficacy of these *global* models on low-level enhancement and restoration problems has not been studied extensively. The pioneering works on Transformers for low-level vision [8, 13] directly applied full self-attention, which only accepts relatively small patches of fixed sizes (e.g.,

*Work done during an internship at Google.

48×48). Such a strategy will inevitably cause patch boundary artifacts when applied on larger images using cropping [13]. Local-attention based Transformers [48,88] ameliorate this issue, but they are also constrained to have limited sizes of receptive field, or to lose non-locality [22,85], which is a compelling property of Transformers and MLP models relative to hierarchical CNNs.

To overcome these issues, we propose a generic image processing network, dubbed **MAXIM**, for low-level vision tasks. A key design element of MAXIM is the use of *multi-axis* approach (Sec. 3.2) that captures both local and global interactions in parallel. By mixing information on *a single axis* for each branch, this MLP-based operator becomes '*fully-convolutional*' and scales linearly with respect to image size, which significantly increases its flexibility for dense image processing tasks. We also define and build a pure MLP-based cross-gating module (Sec. 3.3), which adaptively *gate* the skip-connections in the neck of MAXIM using the same multi-axis approach, and which further boosts performance. Inspired by recent restoration models, we develop a simple but effective multi-stage, multi-scale architecture consisting of a stack of MAXIM backbones. MAXIM achieves strong performance on a range of image processing tasks, while requiring very few number of parameters and FLOPs. Our contributions are:

- A novel and generic architecture for image processing, dubbed MAXIM, using a stack of encoder-decoder backbones, supervised by a multi-scale, multi-stage loss.
- A multi-axis gated MLP module tailored for low-level vision tasks, which always enjoys a global receptive field, with linear complexity relative to image size.
- A cross gating block that cross-conditions two separate features, which is also global and fully-convolutional.
- Extensive experiments show that MAXIM achieves SOTA results on more than 10 datasets including denoising, deblurring, deraining, dehazing, and enhancement.

## 2. Related Work

**Restoration models.** Driven by recent enormous efforts on building vision benchmarks, learning-based models, especially CNN models, have been developed that attain state-of-the-art performance on a wide variety of image enhancement tasks [13–15, 34, 47, 48, 74, 101]. These increased performance gains can be mainly attributed to novel architecture designs, and/or task-specific modules and units. For instance, UNet [73] has incubated many successful encoder-decoder designs [18, 34, 101] for image restoration that improve on earlier single-scale feature processing models [42, 110]. Advanced components developed for high-level vision tasks have been brought into low-level vision tasks as well. Residual and dense connections [40, 86, 86, 110, 111], the multi-scale feature learning [18,38,88], attention mechanisms [60,83,100,101,111], and non-local networks [49, 85, 111] are such good examples. Recently, *multi-stage* networks [14,32,101,103] have attained promising results relative to the aforementioned *single-stage* models on the challenging deblurring and deraining tasks [21, 32, 101]. These multi-stage frameworks are generally inspired by their success on higher-level problems such as pose estimation [16, 45], action segmentation [23,44], and image generation [106,107].

**Low-level vision Transformers.** Transformers were originally proposed for NLP tasks [82], where multi-head self-attention and feed-forward MLP layers are stacked to capture non-local interactions between words. Dosovitskiy *et al*. coined the term Vision Transformer (ViT) [22], and demonstrated the first pure Transformer model for image recognition. Several recent studies explored Transformers for low-level vision problems, *e.g*., the pioneering pre-trained image processing Transformer (IPT) [13]. Similar to ViT, IPT directly applies vanilla Transformers to image patches. The authors of [8] presented a spatial-temporal convolutional self-attention network that exploits local information for video super-resolution. More recently, SwinIR [48] and UFormer [88] apply efficient window-based local attention models on a range of image restoration tasks.

**MLP vision models.** More recently, several authors have argued that when using a patch-based architecture as in ViT, the necessity of complex self-attention mechanisms becomes questionable. For instance, MLP-Mixer [79] adopts a simple token-mixing MLP to replace self-attention in ViT, resulting in an all-MLP architecture. The authors of [50] proposed the gMLP, which applies a spatial gating unit on visual tokens. ResMLP [80] adopts an Affine transformation as a substitute to Layer Normalization for acceleration. Very recent techniques such as FNet [41] and GFNet [69] demonstrate the simple Fourier Transform can be used as a competitive alternative to either self-attention or MLPs.

## 3. Our Approach: MAXIM

We present, to the best of our knowledge, the first effective general-purpose MLP architecture for low-level vision, which we call **M**ulti-**AX**is MLP for **IM**age processing (**MAXIM**). Unlike previous low-level Transformers [8, 13, 48, 88], MAXIM has several desired properties, making it intriguing for image processing tasks. First, MAXIM expresses global receptive fields on arbitrarily large images with linear complexity; Second, it directly supports arbitrary input resolutions, *i.e*., being fully-convolutional; Lastly, it provides a balanced design of local (`Conv`) and global (`MLP`) blocks, outperforming SOTA methods without the necessity for large-scale pre-training [13].

### 3.1. Main Backbone

The MAXIM backbone (Fig. 2a) follows the encoder-decoder design principles that originated with UNet [73].

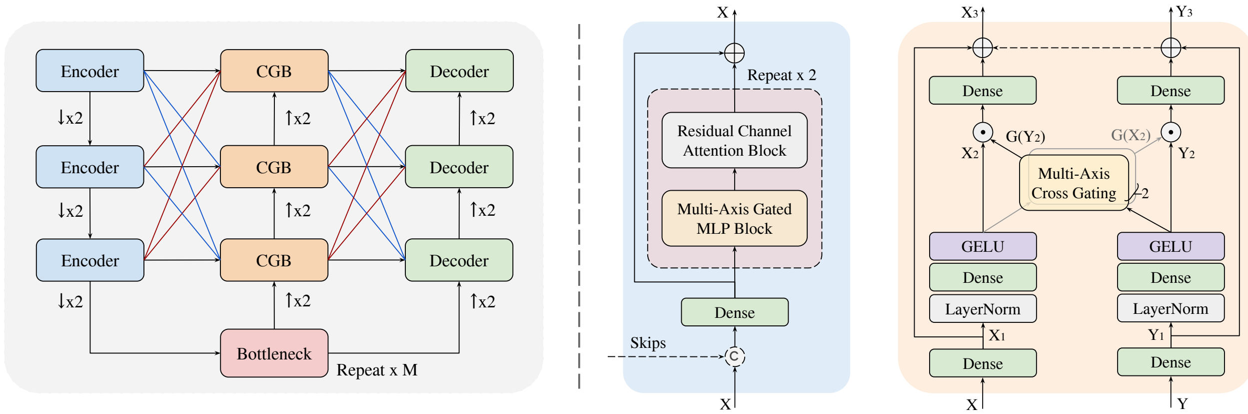



Figure 2. **MAXIM architecture.** We take (a) an encoder-decoder backbone with each (b) encoder, decoder, and bottleneck containing a multi-axis gated MLP block (Fig. 3) as well as a residual channel attention block. The model is further boosted by (c) a cross gating block which allows global contextual features to gate the skip-connections. More details can be found in supplementary materials.

