# MG-MLP: Multi-gated MLP for Restoring Images from Spatially Variant Degradations - Supplementary Material

Jaihyun Koh[1]
julian.koh@samsung.com
Jaihyun Lew[2]
fudojhl@snu.ac.kr
Jangho Lee[3]
ubuntu@inu.ac.kr
Sungroh Yoon[*,2,4]
sryoon@snu.ac.kr

[1] Display Research Center
Samsung Display Corporation
Yongin, South Korea

[2] Interdisciplinary Program in Artificial Intelligence
Seoul National University
Seoul, South Korea

[3] Department of Computer Science and Engineering
Incheon National University
Incheon, South Korea

[4] Department of Electrical and Computer Engineering
Seoul National University
Seoul, South Korea

* Corresponding Author

## 1 Code submission

For reproducibility, we provide python files implemented using PyTorch 1.11.0 and Python 3.9. This implementation is based on BasicSR and the skeleton and macro architecture are borrowed from NAFnet.

- `basicsr/models/archs/mg_mlp_arch.py` contains the architecture of the proposed MG-MLP
- `basicsr/models/archs/NAFNet.py` includes the architecture of the NAFnet [2].
- `basicsr/models/archs/Restormer_arch.py` contains Restormer's networks block [10].
- `options/train/{dataname}/*.yml` contains the learning hyperparameters, network architecture, data location, etc for training.
- `options/test/{dataname}/*.yml` contains the network architecture and data location for a test.

© 2023. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

- `train.sh` includes a command to train.
- `test.sh` includes a command to test.

We trained three restoration models with a unified framework to determine the performance gains from the proposed architecture of the network block. We equalized all hyperparameters used in training the three networks, including macro architecture and learning-rate schedules, which are the same as those of NAFnet. In addition, we set the batch size to 8 to train using a single Nvidia V100 GPU with feeding 256 image patches. The trained weights of the three networks on the GoPro dataset can be downloaded at this link.

# 2 Gating Mechanism

Our gating can learn the multi-modal Gaussian mixture distribution of complex clean images [3]. The existing single-path residual block optimized with the mean squared error (MSE) induces the model to learn the uni-modal Gaussian distribution. By contrast, the proposed multi-path structure provides the ability to learn bi-modal distributions. Intuitively, the patterns in an image comprise a combination of simple patches. Assuming that a single simple patch can be represented by a uni-mode, the combination of several uni-modal distributions can express complex patterns. Empirical results revealed that each gating path handles different degradation levels in an image. This result revealed that the proposed MG-MLP block is effective for spatially varying degradation.

## 2.1 Learning multi-modal Gaussian mixture distribution

A feedforward neural network, such as a ResNet block, approximates the deterministic function $f(y)$. This function, which is learned by empirically minimizing the MSE loss, is a model of a uni-modal Gaussian distribution $p(x|y)$ [3]. Therefore, modeling the complex multi-modal distribution of clean images is difficult. In particular, image-restoration tasks solving an ill-posed inverse problem may have several solutions for the same input $y$, and searching for it with such a uni-modal function is not trivial. Thus, adaptiveness to the input images is critical for obtaining an underdetermined solution $x$. Although the nonlinear approximation capability of deep neural networks can overcome this problem, a structured solution is required. A gating mechanism is an option for addressing this problem [3, 4]. Considering a network that takes an input $y$ and predicts latent $x$, it is a function that models the $p(x|y)$ distribution. If the actual distribution of $x$ is a multi-modal distribution, the conditional mixture model can be expressed as follows:

$$p(x|y) = \sum_g p(y,g|x) = \sum_g p(g|x)p(y|x,g) \tag{1}$$

where $g$ is a gating unit, $p(y|x,g)$ is the probability distribution of the output $y$ predicted by the model, and $p(g|x)$ is the probability distribution of the weights associated with $x$-adaptive gating. In a previous study [3], the multi-modal probability distribution was modeled by multiplying two input projections, each having uni-modal Gaussian distributions, and stacking these operations in several layers. Feed-forward NN blocks, including gatedConv [4] and gMLP [9], have been proposed using this gating mechanism, such that the multiplication of two different projections of the input introduces second-order interaction [9].

