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

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

Abstract

We propose a novel gating mechanism, which can be applied to the MLP mixer-based architecture for image restoration. In the proposed architecture, embedded tokens are subjected to channel and token mixing, which are the primary data flow of the existing MLP mixer. The token vectors are subsequently refined through the proposed intra-token and cross-token gating. Intra-token gating determines the information that is to be propagated or discarded by the interaction of information within each token. By contrast, cross-token gating calculates the propagation weights of local information and recycles information discarded from intra-token gating by comparing the information with adjacent tokens. The two gating paths result in third-order interaction because of cascaded gating multiplication, which is similar to the self-attention of Transformer. However, the proposed method is more efficient than Transformer because it does not involve the quadratic cost of self-attention. The proposed network was applied to various spatially variant deblurring tasks; it outperformed baselines in terms of restoration performance and computational cost.

1 Introduction

Neural network (NN) blocks based on convolutional NNs (CNNs), including ResNet [16] and DenseNet [19], have been widely adopted in image restoration and low-level computer

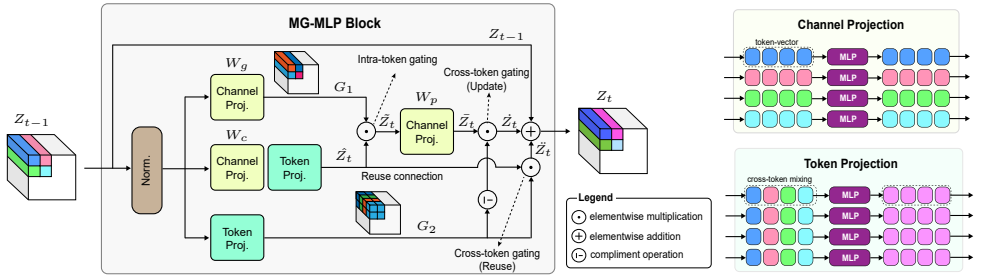

Figure 1: An architecture of the proposed MG-MLP block is presented. The block consists of intra-token gating and cross-token gating. Intra-token gating controls the flow of information through the interaction of data in each token; cross-token gating updates the resulting token from intra-token gating and simultaneously brings back the data discarded by intra-token gating by referring to the adjacent tokens.

vision fields [16, 19, 36, 37, 52, 55]. Because network architectures inheriting that of Transformer [46] and MLP mixer [41] achieved excellent performance in high-level computer vision tasks, such as classification [4, 12, 43], object detection [2, 49, 57, 60], and semantic segmentation [54, 58], their use has been expanded to low-level computer vision [3, 45]. Multi-head self-attention (MHSA) of Vision Transformer (ViT) [12] and MLP mixer's token mixing, which models the relationships between image tokens, exhibit a weaker inductive bias than CNNs. However, despite the high performance of these models, their excessive computational costs have become a critical obstacle in adapting them to low-level vision. Restormer [53] and MAXIM [45] have been proposed to address this inefficiency in image restoration. Restormer uses channel-wise self-attention to reduce the quadratic cost of self-attention, and MAXIM presents two parallel data paths with distinct receptive fields using an efficient permutation technique, as in various token mixing methods [38, 40, 48, 57]. The most recent NAFnet [6] uses simple gating without nonlinear activation and simplifies channel attention. From the perspective of the MLP mixer, this method is a form of gated MLP (gMLP) [25].

Previous MLP-like methods rely on three components that considerably influence performance: (1) a channel mixer [41], which is implemented using 1×1 convolution or MLP along the channel direction, refines the tokens to have task-relevant features in latent space; (2) the token mixer, in which depth-wise convolution or feature permutation followed by MLP, enables information to be exchanged between tokens [7, 14, 18, 24, 39, 50]; (3) gating units propagate only valuable information to efficiently use the capacity of NNs. [11, 25, 34].