We have observed that operators having small footprints such as `Conv3x3` are essential to the performance of UNet-like networks. Thus, we rely on a hybrid model design for each block (Fig. 2b) – `Conv` for local, and `MLP` for long-range interactions – to make the most of them.

To allow long-range spatial mixing at different scales, we insert the multi-axis gated MLP block (MAB) into each encoder, decoder, and bottleneck (Fig. 2b), with a residual channel attention block (RCAB) [91, 101] (`LayerNorm-Conv-LeakyReLU-Conv-SE` [29]) stacked subsequently. Inspired by the gated filtering of skip connections [61, 65], we extend the gated MLP (gMLP) to build a cross gating block (CGB, Fig. 2c), which is an efficient 2nd-order alternative to cross-attention (3rd-order correlations), to interact, or condition two distinct features. We leverage the global features from Bottleneck (Fig. 2a) to gate the skip connections, while propagating the refined global features upwards to the next CGB. Multi-scale feature fusion [18,76,100] (red and blue lines) is utilized to aggregate multi-level information in the Encoder→CGB and CGB→Decoder dataflow.

### 3.2. Multi-Axis Gated MLP

Our work is inspired by the multi-axis blocked self-attention proposed in [113], which performs attention on more than a single axis. The attentions performed on two axes on blocked images correspond to two forms of sparse self-attention, namely regional and dilated attention. Despite capturing local and global information in parallel, this module cannot accommodate image restoration or enhancement tasks where the test images are often of arbitrary sizes.

We improve the '*multi-axis*' concept for image processing tasks, by building a (split-head) multi-axis gated MLP block (MAB), as shown in Fig. 3. Instead of applying multi-axis attention in a single layer [113], we split in half the heads first, each being partitioned independently. In the local branch, the half head of a feature of size $(H, W, C/2)$ is *blocked* into a tensor of shape $(\frac{H}{b} \times \frac{W}{b}, b \times b, C/2)$, representing partitioning into non-overlapping windows each with size of $(b \times b)$; in the global branch, the other half head is *gridded* into the shape $(d \times d, \frac{H}{d} \times \frac{W}{d}, C/2)$ using a fixed $(d \times d)$ grid, with each window having size $(\frac{H}{d} \times \frac{W}{d})$. For visualization, we set $b = 2, d = 2$ in Fig. 3. To make it *fully-convolutional*, we only apply the gated MLP (gMLP) block [50] on *a single axis* of each branch – the 2nd axis for the local branch and the 1st axis for the global branch – while sharing parameters on the other spatial axes. Intuively, applying multi-axis gMLPs in parallel correspond to local and global (dilated) mixing of spatial information, respectively. Finally, the processed heads are concatenated and projected to reduce the number of channels, which are further combined using the long skip-connection from the input. It is worth noting that this approach provides an advantage for our model over methods that process fixed-size image patches [13] by avoiding patch boundary artifacts.

**Complexity analysis.** The computational complexity of our proposed Multi-Axis gMLP block (MAB) is:

$$\Omega(\text{MAB}) = \underbrace{d^2HWC}_{\text{Global gMLP}} + \underbrace{b^2HWC}_{\text{Local gMLP}} + \underbrace{10HWC^2}_{\text{Dense layers}}, \quad (1)$$

which is *linear* with respect to image size $HW$, while other global models like ViT, Mixer, and gMLP are *quadratic*.

**Universality of the multi-axis approach.** Our proposed parallel multi-axis module (Fig. 3) presents a principled way to apply 1D operators on 2D images in a scalable manner. It also allows for significant flexibility and universality. For example, a straightforward replacement of a gMLP with a spatial MLP [79], self-attention [22], or even Fourier Transform [41, 69] leads to a family of MAXIM variants (see Sec. 4.3D), all sharing globality and fully-convolutionality. It is also easily extensible to *any* future 1D operator that may be defined on, *e.g.*, Language models.

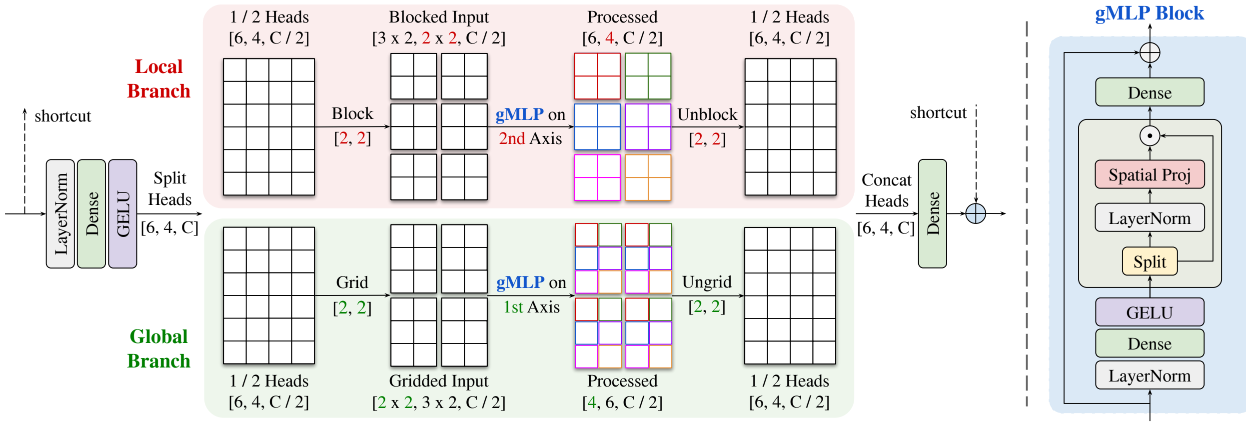



Figure 3. **Multi-axis gated MLP block** (best viewed in color). The input is first projected to a $[6, 4, C]$ feature, then split into two heads. In the local branch, the half head is *blocked* into $3 \times 2$ non-overlapping $[2, 2, C/2]$ patches, while we *grid* the other half using a $2 \times 2$ grid in the global branch. We only apply the gMLP block [50] (illustrated in the right gMLP Block) on *a single axis* of each branch - the 2nd axis for the local branch and the 1st axis for the global branch, while shared along the other spatial dimensions. The gMLP operators, which run in parallel, correspond to local and global (dilated) attended regions, as illustrated with different colors (*i.e.*, the same color are spatially mixed using the gMLP operator). Our proposed block expresses both global and local receptive fields on arbitrary input resolutions.

## 3.3. Cross Gating MLP Block

A common improvement over UNet is to leverage contextual features to selectively *gate* feature propagation in skip-connections [61, 65], which is often achieved by using cross-attention [11, 82]. Here we build an effective alternative, namely cross-gating block (CGB, Fig. 2c), as an extension of MAB (Sec. 3.2) which can only process a single feature. CGB can be regarded as a more general conditioning layer that interacts with multiple features [11, 64, 82]. We follow similar design patterns as those used in MAB.