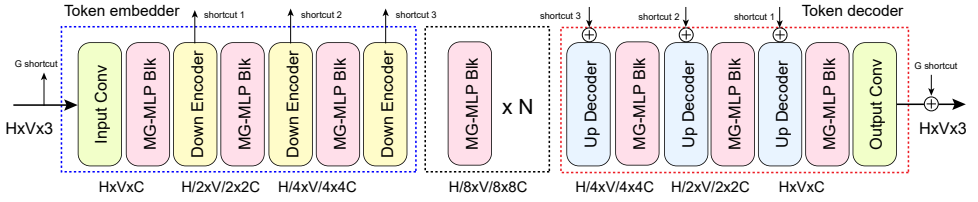


Figure 1: An illustration of a simple Unet structure (macro architecture) for image restoration.

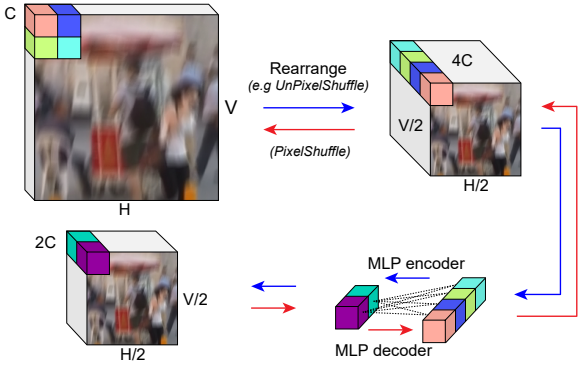


Figure 2: Illustration of token-embedding (blue arrow) and decoding (red arrow) methods.

## 2.2 Analysis on MG-MLP

A network trained using the proposed gating methodology can learn a Gaussian mixture distribution with two modes. Assuming $z$ is a random variable belonging to $Z$, the two unimodal distributions, each having a distinct mode, can be modeled as follows:

$$\begin{aligned} p(z_t|z_{t-1}) &= p(z_t, g_1|z_{t-1}) + p(z_t, g_2|z_{t-1}) \\ &= p(g_1|z_{t-1})p(z_t|z_{t-1}, g_1) \\ &+ p(g_2|z_{t-1})p(z_t|z_{t-1}, g_2). \end{aligned} \tag{2}$$

This result is similar to the analysis conducted in a previous study [3]. However, in the proposed MG-MLP, the unit NN block can instantly learn the multi-modal Gaussian mixture distribution. By deep stacking of these blocks, the model can learn more complex distributions in clean images.

# 3 Macro Architecture

The CNN block-based image restoration architecture generally uses a multi-scale strategy to extend the receptive field of the models. These methods downscale the degraded image with several scale factors, and the resulting images with different sizes are input to multiple network branches. The models are then optimized so that the output image from each branch becomes the same as the downscaled ground truth image. This multi-scale approach

is effective in receptive fields but could not be efficient in terms of memory and computational operation. Many researchers have explored how to efficiently aggregate features from multiple resolution stages without multiple branches. These efforts eventually resulted in complex connections and messy structures. This complexity may come from the inability of the CNN-based blocks to perform adequate feature aggregation due to the spatially invariant biases. Now, it is worth considering whether multi-scale methodologies are necessary for MLP mixer and ViT-based networks where unit NN blocks have sufficient receptive fields and do not have a bias that comes from network structure.

The proposed network blocks are integrated into the Unet as a macro architecture. The feature encoding procedure in the restoration model reduces the feature scale in the spatial direction corresponding to resolution while extending it in the channel direction. Therefore, the embedded token includes the representations of multiple pixels in an adjacent area. After repeating this encoding process, information in a wide area is hierarchically embedded into a token vector. As the number of downscales increases, one token contains more pixel information, so the feature map of the bottleneck layer can have a sufficiently large receptive field. For this reason, the recently proposed MLP mixer-based network architectures more focus on modeling the relation between surrounding tokens rather than developing the multi-scale strategy that widens the receptive field. Transformer and MLP mixer-based networks, which use the Unet structure instead of a complex multi-resolution design, exhibit competitive results. Thus we also do not use multi-scale or stage methods. Instead, we employ the simple Unet structure in which we use a method of encoding spatial to channel by adopting a pixel rearrangement and MLP, as shown in Figure 2. The detailed Unet used in our experiments is shown in Figure 1.

In the ViT or MLP-like architecture, if a single token has an adequate receptive field and interacting tokens are connected densely with free weight, communicating only with adjacent tokens can be acceptable. We empirically found that broad interaction does not lead to further performance improvements. Comparing nine-tokens aggregation implemented by 3x3 depth-wise convolution with more tokens interaction, such as 25 and 49, as increasing aggregation filter size, the performance is degraded due to the error caused by larger padding.

