Journal of Image and Signal Processing 图像与信号处理, 2025, 14(3), 325-332
Published Online July 2025 in Hans. https://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2025.143029

Hans汉斯

# 深度稀疏门控Transformer图像去模糊模型

**路嘉伟**

内蒙古大学数学科学学院，内蒙古 呼和浩特

收稿日期：2025年6月3日；录用日期：2025年6月25日；发布日期：2025年7月4日

## 摘　要

**基于Transformer的图像去模糊方法已经取得了显著的成绩，现阶段已经现有的大多数Transformer图像恢复方法将内部模块设计为自注意力 + 前馈网络的模式。为了降低这样设计带来的巨大计算开销与时间成本，本文提出了一种能够同时融合空间特征与通道特征的深度稀疏门控自注意力求解器。该方法通过Top-*k*稀疏选择与ReLU²稀疏激活将注意力转化为深度稀疏的形式，能够有效地消除令牌全局交互带来的冗余表示，还能增强通道特征融合能力。此外，本文通过设计判别式频域门控模块实现自适应保留与增强对图像恢复有帮助的特征，进一步完成空间特征融合。由这些基本模块组成的神经网络在GoPro基准数据集上取得了先进的结果。**

## 关键词

**图像去模糊，神经网络，Transformer，自注意力，特征融合**

# Deep Sparse Gated Transformer for Image Deblurring

**Jiawei Lu**

School of Mathematical Sciences, Inner Mongolia University, Hohhot Inner Mongolia

Received: Jun. 3rd, 2025; accepted: Jun. 25th, 2025; published: Jul. 4th, 2025

## Abstract

**Transformer-based image deblurring methods have achieved remarkable results. Currently, most existing Transformer-based image restoration approaches adopt a design pattern of self-attention and feed-forward networks for their internal modules. To reduce the substantial computational overhead and time costs associated with such designs, this paper proposes a deep sparse gated self-attention solver capable of simultaneously integrating spatial and channel features. By employing Top-*k* sparse selection and ReLU² sparse activation, this method transforms attention into a deep**

文章引用: 路嘉伟. 深度稀疏门控 Transformer 图像去模糊模型[J]. 图像与信号处理, 2025, 14(3): 325-332.
DOI: 10.12677/jisp.2025.143029

**sparse form, effectively eliminating redundant representations caused by token-wise global interactions while enhancing channel feature fusion capabilities. Furthermore, this paper designs a discriminative frequency-domain gating module to adaptively preserve and enhance features beneficial for image restoration, thereby further improving spatial feature fusion. The neural network composed of these fundamental modules achieves state-of-the-art results on the GoPro benchmark dataset.**

## Keywords

**Image Deblurring, Neural Networks, Transformer, Self-Attention, Feature Fusion**



## 1. 引言

图像去模糊是经典的底层视觉任务，它的目标是从观察到的模糊图像中恢复出干净图像。早期的方法[1] [2]通过模糊图像与清晰图像之间的统计学特征差异来构造各种先验模型。虽然先验模型能够处理简单的均匀模糊场景，但对于复杂退化场景的去模糊性能不佳且算法不够鲁棒。

为了克服模型驱动方法的缺点，许多先进的方法都采用 CNN 架构(例如[3]-[5])来实现高质量的图像去模糊。由于卷积运算局部连接与权重共享的特性，基于 CNN 的方法不具有消除远程模糊退化的能力。为了克服该缺点，具有全局建模能力的视觉 Transformer 架构(Vision Transformer, ViT)被用于图像去模糊，并且取得了良好的性能表现。

本文注意到，标准的 ViT 架构[6]通常通过所有的查询–键组合来计算自注意力特征以此达到融合全局特征的目的。事实上，来自于查询的令牌与来自键的令牌并不总是具有相关性[4]。因此在融合特征的过程中使用这些互不相关的令牌来计算注意力特征时会直接干扰接下来的图像恢复过程。这些发现激励我们探索最有用的自注意力值，以便能够充分利用这些注意力来进行图像恢复。

为此，本文开发了一种端到端的判别式稀疏门控 Transformer 网络(Discriminative Sparse Gated Transformer, DSGformer)用于图像去模糊任务。具体而言，DSGformer 框架通过判别式稀疏注意力模块(Discriminative Sparse Gated Attention, DSGA)来保留最有用的自注意力分数，以此提升去模糊性能。进一步地，DSGA 模块包含两个核心部分：动态加权的 Top-$k$ 稀疏注意力分数(Dynamic Top-$k$ Sparse Attention Score, DTSAS)与判别式频域门控模块(Discriminative Frequency-domain Gated Block, DFGB)。其中，DTSAS 用于代替传统的注意力分数，通过 Top-$k$ 稀疏选择与 $\mathrm{ReLU}^2$ 稀疏激活保留相关性较强的注意力分数，实现有效的通道特征融合。

此外，在特征融合的过程中，并不是所有的高频与低频信息都对图像恢复有帮助。为了产生更有效的特征用于恢复图像，本文设计了一种简单有效的基于频域的判别式门控模块 DFGB。该模块是由联合摄影专家组压缩算法(Joint Photographic Experts Group, JPEG)驱动的。它在门控模块中引入一种频域判别学习机制，以确定应该保留哪些高频与低频信息用于图像恢复。