To be more specific, let $\mathbf{X}, \mathbf{Y}$ be two input features, and $\mathbf{X}_1, \mathbf{Y}_1 \in \mathbb{R}^{H \times W \times C}$ be the features projected after the first `Dense` layers in Fig. 2c. Input projections are then applied:

$$\mathbf{X}_2 = \sigma(\mathbf{W}_1 \mathsf{LN}(\mathbf{X}_1)) \ , \ \ \mathbf{Y}_2 = \sigma(\mathbf{W}_2 \mathsf{LN}(\mathbf{Y}_1)) \tag{2}$$

where $\sigma$ is the $\mathsf{GELU}$ activation [28], $\mathsf{LN}$ is Layer Normalization [4], and $\mathbf{W}_1, \mathbf{W}_2$ are `MLP` projection matrices. The multi-axis blocked gating weights are computed from $\mathbf{X}_2, \mathbf{Y}_2$, respectively, but applied *reciprocally*:

$$\hat{\mathbf{X}} = \mathbf{X}_2 \odot \mathrm{G}(\mathbf{Y}_2) \ , \ \ \hat{\mathbf{Y}} = \mathbf{Y}_2 \odot \mathrm{G}(\mathbf{X}_2) \tag{3}$$

where $\odot$ represents element-wise multiplication, and the function $\mathrm{G}(\cdot)$ extracts multi-axis cross gating weights from the input using our proposed multi-axis approach (Sec. 3.2):

$$\mathrm{G}(\mathbf{x}) = \mathbf{W}_5([\mathbf{W}_3 \mathsf{Block}_b(\mathbf{z_1}), \mathbf{W}_4 \mathsf{Grid}_d(\mathbf{z_2})]) \tag{4}$$

where $[\cdot, \cdot]$ denotes concatenation. Here $(\mathbf{z_1}, \mathbf{z_2})$ are two independent heads split from $\mathbf{z}$ along the channel dimension, where $\mathbf{z}$ represents the projected features $\mathbf{x}$ after activation:

$$[\mathbf{z_1}, \mathbf{z_2}] = \mathbf{z} = \sigma(\mathbf{W}_6 \mathsf{LN}(\mathbf{x})), \tag{5}$$

and $\mathbf{W}_3, \mathbf{W}_4$ are spatial projection matrices applied on the 2nd and 1st axis of the blocked/gridded features having fixed window size $b \times b$ ($\mathsf{Block}_b$), and fixed grid size of $d \times d$ ($\mathsf{Grid}_d$), respectively. Finally, we adopt residual connection from the inputs, following an output channel-projection that maintains the same channel dimensions as the inputs $(\mathbf{X}_1, \mathbf{Y}_1)$, using projection matrices $\mathbf{W}_7, \mathbf{W}_8$, denoted by

$$\mathbf{X}_3 = \mathbf{X}_1 + \mathbf{W}_7 \hat{\mathbf{X}} \ , \ \ \mathbf{Y}_3 = \mathbf{Y}_1 + \mathbf{W}_8 \hat{\mathbf{Y}}. \tag{6}$$

The complexity of CGB is also tightly-bounded by Eq. (1).

## 3.4. Multi-Stage Multi-Scale Framework

We further adopt a multi-stage framework because we find it more effective, as compared to scaling up the model width or height (see ablation Sec. 4.3A). We deem full resolution processing [14, 63, 70] a better approach than a multi-patch hierarchy [75, 101, 103], since the latter would potentially induce boundary effects across patches. To impose stronger supervision, we apply a multi-scale approach [16, 18, 45] at each stage to help the network learn. We leverage the supervised attention module [101] to propagate attentive features progressively along the stages. We leverage the cross-gating block (Sec. 3.3) for cross-stage feature fusion. See supplementary materials for details.

Formally, given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we first extract its multi-scale variants by downscaling: $\mathbf{I}_n$, $n = 1, ..., N$. MAXIM predicts multi-scale restored outputs at each stage $s$ of $S$ stages, yielding a total of $S \times N$ outputs: $\mathbf{R}_{s,n}$. Despite being multi-stage, MAXIM is trained *end-to-end* with losses accumulating across stages and scales:

$$\mathcal{L} = \sum_{s=1}^{S} \sum_{n=1}^{N} [\mathcal{L}_{char}(\mathbf{R}_{s,n}, \mathbf{T}_n) + \lambda \mathcal{L}_{freq}(\mathbf{R}_{s,n}, \mathbf{T}_n)], \tag{7}$$

where $\mathbf{T}_n$ denotes (bilinearly-rescaled) multi-scale target

| Method | SIDD [1] PSNR↑ | SIDD [1] SSIM↑ | DND [66] PSNR↑ | DND [66] SSIM↑ | Average PSNR↑ | Average SSIM↑ |
|---|---|---|---|---|---|---|
| DnCNN [110] | 23.66 | 0.583 | 32.43 | 0.790 | 28.04 | 0.686 |
| MLP [5] | 24.71 | 0.641 | 34.23 | 0.833 | 29.47 | 0.737 |
| BM3D [19] | 35.65 | 0.685 | 34.51 | 0.851 | 35.08 | 0.768 |
| CBDNet* [27] | 30.78 | 0.801 | 38.06 | 0.942 | 34.42 | 0.872 |
| RIDNet* [2] | 38.71 | 0.951 | 39.26 | 0.953 | 38.99 | 0.952 |
| AINDNet* [35] | 38.95 | 0.952 | 39.37 | 0.951 | 39.16 | 0.952 |
| VDN [97] | 39.28 | 0.956 | 39.38 | 0.952 | 39.33 | 0.954 |
| SADNet* [10] | 39.46 | 0.957 | 39.59 | 0.952 | 39.53 | 0.955 |
| CycleISP* [99] | 39.52 | 0.957 | 39.56 | **0.956** | 39.54 | 0.957 |
| MIRNet [100] | 39.72 | 0.959 | **39.88** | **0.956** | 39.80 | **0.958** |
| MPRNet [101] | 39.71 | 0.958 | 39.80 | 0.954 | 39.76 | 0.956 |
| MAXIM-3S | **39.96** | **0.960** | 39.84 | 0.954 | **39.90** | 0.957 |

Table 1. Denoising results. Our model is only trained on SIDD [1] and evaluated on SIDD [1] and DND [66], where * denotes methods using additional training data.

images, and $\mathcal{L}_{char}$ is the Charbonnier loss [101]:

$$\mathcal{L}_{char}(\mathbf{R}, \mathbf{T}) = \sqrt{\|\mathbf{R} - \mathbf{T}\|^2 + \epsilon^2}, \quad (8)$$

where we set $\epsilon = 10^{-3}$. $\mathcal{L}_{freq}$ is the frequency reconstruction loss that enforces high-frequency details [18, 33]:

$$\mathcal{L}_{freq}(\mathbf{R}, \mathbf{T}) = \|\mathcal{F}(\mathbf{R}) - \mathcal{F}(\mathbf{T})\|_1 \quad (9)$$

where $\mathcal{F}(\cdot)$ represents the 2D Fast Fourier Transform. We used $\lambda = 0.1$ as the weighting factor in all experiments.