# 4 Additional Results

## 4.1 Evaluation on DPDD Dataset

**Defocus Deblurring** We evaluated the models on the DPDD dataset [1] to confirm the performance of spatially variant defocus deblurring. The quantitative results are presented in Table 4.2. The MG-MLP outperformed others in terms of SSIM, LPIPS [11], and DISTS [5], which evaluate the perceptual quality. This result revealed that the proposed gating mechanism effectively reconstructs the structural information of an object. The qualitative comparison in Figure 3 supports this argument.

## 4.2 Additional Results

Figure 4 - 7 shows the addition qualitative comparisons. All figures are best viewed in zoom.

| GoPro-trained | Restormer [10] | | NAFNet [2] | | MG-MLP (ours) | |
|---|---|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| GoPro-test | 32.07 | 0.9513 | 32.23 | 0.9543 | **32.87** | **0.9604** |

Table 1: Average PSNR and SSIM on GoPro testset [6]: two baselines [2, 10] are trained using macro architecture and learning hyper-parameters proposed in the original paper, whereas the proposed MG-MLP is trained with the unified framework.

| DPDD-trained | DPDD-test [1] | | | |
|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | LPIPS ↓ | DISTS ↓ |
| Restormer [10] | 25.72 | 0.8561 | 0.2055 | 0.1322 |
| NAFNet [2] | **25.77** | 0.8589 | 0.2145 | 0.1326 |
| MG-MLP (ours) | 25.50 | **0.8592** | **0.1889** | **0.1226** |

Table 2: Average PSNR, SSIM, LPIPS [11], and DISTS [5] for the images in the spatially variant defocus deblurring dataset (DPDD) [1]. Up-arrows denote that higher values are better, and down-arrow indicate the opposite.

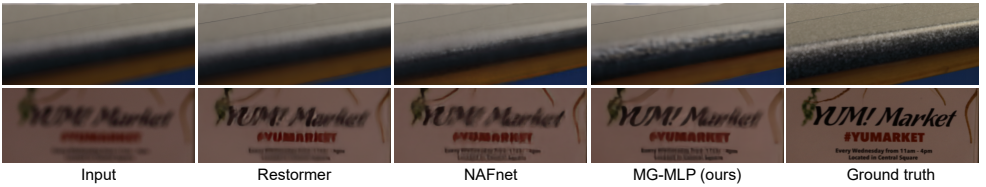


Figure 3: Two example images from the DPDD dataset [1] restored by three different networks.


Figure 4: Additional example images from the GoPro dataset [6] restored by three different networks. The first column contains the degraded input images. The next three columns show the reconstructed images obtained using Restormer [10], NAFnet [2], and our MG-MLP. The final column contains the ground truth images.


Figure 5: Additional example images from the RealBlur-J dataset [7] restored by three different networks.


Figure 6: Additional example images from the DVD dataset [8] restored by three different networks.

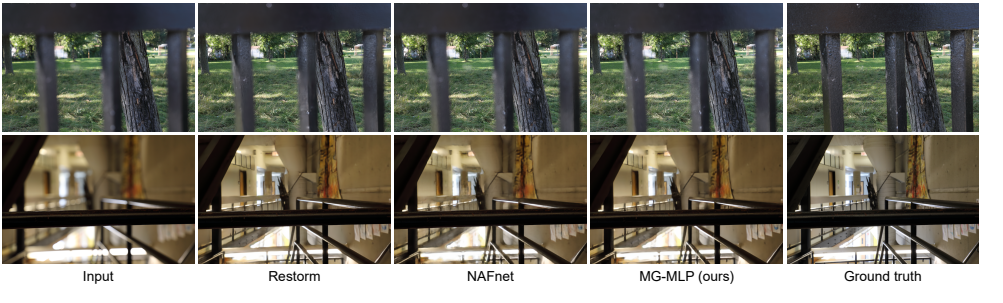


Figure 7: Additional example images from the DPDD dataset [1] restored by three different networks.

# References


[1] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, pages 111–126. Springer, 2020.

[2] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*, 2022.

[3] Yann N Dauphin and David Grangier. Predicting distributions with linearizing belief networks. *arXiv preprint arXiv:1511.05622*, 2015.

[4] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

[5] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[6] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017.

[7] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European Conference on Computer Vision*, pages 184–201. Springer, 2020.

[8] Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1279–1288, 2017.

[9] Asher Trockman and J Zico Kolter. Patches are all you need? *arXiv preprint arXiv:2201.09792*, 2022.

[10] 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.

[11] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.