This study focuses on the effectiveness of the gating mechanism. Existing gating methods, expressed as the product of two projections of input features, are used for information flow control. Inspired by the multiple gatings in GRU [9] and LSTM [51], we propose a novel MLP mixer-based NN block, namely, *MG-MLP*, that has two gating units. In this block, update and reuse gating are added to propagation gating. The proposed gating configuration comprises a cascade of two types of gating, namely intra-token gating and cross-token gating. Intra-token gating is a soft weighting of the primary token pipeline to control information flow. In cross-token gating, the attention score based on the correlation between adjacent tokens is calculated. The added "reuse connection" in cross-token gating can restore the information discarded by intra-token gating: the masked features, by referring to the context in a single token, can be recycled reconsidering this with the adjacent tokens.

The proposed gating mechanism can result in third-order interactions, such as Transformer, whereas vanilla gating relies on second-order interactions [25].

The contributions of the study can be summarized as follows:

- A novel NN architecture block utilizing a multi-gating mechanism based on an MLP mixer for image restoration is proposed; this block improves the nonlinear interaction capability [25].
- The two proposed gatings control the complementary information flows through intra-token and cross-token interactions, and the information masked by intra-token gating but necessary to reconstruct the object's edge can be recycled based on the correlation between tokens in cross-token gating.
- The superiority of the proposed method over existing state-of-the-art blocks is demonstrated through various comparative experiments, including four motion deblurring datasets. In particular, the results reveal that the proposed architecture can effectively restore images with spatially variant degradation.

2 Related Work

2.1 Vision Transformer and MLP Mixer

ViT has attracted considerable research attention in image classification and segmentation tasks [15], and studies have focused on the development of MLP mixer-like architecture [26] to avoid the strong local inductive bias of CNNs. ViT and MLP mixers divide an image into patches called "tokens," and subsequently refine the token and model the relation between them. To understand the relationship between tokens, MHSA is used in ViT, which includes the inner product of the token vectors as its primary operation. The vanilla MLP mixer combines the elements of a specific order of tokens using fully connected weighting. Representative variants based on ViT are as follows: DeiT [42], in which basic attention and knowledge distillation for dataset efficiency are used; Swin Transformer [27] in which hierarchical patch attention is used for computational efficiency; and CMT [13], which is a hybrid network combining the advantages of local information collection of CNNs and long-range dependency modeling of ViT. In MLP mixer-based architectures, token-mixing methods have been developed for minimizing the number of parameters and floating-point operations while gathering information from as many tokens as possible. ConvMixer [44] implements the MLP mixing mechanism using CNNs, and ResMLP [43] uses two shortcut connections and introduced affine normalization. In ViP [18], tokens in multiple directions are mixed by swapping the feature axis, and S2MLP [50] widens the token receptive fields through a feature map shift.

2.2 Image Restoration

Unlike high-level vision, low-level vision tasks, such as deblurring [21] and super resolution [47], require the following properties: modeling the low-level semantic is critical, that is, the relationship between pixels rather than the abstract of the image is considered; a sufficient receptive field is essential to search for similar patterns in the own-image; the NN architecture used for image restoration should be "fully convolutional" [45] because the resolution of the input image is not fixed in most cases.

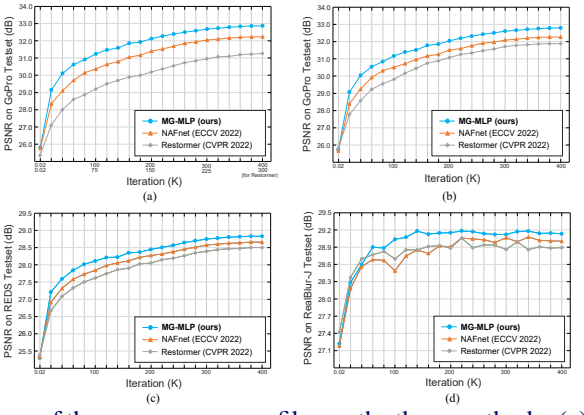

Figure 2: Comparison of the convergence profiles on the three methods: (a) average PSNR on GoPro testset [30], two baselines (NAFNet [6], Restormer [53]) are trained using the macro architecture and learning hyperparameters proposed in the original paper, and average PSNR on (b) GoPro [30], (c) REDS [31], and (d) RealBlur-J [32] test sets, which are all trained using the unified framework.