# 4. Experiments

We aim at building a generic backbone for a broad spectrum of image processing tasks. Thus, we evaluated MAXIM on five different tasks: (1) denoising, (2) deblurring, (3) deraining, (4) dehazing, and (5) enhancement (retouching) on **17** different datasets. More comprehensive results and visualizations can be found in Appendix.

## 4.1. Experimental Setup

**Datasets and metrics.** We measured PSNR and SSIM [87] metrics between ground truth and predicted images to make quantitative comparisons. We used SIDD [1] and DND [66] for denoising, GoPro [57], HIDE [74], and RealBlur [72] for debluring, a combined dataset Rain13k used in [101] for deraining. The RESIDE [43] is used for dehazing, while Five-K [6] and LOL [89] are evaluated for enhancement.

**Training details.** Our proposed MAXIM model is end-to-end trainable and requires neither large-scale pretraining nor progressive training. The network is trained on 256×256 random-cropped patches. We train different iterations for each task. We used random horizontal and vertical flips, 90° rotation, and MixUp [102] with probability 0.5 for data augmentation. We used the Adam optimizer [36] with an initial learning rate of $2{\times}10^{-4}$, which are steadily decreased to $10^{-7}$ with the cosine annealing decay [55]. When testing, we padded the input images to be a multiplier of 64×64 using symmetric padding on both sides. After inference, we cropped the padded image back to original size.

| Method | GoPro [57] PSNR↑ | GoPro [57] SSIM↑ | HIDE [74] PSNR↑ | HIDE [74] SSIM↑ | Average PSNR↑ | Average SSIM↑ |
|---|---|---|---|---|---|---|
| DeblurGAN [37] | 28.70 | 0.858 | 24.51 | 0.871 | 26.61 | 0.865 |
| Nah *et al.* [57] | 29.08 | 0.914 | 25.73 | 0.874 | 27.41 | 0.894 |
| Zhang *et al.* [108] | 29.19 | 0.931 | - | - | - | - |
| DeblurGAN-v2 [38] | 29.55 | 0.934 | 26.61 | 0.875 | 28.08 | 0.905 |
| SRN [78] | 30.26 | 0.934 | 28.36 | 0.915 | 29.31 | 0.925 |
| Shen *et al.* [74] | - | - | 28.89 | 0.930 | - | - |
| Gao *et al.* [26] | 30.90 | 0.935 | 29.11 | 0.913 | 30.01 | 0.924 |
| DBGAN [109] | 31.10 | 0.942 | 28.94 | 0.915 | 30.02 | 0.929 |
| MT-RNN [63] | 31.15 | 0.945 | 29.15 | 0.918 | 30.15 | 0.932 |
| DMPHN [103] | 31.20 | 0.940 | 29.09 | 0.924 | 30.15 | 0.932 |
| Suin *et al.* [75] | 31.85 | 0.948 | 29.98 | 0.930 | 30.92 | 0.939 |
| MPRNet [101] | 32.66 | 0.959 | 30.96 | 0.939 | 31.81 | 0.949 |
| Pretrained-IPT [13] | 32.58 | - | - | - | - | - |
| MIMO-UNet+ [18] | 32.45 | 0.957 | 29.99 | 0.930 | 31.22 | 0.944 |
| HINet [14] | 32.71 | 0.959 | 30.32 | 0.932 | 31.52 | 0.946 |
| MAXIM-3S | **32.86** | **0.961** | **32.83** | **0.956** | **32.85** | **0.959** |

Table 2. Deblurring results. Our model is trained on GoPro [57] and evaluated on the GoPro and the HIDE dataset [74].

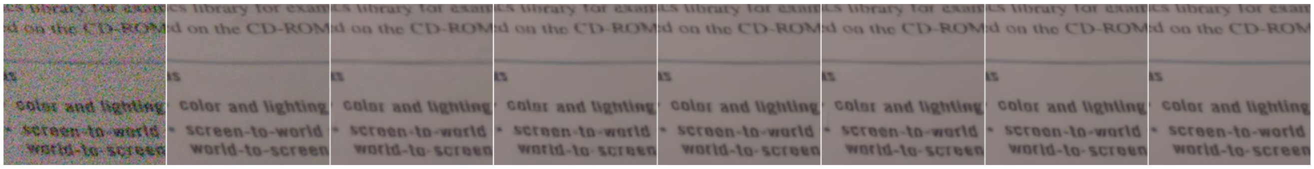
| Method | RealBlur-R [72] PSNR↑ | RealBlur-R [72] SSIM↑ | RealBlur-J [72] PSNR↑ | RealBlur-J [72] SSIM↑ | Average PSNR↑ | Average SSIM↑ |
|---|---|---|---|---|---|---|
| Hu *et al.* [31] | 33.67 | 0.916 | 26.41 | 0.803 | 30.04 | 0.860 |
| Nah *et al.* [57] | 32.51 | 0.841 | 27.87 | 0.827 | 30.19 | 0.834 |
| DeblurGAN [37] | 33.79 | 0.903 | 27.97 | 0.834 | 30.88 | 0.869 |
| Pan *et al.* [62] | 34.01 | 0.916 | 27.22 | 0.790 | 30.62 | 0.853 |
| Xu *et al.* [94] | 34.46 | 0.937 | 27.14 | 0.830 | 30.8 | 0.884 |
| DeblurGAN-v2 [38] | 35.26 | 0.944 | 28.70 | 0.866 | 31.98 | 0.905 |
| Zhang *et al.* [108] | 35.48 | 0.947 | 27.80 | 0.847 | 31.64 | 0.897 |
| SRN [78] | 35.66 | 0.947 | 28.56 | 0.867 | 32.11 | 0.907 |
| DMPHN [103] | 35.70 | 0.948 | 28.42 | 0.860 | 32.06 | 0.904 |
| MPRNet [101] | **35.99** | **0.952** | 28.70 | 0.873 | **32.35** | **0.913** |
| MAXIM-3S | 35.78 | 0.947 | **28.83** | **0.875** | 32.31 | 0.911 |
| †DeblurGAN-v2 | 36.44 | 0.935 | 29.69 | 0.870 | 33.07 | 0.903 |
| †SRN [78] | 38.65 | 0.965 | 31.38 | 0.909 | 35.02 | 0.937 |
| †MPRNet [101] | 39.31 | **0.972** | 31.76 | 0.922 | 35.54 | 0.947 |
| †MIMO-UNet+ [18] | - | - | 32.05 | 0.921 | - | - |
| †MAXIM-3S | **39.45** | 0.962 | **32.84** | **0.935** | **36.15** | **0.949** |

Table 3. Deblurring results on RealBlur [72]. † denotes methods that are trained on RealBlur, while those without † indicate methods trained only on GoPro.

**Architectural configuration.** We designed two MAXIM variants: a two-stage model called MAXIM-2S, and a three-stage model, MAXIM-3S, for different tasks. We start with 32 initial channels for feature extraction, with 3 downsampling layers, where the features contract from $256^2 \times 32$, $128^2 \times 64$, $64^2 \times 128$, to $32^2 \times 256$ processed by *two* Bottlenecks (Fig. 2a), then symmetrically expanded back to full resolution. The number of parameters and required FLOPs of MAXIM-2S and MAXIM-3S, when applied on a $256 \times 256$ image are shown in the last two rows of Tab. 7A.

## 4.2. Main Results

**Denoising.** We report in Tab. 1 numerical comparisons on the SIDD [1] and DND [66] datasets. As may be seen,




Input | Target | VDN [97] | DANet [98] | MIRNet [100] | CycleISP [99] | MPRNet [101] | MAXIM-3S

Figure 4. Denoising comparisons. The example from SIDD [1] shows that our method produces cleaner denoising results.


Input | Target | DMPHN [103] | Suin *et al.* [75] | MPRNet [101] | HINet [14] | MIMO-UNet [18] | MAXIM-3S

Figure 5. Deblurring comparisons. The top row shows an example from GoPro [57] while the second row shows one from HIDE [74].

our method outperformed previous SOTA techniques, *e.g.*, MIRNet [100] by **0.24 dB** of PSNR on SIDD while obtaining competitive PSNR (39.84 dB) on DND. Fig. 4 shows visual results on SIDD. Our method clearly removes real noise while maintaining fine details, yielding visually pleasant results to the other methods.

**Deblurring.** Tab. 2 shows the quantitative comparison of MAXIM-3S against SOTA deblurring methods on two synthetic blur datasets: GoPro [57] and HIDE [74]. Our method achieves **0.15** dB gain in PSNR over the previous best model HINet [14]. It is notable that the GoPro-trained MAXIM-3S model generalizes extremely well on the HIDE dataset, setting new SOTA PSNR values: **32.83** dB. We also evaluated on real-world blurry images from RealBlur [72] under two settings: (1) directly applied the GoPro-trained model on RealBlur, and (2) fine-tuned the model on RealBlur. Under setting (1), MAXIM-3S ranked *first* on RealBlur-J subset while obtaining the top two performance on RealBlur-R. Fig. 5 shows visual comparisons of the evaluated models on GoPro [57], HIDE [74] and RealBlur [72], respectively. It may be observed that our model recovers **text** extremely well, which may be attributed to the use of multi-axis MLP module within each block that globally aggregates repeated patterns across various scales.

**Deraining.** Following previous work [32, 101], we computed the performance metrics using the Y channel (in YCbCr color space). Tab. 4 shows quantitative comparisons with previous methods. As may be seen, our model improved over the SOTA performances on all datasets. The average PSNR gain of our model over the previous best model HINet [14] is **0.24** dB. We demonstrate some challenging examples in Fig. 6, which demonstrates that our method consistently delivered faithfully recovered images without introducing any noticeable visual artifacts.

**Dehazing.** We report our comparisons against SOTA models in Tab. 5. Our model surpassed the previous best model by **0.94** dB and **0.62** dB of PSNR on the SOTS [43] indoor and outdoor sets. Fig. 7 shows that our model recovered images of better quality on both flat regions as well as textures, while achieving a harmonious global tone.

**Enhancement / Retouching.** As Tab. 6 illustrates, our model achieved the best PSNR and SSIM values on FiveK [6] and LOL [89], respectively. As the top row of Fig. 8 suggests, MAXIM recovered diverse naturalistic colors as compared to other techniques. Regarding the bottom example, while MIRNet [100] obtained a higher PSNR, we consistently observed that our model attains visually better quality with sharper details and less noise. Moreover, the far more perceptually relevant SSIM index indicates a significant advantage of MAXIM-2S relative to MIRNet.

**Other benchmarks.** Due to space limitations, we detail the outcomes of our experiments on the REDS deblurring [58] and the Raindrop removal task [67] in Appendix.

### 4.3. Ablation

We conduct extensive ablation studies to validate the proposed multi-axis gated MLP block, cross-gating block, and multi-stage multi-scale architecture. The evaluations were performed on the GoPro dataset [57] trained on image patches of size $256 \times 256$ for $10^6$ iterations. We used the

Input Target RESCAN [46] PreNet [70] MSPFN [32] MPRNet [101] HINet [14] MAXIM-2S

Figure 6. Deraining comparisons. The top and bottom rows present examples from Rain100L [95] and Test100 [105], respectively, demonstrating the ability of MAXIM to remove rain streaks while recovering more details, hence yielding more visually pleasant results.

Input Target GCANet [12] GridDehaze [51] DuRN [52] MSBDN [21] FFA-Net [68] MAXIM-2S

Figure 7. Dehazing comparisons. The top and bottom rows exemplify visual results from the SOTS indoor and outdoor sets [43].

| Method | Rain100L [95] | | Rain100H [95] | | Test100 [105] | | Test1200 [104] | | Test2800 [25] | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| DerainNet [24] | 27.03 | 0.884 | 14.92 | 0.592 | 22.77 | 0.810 | 23.38 | 0.835 | 24.31 | 0.861 | 22.48 | 0.796 |
| SEMI [90] | 25.03 | 0.842 | 16.56 | 0.486 | 22.35 | 0.788 | 26.05 | 0.822 | 24.43 | 0.782 | 22.88 | 0.744 |
| DIDMDN [104] | 25.23 | 0.741 | 17.35 | 0.524 | 22.56 | 0.818 | 29.65 | 0.901 | 28.13 | 0.867 | 24.58 | 0.770 |
| UMRL [96] | 29.18 | 0.923 | 26.01 | 0.832 | 24.41 | 0.829 | 30.55 | 0.910 | 29.97 | 0.905 | 28.02 | 0.880 |
| RESCAN [46] | 29.80 | 0.881 | 26.36 | 0.786 | 25.00 | 0.835 | 30.51 | 0.882 | 31.29 | 0.904 | 28.59 | 0.857 |
| PreNet [70] | 32.44 | 0.950 | 26.77 | 0.858 | 24.81 | 0.851 | 31.36 | 0.911 | 31.75 | 0.916 | 29.42 | 0.897 |
| MSPFN [32] | 32.40 | 0.933 | 28.66 | 0.860 | 27.50 | 0.876 | 32.39 | 0.916 | 32.82 | 0.930 | 30.75 | 0.903 |
| MPRNet [101] | 36.40 | 0.965 | 30.41 | 0.890 | 30.27 | 0.897 | 32.91 | 0.916 | 33.64 | 0.938 | 32.73 | 0.921 |
| HINet [14] | 37.20 | 0.969 | 30.63 | 0.893 | 30.26 | 0.905 | **33.01** | 0.918 | **33.87** | 0.940 | 33.00 | 0.925 |
| MAXIM-2S | **38.06** | **0.977** | **30.81** | **0.903** | **31.17** | **0.922** | 32.37 | **0.922** | 33.80 | **0.943** | **33.24** | **0.933** |