CNN-based Blocks Most CNN-based restoration networks focus on hierarchical architecture with multi-scale images or feature pyramids to widen the receptive field of the model [8]. Several methods have been proposed for novel connection, feature aggregation, and propagation to effectively transfer task-relevant features [8, 52, 59]. The residual dense network [56], which directly transfers shallow features to the deep level, has been widely used in various restoration tasks. However, the high memory usage of this method should be addressed. Parallel dilated convolution [5, 22] effectively obtains a wider receptive field by using multiple dilation factors in the same level features; it is also efficient because of the use of depth-wise convolution.

ViT and MLP Mixer-based Blocks To improve spatial inductive bias of CNNs and limited receptive field, architectures such as ViT and MLP mixer have been introduced in image restoration. Restormer [53] introduces the channel-wise self-attention mechanism that improves the quadratic computational complexity of MHSA. This mechanism has an invariant architecture to the resolution of an input image; MAXIM [45] updates the tokens by connecting two branches with various receptive fields by using different feature permutations and a gating mechanism; NAFnet [6] implements the gating operation using element-wise multiplication without nonlinear activation. Channel mixing and token mixing are critical components of these methods. Another component, the gating unit, was invented in the natural language process and developed for vision tasks. gMLP [44] reveals that spatial projection followed by gating can effectively improve the performance of the MLP mixer. NAFnet [6] proposed a simple gating method for low-level vision. The gating unit is simple to implement but can propagate important features well in the forward direction.

3 Method

The translation-invariant biases in CNN blocks may be suitable for high-level vision tasks that extract abstract or high-level spatial features in the images. However, these may not be appropriate for low-level vision tasks that model the relation of adjacent pixels or patches

because the features extracted by the deep CNN blocks tend to include abstract, pattern, and texture rather than fine-grained pixel arrangements [17, 23]. By contrast, the MLP mixer-based block, which has fewer inductive biases, can be suitable for image restoration, which requires understanding low-level features than abstracts and textures. However, the vanilla MLP mixer cannot be applied to image restoration tasks because it is not fully convolutional, the number of free parameters varies depending on the resolution of the input image, and it is computationally expensive because of the token-wise fully connected scheme. These two problems can be addressed by replacing the token-mixing MLP with depth-wise convolution.

3.1 Proposed Gated MLP Block

The multiple layers of gating blocks in existing methods [10, 11, 24] do not include the marginal for the multiple gating component. Taking this into account, we propose a novel network block, namely MG-MLP, with two gating components. In this method, multiple gating allows the block to have an opportunity to model multi-modal distributions (see Supplementary Material). Figure 1 reveals the proposed architecture of the MG-MLP. A network block that receives a set of intermediate tokens $Z_{t-1} \in \mathbb{R}^{h \times v \times c}$ and outputs $Z_t \in \mathbb{R}^{h \times v \times c}$ is considered. Here, h and v are the horizontal and vertical lengths of the token arrays, respectively, and c is the dimension of each token vector. The primary pipeline of the proposed MG-MLP is a series of two projections, namely channel mixing and token-mixing behind normalization, similar to the existing MLP mixer, and this can be formally expressed as follows:

$$\hat{Z}_t = \mathrm{TokenMix}(W_c(\mathrm{Norm}(Z_{t-1}))), \tag{1}$$

where $\hat{Z}_t$ is the resulting token from the primary pipeline, $W_c \in \mathbb{R}^{c \times c}$ is a learnable matrix for channel-mixing: $\mathbb{R}^c \to \mathbb{R}^c$, and $\mathrm{TokenMix}(\cdot)$ is the operation of token-mixing in which c kernels with $k \times k$ size aggregate the k^2 tokens in the same element group. Two gating operations, namely intra-token gating and cross-token gating, are added to the primary pipeline. In intra-token gating, the token resulting from the primary pipeline $\hat{Z}_t$ is multiplied by the projected input token G_1 by $W_g \in \mathbb{R}^{c \times c}$ along the channel direction as follows:

$$\begin{aligned} G_1 &= W_g(\mathrm{Norm}(Z_{t-1})), \text{and} \\ \tilde{Z}_t &= \hat{Z}_t \odot G_1, \end{aligned} \tag{2}$$

where $\tilde{Z}_t$ is the resulting token from the intra-token gating and $\odot$ is the element-wise multiplication. In this procedure, only the important information in the token is propagated by referring to the values of the intra-token. The rest is discarded. Next, another channel mixing matrix $W_p \in \mathbb{R}^{c \times c}$ rearranges and refines the gated token vector as follows:

$$\bar{Z}_t = W_p(\tilde{Z}_t). \tag{3}$$

Cross-token gating determines the values to be forgetten and values to be delivered through an information exchange between the surrounding tokens. The operation for obtaining the cross-gating weight G_2 can be expressed as follows:

$$G_2 = \mathrm{TokenMix}(\mathrm{Norm}(Z_{t-1})). \tag{4}$$

The cross-token gating weight G_2 partially extracts the reusable features from the token in the primary pipeline $\hat{Z}_t$. Simultaneously, the forgetting weight, that is, the complement of

G_2, is multiplied by the token from the intra-token gating $\bar{Z}_t$ to softly erase irrelevant features in $\bar{Z}_t$. The operation of cross-token gating can be expressed as follows:

$$\ddot{Z}_t = \hat{Z}_t \odot G_2 \text{ (reuse gating)}, \tag{5}$$

and

$$\dot{Z}_t = \bar{Z}_t \odot (1 - G_2) \text{ (update gating)}, \tag{6}$$

where $\ddot{Z}_t$ is the recycled token from the discarded feature by intra-token gating, and $\dot{Z}_t$ is the updated token by cross-token gating. This operation returns the values that can positively affect adjacent tokens among the values discarded by intra-token gating. This result is consistent with the multi-gating strategies of the LSTM and GRU. However, this method is unique such that two gating units estimate each gating weight differently based on the interaction between values within the token or cross-tokens.

To improve the computational complexity of the original token mixing and MHSA and to support any number of tokens, the proposed token-mixing method connects tokens within a fixed range, implemented by depth-wise convolution. Finally, the resulting two gated outputs are connected to the input token with a the residual connection as follows:

$$Z_t = Z_{t-1} \oplus \dot{Z}_t \oplus \ddot{Z}_t, \tag{7}$$

where $\oplus$ expresses element-wise addition.

The proposed network blocks are integrated into the Unet as a macro architecture. The token embedding procedure reduces the feature scale in the spatial direction corresponding to the resolution and extends it in the channel direction (see Supplementary Material). Therefore, the embedded token includes representations of multiple pixels in adjacent areas. After repeating this encoding process, the information in a wide area is hierarchically embedded into a token vector. Therefore, recently proposed MLP mixer-based network architectures emphasize modeling the influence between surrounding tokens [7, 39] rather than developing a multi-scale strategy that widens the receptive field. Networks based on the transformer and MLP mixers in which the Unet structure is used instead of a complex multi-resolution design, exhibit competitive results [6, 53].

In our macro architecture, the input degradation image $Y \in \mathbb{R}^{H \times V \times 3}$ is first expanded to the feature dimension C through shallow convolution. The expanded features are then refined with MG-MLP and downscaled through rearrangement followed by encoding (see Supplementary Material). This process is performed three times to obtain a bottleneck feature map $Z_0 \in \mathbb{R}^{\frac{H}{2^3} \times \frac{V}{2^3} \times 2^3 C}$, as

$$Z_0 = \text{TokenEmb}^3(\text{Conv}(Y)), \tag{8}$$

where H and V are the horizontal and vertical resolutions of the degraded input image Y, respectively. Here, Z_0 passes through N MG-MLPs and becomes a refined features Z_N. After performing three inverse operations of the encoder, that is, decoder, the restored image $X \in \mathbb{R}^{H \times V \times 3}$ is obtained by output convolution as follows:

$$X = \text{Conv}(\text{TokenDec}^3(Z_N)). \tag{9}$$

The detailed Unet used in our experiments is presented in Supplementary Material.