Table 4. Deraining comparisons. Our method consistently yields better quality metrics with respect to both PSNR or SSIM on all the tested datasets: Rain100L [95], Rain100H [95], Test100 [105], Test1200 [104], Test2800 [25]

| Method | SOTS-Indoor | | SOTS-Outdoor | |
|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| DehazeNet [7] | 21.14 | 0.847 | 22.46 | 0.851 |
| GFN [71] | 22.30 | 0 880 | 21.55 | 0.844 |
| GCANet [12] | 30.23 | 0.959 | 19.98 | 0.704 |
| GridDehaze [51] | 32.14 | 0.983 | 30.86 | 0.981 |
| GMAN [54] | 27.93 | 0.896 | 28.47 | 0.944 |
| MSBDN [21] | 33.79 | 0.984 | 23.36 | 0.875 |
| DuRN [52] | 32.12 | 0.980 | 24.47 | 0.839 |
| FFA-Net [68] | 36.39 | 0.989 | 33.57 | 0.984 |
| AECR-Net [92] | 37.17 | 0.990 | - | - |
| MAXIM-2S | **38.11** | **0.991** | **34.19** | **0.985** |

Table 5. Dehazing comparisons. Our model achieved the best results on both indoor and outdoor scenes.

MAXIM-2S model as the test-bed for Ablation-A and -B.

**A. Individual components.** We conducted an ablation by progressively adding (1) inter-stage cross-gating blocks ($\text{CGB}_{\text{IS}}$), (2) a supervised attention module (SAM), (3) cross-stage cross-gating blocks ($\text{CGB}_{\text{CS}}$, and (4) the multi-scale supervision (MS-Sp). Tab. 7A indicates a PSNR gain of 0.25, 0.63, 0.36, 0.26 dB for each respective component.

**B. Effects of multi-axis approach.** We further examined the necessity of our proposed multi-axis approach, as shown in Tab. 7B. We conducted experiments over (1) baseline UNet, (2) by adding the local branch of MAB ($\text{MAB}_\ell$), (3) by adding the global branch of MAB ($\text{MAB}_g$), (4) by adding the local branch of CGB ($\text{CGB}_\ell$), (5) by adding the global branch of CGB ($\text{CGB}_g$). Note that the huge jump (+1.04 dB) of PSNR by adding $\text{MAB}_\ell$ can be largely attributed to the addition of input and output channel projection layers, because we also observe a high performance of **31.42** dB PSNR if only $\text{MAB}_g$ is added. Overall, we observed a *major* improvement when including MAB, and a relatively *minor* gain when adding CGB.

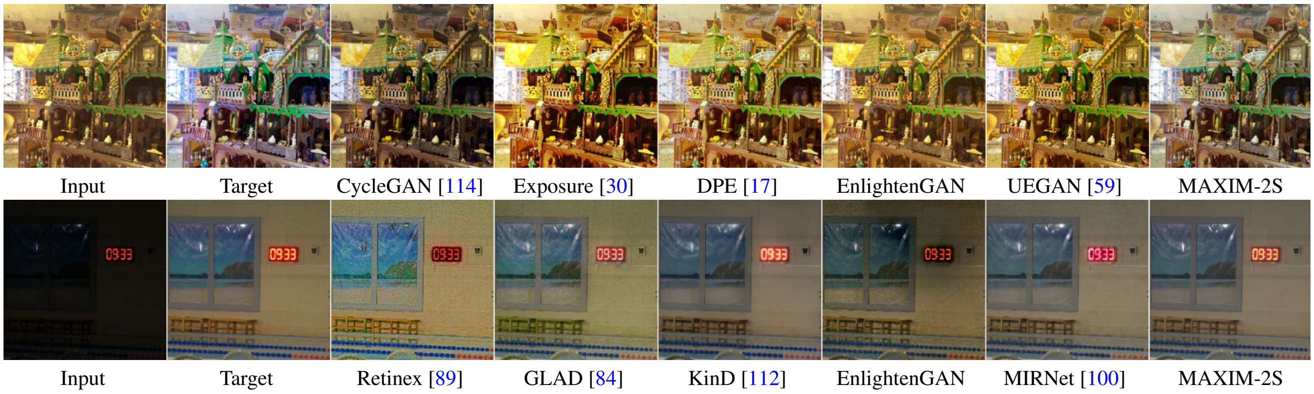



Figure 8. Retouching and low-light enhancement comparisons. The top row shows an example from the MIT-Adobe FiveK dataset [6], while the bottom row exemplifies a comparison from LOL [89]. Our model generated variegated and more naturalistic colors (top) for retouching, while achieving clearer and brighter visual enhancements in the bottom example.

| Method | FiveK [6] PSNR↑ | FiveK [6] SSIM↑ | Method | LOL [89] PSNR↑ | LOL [89] SSIM↑ |
|---|---|---|---|---|---|
| CycleGAN [114] | 18.23 | 0.835 | Retinex [89] | 16.77 | 0.559 |
| Exposure [30] | 22.35 | 0.861 | GLAD [84] | 19.71 | 0.703 |
| EnlightenGAN | 17.74 | 0.828 | EnlightenGAN | 17.48 | 0.657 |
| DPE [17] | 24.08 | 0.922 | KinD [112] | 20.37 | 0.804 |
| UEGAN [59] | 25.00 | 0.929 | MIRNet [100] | **24.14** | 0.830 |
| MAXIM-2S | **26.15** | **0.945** | MAXIM-2S | 23.43 | **0.863** |

Table 6. Enhancement results on FiveK [6] and LOL [89].

| $CGB_{IS}$ | SAM | $CGB_{CS}$ | MS-Sp | PSNR |
|---|---|---|---|---|
| | | | | 30.73 |
| ✓ | | | | 30.98 |
| ✓ | ✓ | | | 31.61 |
| ✓ | ✓ | ✓ | | 31.97 |
| ✓ | ✓ | ✓ | ✓ | 32.23 |

A. Individual components.

| $MAB_\ell$ | $MAB_g$ | $CGB_\ell$ | $CGB_g$ | PSNR |
|---|---|---|---|---|
| | | | | 30.48 |
| ✓ | | | | 31.52 |
| ✓ | ✓ | | | 31.68 |
| ✓ | ✓ | ✓ | | 31.84 |
| ✓ | ✓ | ✓ | ✓ | 31.91 |

B. Effects of multi-axis approach.

| | S | W | D | PSNR | Params | FLOPs |
|---|---|---|---|---|---|---|
| Base | 1 | 32 | 3 | 31.08 | 6.1M | 93.6G |
| Wider | 1 | 64 | 3 | 32.09 | 19.4M | 309.9G |
| | 1 | 96 | 3 | 32.31 | 41.7M | 648.9G |
| Deeper | 1 | 32 | 4 | 31.17 | 19.8M | 121.6G |
| | 1 | 32 | 5 | 31.43 | 75.0M | 153.4G |
| More stages | 2 | 32 | 3 | 31.82 | 14.1M | 216.4G |
| | 3 | 32 | 3 | 32.44 | 22.2M | 339.2G |

C. Why multi-stage?

| Variant | PSNR | Params | FLOPs |
|---|---|---|---|
| M1-FFT | 39.67 | 4.1M | 71G |
| M1-MLP | 39.75 | 5.4M | 83G |
| M1-gMLP | 39.80 | 6.1M | 93G |
| M1-SA | 39.79 | 5.3M | 111G |
| M2-FFT | 39.74 | 10.1M | 172G |
| M2-MLP | 39.70 | 12.7M | 195G |
| M2-gMLP | 39.83 | 14.1M | 216G |
| M2-SA | 39.85 | 12.5M | 250G |

D. Beyond gMLP.

Table 7. Ablation studies. Components in subtable A and B are defined in Sec. 4.3. S, W, and D denote the number of stages, width, and depth, respectively. M1 and M2 in subtable D denote MAXIM-1S and MAXIM-2S models, respectively.

**C. Why multi-stage?** Towards understanding this, we scaled up MAXIM in terms of width (channels), depth (downscaling steps), and the number of stages. Tab. 7C suggests that packing the backbone into multi-stages yields the best performance *vs*. complexity tradeoff (32.44 dB, 22.2 M, 339.2 G), compared to making it wider or deeper.

**D. Beyond gMLP: the MAXIM families.** As described in Sec. 3.2, our proposed multi-axis approach (Fig. 3) offers a scalable way of applying *any* 1D operators on (high-resolution) images, with linear complexity relative to image size while maintaining fully-convolutional. We conducted a pilot study using MAXIM-1S and -2S on SIDD [1] to explore the MAXIM families: MAXIM-FFT, -MLP, -gMLP (modeled in this paper), -SA, where we use the Fourier Transform filter [41, 69], spatial MLP [79], gMLP [50], and self-attention [22] on spatial axes using the same multi-axis approach (Fig. 3). As Tab. 7D shows, the gMLP and self-attention variants achieved the best performance, while the FFT and MLP families were more computationally efficient. We leave deeper explorations to future works.

## 5. Conclusion

We have presented a generic network for restoration or enhancement tasks, dubbed MAXIM, inspired by recently popular MLP-based global models. Our work suggests an effective and efficient approach for applying gMLP to low-level vision tasks to gain global attention, a missing attribute of basic CNNs. Our gMLP instantiation of the MAXIM family significantly advances state-of-the-arts in several image enhancement and restoration tasks with moderate complexity. We demonstrate a few applications, but there are many more possibilities beyond the scope of this work which could significantly benefit by using MAXIM. Our future work includes exploring more efficient models for extremely high-resolution image processing, as well as training large models that can adapt on multiple tasks.

**Broader impacts.** The proposed model can be used as an effective tool to enhance and retouch daily photos. However, enhancing techniques such as denoising and deblurring are vulnerable to malicious use for privacy concerns. The models trained on specific data may express bias. These issues should be responsibly taken care of by researchers.

## References


[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, pages 1692–1700, 2018. 1, 5, 6, 8

[2] Saeed Anwar and Nick Barnes. Real image denoising with feature attention. In *CVPR*, pages 3155–3164, 2019. 5

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 1

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[5] Harold C Burger, Christian J Schuler, and Stefan Harmeling. Image denoising: Can plain neural networks compete with bm3d? In *CVPR*, pages 2392–2399. IEEE, 2012. 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*, pages 97–104. IEEE, 2011. 1, 5, 6, 8

[7] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE TIP*, 25(11):5187–5198, 2016. 7

[8] Jiezhang Cao, Yawei Li, Kai Zhang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 1, 2

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 1

[10] Meng Chang, Qi Li, Huajun Feng, and Zhihai Xu. Spatial-adaptive network for single image denoising. In *ECCV*, pages 171–187. Springer, 2020. 5

[11] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *arXiv preprint arXiv:2103.14899*, 2021. 4

[12] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1375–1383. IEEE, 2019. 7

[13] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, pages 12299–12310, 2021. 1, 2, 3, 5

[14] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *CVPRW*, pages 182–192, 2021. 1, 2, 4, 5, 6, 7

[15] Li-Heng Chen, Christos G Bampis, Zhi Li, Andrey Norkin, and Alan C Bovik. Proxiqa: A proxy approach to perceptual optimization of learned image compression. *IEEE TIP*, 30:360–373, 2020. 1, 2

[16] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, pages 7103–7112, 2018. 2, 4

[17] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *CVPR*, pages 6306–6314, 2018. 8

[18] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *ICCV*, pages 4641–4650, 2021. 1, 2, 3, 4, 5, 6

[19] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE TIP*, 16(8):2080–2095, 2007. 5

[20] Mauricio Delbracio, Hossein Talebi, and Peyman Milanfar. Projected distribution loss for image enhancement. *ICCP*, 2021. 1

[21] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *CVPR*, pages 2157–2167, 2020. 2, 7

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 2, 3, 8

[23] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *CVPR*, pages 3575–3584, 2019. 2

[24] Xueyang Fu, Jiabin Huang, Xinghao Ding, Yinghao Liao, and John Paisley. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE TIP*, 26(6):2944–2956, 2017. 7

[25] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, pages 3855–3863, 2017. 7

[26] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *CVPR*, pages 3848–3856, 2019. 5

[27] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *CVPR*, pages 1712–1722, 2019. 5

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

[29] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[30] Yuanming Hu, Hao He, Chenxi Xu, Baoyuan Wang, and Stephen Lin. Exposure: A white-box photo post-processing framework. *ACM TOG*, 37(2):1–17, 2018. 8

[31] Zhe Hu, Sunghyun Cho, Jue Wang, and Ming-Hsuan Yang. Deblurring low-light images with light streaks. In *CVPR*, pages 3382–3389, 2014. 5

[32] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *CVPR*, pages 8346–8355, 2020. 1, 2, 6, 7

[33] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *ICCV*, pages 13919–13929, 2021. 5

[34] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE TIP*, 30:2340–2349, 2021. 1, 2

[35] Yoonsik Kim, Jae Woong Soh, Gu Yong Park, and Nam Ik Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *CVPR*, pages 3482–3492, 2020. 5

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[37] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, pages 8183–8192, 2018. 5

[38] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, pages 8878–8887, 2019. 1, 2, 5

[39] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, pages 624–632, 2017. 1

[40] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 1, 2

[41] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*, 2021. 1, 2, 3, 8

[42] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *ICCV*, pages 4770–4778, 2017. 2