Models	Computational Budget					GoPro [30]		REDS [33]		RealBlur-J [32]		DVD [35]	
	C	N	MACs	Params	Mems	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Restormer [53]	64	16	75.9G	54.3M	3.6G	31.96	0.9527	28.51	0.8570	28.87	0.9094	31.28	0.9367
NAFNet [6]	64	28	59.0G	54.5M	2.2G	32.34	0.9558	28.65	0.8588	28.99	0.9089	31.48	0.9393
MG-MLP (ours)	64	56	54.4G	46.7M	2.1G	**32.87**	**0.9604**	**28.80**	**0.8613**	**29.14**	**0.9110**	**31.59**	**0.9409**

Table 1: Blind deblurring performance on GoPro [30], REDS [31], RealBlur [32], and DVD [35] datasets: PSNR and SSIM are averaged over the test image in each dataset. The best results are printed in boldface.

RealBlur-J trained	GoPro [30]		HIDE [34]		RealBlur-R [32]		REDS [33]		DVD [35]		Throughput (img/s)
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
Restormer [53]	24.29	0.8468	23.58	0.8197	35.02	0.9298	25.01	0.7700	25.79	0.8687	0.575
NAFNet [6]	24.17	0.8578	23.31	0.8241	35.31	0.9309	24.51	**0.7757**	25.20	0.8783	0.690
MG-MLP (ours)	**24.54**	**0.8596**	**23.95**	**0.8298**	**35.46**	**0.9337**	**25.16**	0.7745	**26.20**	**0.8854**	**0.787**

Table 2: Deblurring performance of the model trained with the RealBlur-J [32] dataset: the trained model is evaluated on GoPro [30], HIDE [34], RealBlur-R [32], and DVD [35] test sets. The throughput is tested using 1280x720 images.

4 Experiment

We trained the three networks, including two baselines, NAFnet [6], Restormer [53], and the proposed MG-MLP, on four public spatially variant deblurring datasets, namely, GoPro [30], REDS [33], RealBlur [32], and DVD [35], to compare image restoration performance. State-of-the-art methods present unique components, including network structures and learning-related hyperparameters, and finally suggest the optimal combination. However, in this case, identifying where the superiority of restoration performance comes from is difficult. Indeed, performance highly depends on learning hyperparameters such as the learning rate scheduling and batch size. To determine the performance gains from the proposed architecture of the NN block, we equalized all hyperparameters used in all comparative experiments, such as macro architecture, learning-rate schedule, batch size, and patch size.

Unified Framework The simple Unet was used in the macro architecture as we discussed in Section 3, in which the dimensions of the tokens of the three evaluated NN blocks were the same as $C = 64$ and the number of blocks in the bottleneck layers N were set to 56 for our MG-MLP, 28 for NAFnet, and 16 for Restormer. These values were set to match the number of parameters and floating-point operations. Therefore, the required number of multiply-accumulate operations (MACs) when feeding a 256x256 image is 54.4G, 59.0G, and 75.9G for MG-MLP, NAFNet, and Restormer, respectively. The computational cost of the proposed block was 8.5% and 39.5% lower than those of NAFnet and Restormer, respectively. The number of parameters in the proposed MG-MLP was 16% less than that of the two baselines, and this configuration reduces memory usage during training by 42% compared to the Restormer (see Table 1). We adopted cosine annealing [28] to decay the learning rate from 1×10^3 to 1×10^7 for a total of 400k iterations. The batch size was set to 8 for all experiments. The training image size was 256×256, and the random horizontal and vertical flips were applied to image pairs. We used AdamW [29] with $\beta_1 = 0.9$, $\beta_2 = 0.9$, and optimized the models with peak signal-to-noise ratio (PSNR) loss

Motion Deblurring Before evaluating the restoration models using the unified framework, we trained NAFnet and Restormer with the GoPro dataset using the optimal setup and learning methods suggested in each paper, except the batch size, which was changed to 8. These

Figure 3: Three example images from the GoPro dataset [30] restored by three different networks are demonstrated. The first column contains the degraded input images. The next three columns show the reconstructed images obtained using Restormer [53], NAFnet [6], and our MG-MLP. The final column contains the ground truth images.