[43] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE TIP*, 28(1):492–505, 2019. 1, 5, 6, 7

[44] Shi-Jie Li, Yazan AbuFarha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[45] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019. 2, 4

[46] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, pages 254–269, 2018. 7

[47] Yinxiao Li, Pengchong Jin, Feng Yang, Ce Liu, Ming-Hsuan Yang, and Peyman Milanfar. Comisr: Compression-informed video super-resolution. In *ICCV*, 2021. 1, 2

[48] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *IEEE Int. Conf. Comput. Vis. Worksh.*, 2021. 1, 2

[49] Ding Liu, Bihan Wen, Yuchen Fan, Chen Change Loy, and Thomas S Huang. Non-local recurrent network for image restoration. *arXiv preprint arXiv:1806.02919*, 2018. 2

[50] Hanxiao Liu, Zihang Dai, David R So, and Quoc V Le. Pay attention to mlps. *arXiv preprint arXiv:2105.08050*, 2021. 1, 2, 3, 4, 8

[51] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *ICCV*, pages 7314–7323, 2019. 7

[52] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *CVPR*, pages 7007–7016, 2019. 7

[53] 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. *ICCV*, 2021. 1

[54] Zheng Liu, Botao Xiao, Muhammad Alrabeiah, Keyan Wang, and Jun Chen. Single image dehazing with a generic model-agnostic convolutional neural network. *IEEE Signal Processing Letters*, 26(6):833–837, 2019. 7

[55] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[56] Erik Matlin and Peyman Milanfar. Removal of haze and noise from a single image. In *Computational Imaging X*, volume 8296, page 82960T. International Society for Optics and Photonics, 2012. 1

[57] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, pages 3883–3891, 2017. 1, 5, 6

[58] Seungjun Nah, Sanghyun Son, Suyoung Lee, Radu Timofte, and Kyoung Mu Lee. Ntire 2021 challenge on image deblurring. In *CVPR Workshops*, pages 149–165, June 2021. 6

[59] Zhangkai Ni, Wenhan Yang, Shiqi Wang, Lin Ma, and Sam Kwong. Towards unsupervised deep image enhancement with generative adversarial network. *IEEE TIP*, 29:9140–9151, 2020. 8

[60] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *ECCV*, pages 191–207. Springer, 2020. 1, 2

[61] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 3, 4

[62] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *CVPR*, pages 1628–1636, 2016. 5

[63] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *ECCV*, pages 327–343. Springer, 2020. 4, 5

[64] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, volume 32, 2018. 4

[65] Olivier Petit, Nicolas Thome, Clement Rambour, Loic Themyr, Toby Collins, and Luc Soler. U-net transformer: self and cross attention for medical image segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 267–276. Springer, 2021. 3, 4

[66] Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. In *CVPR*, pages 1586–1595, 2017. 5

[67] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *CVPR*, pages 2482–2491, 2018. 6

[68] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI*, volume 34, pages 11908–11915, 2020. 7

[69] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *arXiv preprint arXiv:2107.00645*, 2021. 1, 2, 3, 8

[70] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, pages 3937–3946, 2019. 4, 7

[71] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *CVPR*, 2018. 7

[72] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, pages 184–201. Springer, 2020. 5, 6

[73] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[74] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *ICCV*, pages 5572–5581, 2019. 2, 5, 6

[75] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*, pages 3606–3615, 2020. 4, 5, 6

[76] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, pages 5693–5703, 2019. 3

[77] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple-object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 1

[78] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, pages 8174–8182, 2018. 5

[79] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021. 1, 2, 3, 8

[80] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *arXiv preprint arXiv:2105.03404*, 2021. 2

[81] Ashish Vaswani, Prajit Ramachandran, Aravind Srinivas, Niki Parmar, Blake Hechtman, and Jonathon Shlens. Scaling local self-attention for parameter efficient visual backbones. In *CVPR*, pages 12894–12904, 2021. 1

[82] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2, 4

[83] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *CVPR*, pages 12270–12279, 2019. 1, 2

[84] Wenjing Wang, Chen Wei, Wenhan Yang, and Jiaying Liu. Gladnet: Low-light enhancement network with global awareness. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 751–755. IEEE, 2018. 8

[85] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 2

[86] 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 *IEEE Eur. Conf. Comput. Vis. Worksh.*, pages 0–0, 2018. 1, 2

[87] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 5

[88] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021. 2

[89] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 5, 6, 8

[90] Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, and Ying Wu. Semi-supervised transfer learning for image rain removal. In *CVPR*, pages 3877–3886, 2019. 7

[91] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 3

[92] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *CVPR*, pages 10551–10560, 2021. 7

[93] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *arXiv preprint arXiv:2105.15203*, 2021. 1

[94] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural l0 sparse representation for natural image deblurring. In *CVPR*, pages 1107–1114, 2013. 5

[95] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *CVPR*, pages 1357–1366, 2017. 1, 7

[96] Rajeev Yasarla and Vishal M Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *CVPR*, pages 8405–8414, 2019. 7

[97] Zongsheng Yue, Hongwei Yong, Qian Zhao, Lei Zhang, and Deyu Meng. Variational denoising network: Toward blind noise modeling and removal. *arXiv preprint arXiv:1908.11314*, 2019. 5, 6

[98] Zongsheng Yue, Qian Zhao, Lei Zhang, and Deyu Meng. Dual adversarial network: Toward real-world noise removal and noise generation. In *ECCV*, pages 41–58. Springer, 2020. 6

[99] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Cycleisp: Real image restoration via improved data synthesis. In *CVPR*, pages 2696–2705, 2020. 5, 6

[100] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *ECCV*, pages 492–511. Springer, 2020. 1, 2, 3, 5, 6, 8

[101] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, pages 14821–14831, 2021. 1, 2, 3, 4, 5, 6, 7

[102] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 5

[103] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, pages 5978–5986, 2019. 1, 2, 4, 5, 6

[104] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *CVPR*, pages 695–704, 2018. 7

[105] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *IEEE TCSVT*, 30(11):3943–3956, 2019. 7

[106] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, pages 5907–5915, 2017. 2

[107] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE TPAMI*, 41(8):1947–1962, 2018. 2

[108] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson WH Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *CVPR*, pages 2521–2529, 2018. 5

[109] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *CVPR*, pages 2737–2746, 2020. 5

[110] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE TIP*, 26(7):3142–3155, 2017. 1, 2, 5

[111] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, pages 2472–2481, 2018. 1, 2

[112] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *ACM MM*, pages 1632–1640, 2019. 8

[113] Long Zhao, Zizhao Zhang, Ting Chen, Dimitris N Metaxas, and Han Zhang. Improved transformer for high-resolution gans. *arXiv preprint arXiv:2106.07631*, 2021. 3

[114] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017. 8

[115] Xiang Zhu, Filip Šroubek, and Peyman Milanfar. Deconvolving psfs for a better motion deblurring using multiple images. In *European Conference on Computer Vision*, pages 636–647. Springer, 2012. 1