Figure 4: Two example images from the RealBlur-J dataset [32] restored by three different networks are presented.

two baselines were then compared with our MG-MLP trained using the unified framework. The results are presented in Figure 2 (a) and Table 1 in the Supplementary Material. The validation results during the training of the proposed block steadily outperformed the baselines. The final PSNR result of MG-MLP improved by 0.64 dB compared with NAFNet and 0.80 dB higher than that of Restormer. The SSIM was improved by 0.0064 and 0.0091 compared with NAFnet and Restormer, respectively.

Next, we replaced the NN blocks of the unified Unet with those of Restormer and NAFNet and trained the models with four public spatially variant deblurring datasets, including GoPro, REDS, RealBlur-J, and DVD. We then evaluated the learned models on each test set. The comparative results are reported in Table 1 and the convergence characteristics for the GoPro, REDS, and RealBlur-J datasets are displayed in Figure 2 (b), (c), and (d), respectively. The proposed network block outperformed the baselines for all datasets, with the smallest number of operations in a given computational budget. Figure 3, 4, and 5 display a qualitative comparison of the images restored by the three networks.

A cross-dataset evaluation was conducted to compare the generalization performance of the deblurring of models trained on a specific dataset. As a training dataset, we selected the JPEG version of the RealBlur dataset, which is the most recent and known to be more realistic than other synthetic datasets. Next, we evaluated the three trained models on four other test sets (GoPro, REDS, RealBlur-J, and DVD). Finally, the results are presented in Table 2. This result also revealed that the proposed MG-MLP has an outstanding generalization performance compared with the baselines. We evaluated the inference speeds of the networks. The throughput of our MG-MLP, evaluated by the 1280×720 images, was 0.787

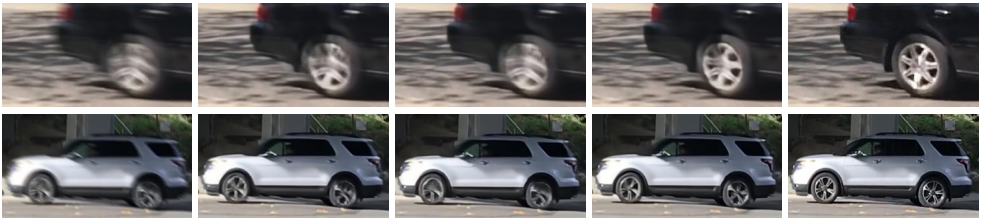

Figure 5: Two example images from the DVD dataset [35] restored by three different networks are presented.

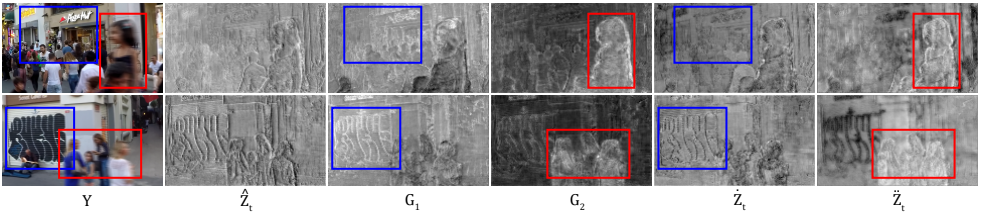

Figure 6: Visualizations of intermediate features in the MG-MLP block are demonstrated. The first column contains the input images. The following three columns respectively show the relative intensity of the primary pipeline, intra-token gating, and cross-token gating weight. The final two columns contain the learned two residuals. The blue and red boxes indicate the less blurry and more blurry region. Refer to the notations in Figure 1 and Eq. (1)-(6).

images/s, which is 1.36 times faster than Restormer, which requires channel-wise attention calculation.

Feature Visualization The intermediate token map was analyzed to investigate information contained in the tokens of each node. We averaged and normalized each token of the five nodes along the channel direction in the final MG-MLP block: $\hat{Z}_t$, which is the tokens of the primary pipeline; G_1 intra-token gating weights; G_2 cross-token gating weights; $\dot{Z}_t$ updated the feature by $(1-G_2)$, and $\ddot{Z}_t$, which is brought back from $\hat{Z}_t$ by reuse gating. Figure 6 reveals that (1) the feature intensity refined by intra-token gating is concentrated in a relatively less blurry region, and (2) the cross-token gating weight highlights blurry region. This phenomenon can be attributed to the cross-token gating mechanism, which operates on a blurry area that requires a wider receptive field. (3) The two residuals refined by the two gating mechanisms contain complementary features. The two gating paths perform distinct feature refinements by dividing a region into more blurry and less blurry regions.

Ablation The effectiveness of the proposed components was verified through an ablation experiment. The ablation setup includes the projection methods used for generating the two gating weights: the combination of channel mixing and token mixing, and the gating connection, that is, the presence of reuse and update connections. The results are displayed in Table 3.

Standard MLP-mixer The standard MLP is not fully convolutional, thus we implemented it using 1024 to 1024 linear MLP after flattening the tokens (32x32 in our case), then compared its performance with those of two full convolutional models with different kernel sizes (3x3, 5x5) using center cropped 256x256 images. Even as its complexity increases, standard

Ablation (C = 32)				GoPro testset [40]	
G_1	G_2	Reuse	Update	PSNR	SSIM
Channel	N/A			31.54	0.9465
Channel	Token	✓		31.09	0.9433
Channel	Token		✓	31.30	0.9452
Token	Channel	✓	✓	31.36	0.9459
Channel	Token	✓	✓	31.76	0.9500

Table 3: Results of the ablation study: the first row shows PSNR and SSIM values for intra-token gating alone. This is then augmented by adding cross-token gating with reuse connection (2nd row) and update connection (3rd row). The 4th row shows results when swapping the projection method of G1 and G2. The final row shows the results of the proposed architecture.

MLP performs worse than MG-MLP. This indicates that our token embedding has allowed a single token to contain enough receptive fields. In addition, after careful analysis, we determined that padding is the source of the large kernel MG-MLP's performance degradation.

GoPro	MG-MLP 3x3		MG-MLP 5x5		Std. MLP	
	PSNR	Macs(G)	PSNR	Macs(G)	PSNR	Macs(G)
RST./CPX.	**32.65**	**53.45**	32.45	56.12	32.44	84.70

Table 4: Comparison with standard MLP

5 Limitation and Conclusion

The proposed network can be applied to the image restoration of spatially uniform degradations such as denoising and deraining. The denoising and deraining MG-MLP trained with the SIDD dataset [1] and Rain13 dataset [20] exhibited a lower performance than NAFnet. These results empirically determined that the proposed method is appropriate for spatially variant degradation. This can be attributed to the two gating paths, which can understand the degradation level of regions. In the future, a new NN block that is applicable to various restoration problems, including spatially uniform degradation, should be developed.

6 Acknowledgement

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.20210-01343, Artificial Intelligence Graduate School Program (Seoul National University)] and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A3B1077720).

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

[3] George Cazenavette and Manuel Ladron De Guevara. Mixergan: An mlp-based architecture for unpaired image-to-image translation. *arXiv preprint arXiv:2105.14110*, 2021.

[4] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021.

[5] 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.

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

[7] Shoufa Chen, Enze Xie, Chongjian Ge, Ding Liang, and Ping Luo. Cyclemlp: A mlp-like architecture for dense prediction. *arXiv preprint arXiv:2107.10224*, 2021.

[8] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021.

[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

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

[11] 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.

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

[13] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12175–12185, 2022.

[14] Jianyuan Guo, Yehui Tang, Kai Han, Xinghao Chen, Han Wu, Chao Xu, Chang Xu, and Yunhe Wang. Hire-mlp: Vision mlp via hierarchical rearrangement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 826–836, 2022.

[15] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33:19000–19015, 2020.

[18] Qibin Hou, Zihang Jiang, Li Yuan, Ming-Ming Cheng, Shuicheng Yan, and Jiashi Feng. Vision permutator: A permutable mlp-like architecture for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[20] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8346–8355, 2020.

[21] Jaihyun Koh, Jangho Lee, and Sungroh Yoon. Single-image deblurring with neural networks: A comparative survey. *Computer Vision and Image Understanding*, 203: 103134, 2021.

[22] Jaihyun Koh, Jangho Lee, and Sungroh Yoon. Bnudc: A two-branched deep neural network for restoring images from under-display cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1950–1959, 2022.

[23] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie. Shape-texture debiased neural network training. *arXiv preprint arXiv:2010.05981*, 2020.

[24] Dongze Lian, Zehao Yu, Xing Sun, and Shenghua Gao. As-mlp: An axial shifted mlp architecture for vision. *arXiv preprint arXiv:2107.08391*, 2021.

[25] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlps. *Advances in Neural Information Processing Systems*, 34:9204–9215, 2021.

[26] Ruiyang Liu, Yinghui Li, Linmi Tao, Dun Liang, and Hai-Tao Zheng. Are we ready for a new paradigm shift? a survey on visual deep mlp. *Patterns*, 3(7):100520, 2022.

[27] 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.

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

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[30] 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.

[31] Seungjun Nah, Sungyong Baik, Seokil Hong, Gyeongsik Moon, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2019 challenge on video deblurring and super-resolution: Dataset and study. In *CVPR Workshops*, June 2019.

[32] 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.

[33] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5572–5581, 2019.

[34] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[35] 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.

[36] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017.

[37] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE international conference on computer vision*, pages 4539–4547, 2017.

[38] Chuanxin Tang, Yucheng Zhao, Guangting Wang, Chong Luo, Wenxuan Xie, and Wenjun Zeng. Sparse mlp for image recognition: Is self-attention really necessary? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2344–2351, 2022.

[39] Yehui Tang, Kai Han, Jianyuan Guo, Chang Xu, Yanxi Li, Chao Xu, and Yunhe Wang. An image patch is a wave: Phase-aware vision mlp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10935–10944, 2022.

[40] Yuki Tatsunami and Masato Taki. Raftmlp: Do mlp-based models dream of winning over computer vision? *arXiv e-prints*, pages arXiv–2108, 2021.

[41] Ilya O 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. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021.

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

[43] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

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

[45] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5780, 2022.

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

[47] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3365–3387, 2020.

[48] Ziyu Wang, Wenhao Jiang, Yiming M Zhu, Li Yuan, Yibing Song, and Wei Liu. Dynamixer: a vision mlp architecture with dynamic mixing. In *International Conference on Machine Learning*, pages 22691–22701. PMLR, 2022.

[49] Guoqiang Wei, Zhizheng Zhang, Cuiling Lan, Yan Lu, and Zhibo Chen. Activemlp: An mlp-like architecture with active token mixer. *arXiv preprint arXiv:2203.06108*, 2022.

[50] Tan Yu, Xu Li, Yunfeng Cai, Mingming Sun, and Ping Li. S2-mlp: Spatial-shift mlp architecture for vision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 297–306, 2022.

[51] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7): 1235–1270, 2019.

[52] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021.

[53] 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.

[54] David Junhao Zhang, Kunchang Li, Yunpeng Chen, Yali Wang, Shashwat Chandra, Yu Qiao, Luoqi Liu, and Mike Zheng Shou. Morphmlp: A self-attention free, mlp-like backbone for image and video. *arXiv preprint arXiv:2111.12527*, 2021.

[55] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2472–2481, 2018.

[56] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2020.

[57] Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Mixing and shifting: Exploiting global and local dependencies in vision mlps. *arXiv preprint arXiv:2202.06510*, 2022.

[58] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[59] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, Haozhe Xie, Jinshan Pan, and Jimmy S Ren. Davanet: Stereo deblurring with view aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10996–11005, 2019.

[60] X Zhu, W Su, LW Lu, B Li, XG Wang, and Deformable DETR Dai J F. Deformable transformers for end-to-end object detection. In *Proceedings of the 9th International Conference on Learning Representations. Virtual Event, Austria: OpenReview. net*, 2021.