综上，本文提出的 DSGformer 模型的主要贡献如下：

(1) 本文提出了一种动态加权的 Top-$k$ 稀疏注意力分数模块 DTSAS 用于计算注意力分数。DTSAS 同时结合了 Top-$k$ 稀疏选择与 $\mathrm{ReLU}^2$ 稀疏激活，能够有效地降低全局令牌交互带来的冗余表示，消除不相关或弱相关的注意力分数，保留相关性较强的注意力分数，使模型实现有效的通道特征融合。

(2) 本文提出一种判别式频域门控模块 DFGB 用于融合空间特征。DFGB 能够在频域中自适应地保留并且增强对图像恢复最有用的频率分量。此外，通过判别式频域学习后的门控特征，能够直接与注意力特征进行融合，实现高效的空间信息交互。

(3) 本文将 DTSAS 与 DFGB 组成了一个能够同时融合通道与空间特征的深度稀疏门控注意力模块 DSGA，并且利用 DGSA 模块构建了一个对称的编码－解码网络 DSGformer 用于图像去模糊。在基准数据集上的实验证明了本章方法的有效性。

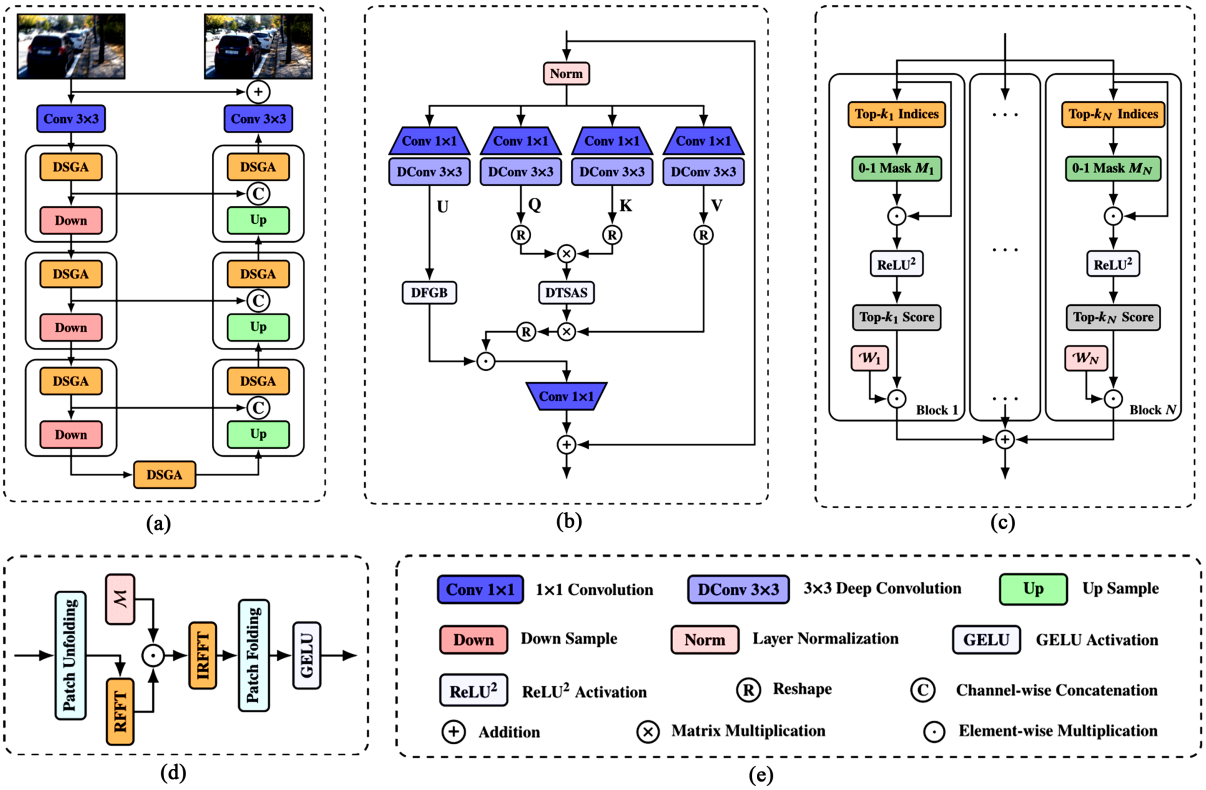


**Figure 1.** The neural network architecture. (a) The overall architecture of the deep sparse gated transformer. (b) Deep sparse gated attention. (c) Dynamic Top-$k$ sparse attention scores. (d) Discriminative frequency-domain gated block. (e) Symbol explanation

**图 1.** 本章提出的神经网络架构。(a) 深度稀疏门控 Transformer 的整体架构。(b) 深度稀疏门控注意力。(c) 动态 Top-$k$ 稀疏注意力分数。(d) 判别式频域门控模块。(e) 符号说明

## 2. 算法框架

### 2.1. 判别式稀疏门控注意力

与一般的 ViT [6]不同的是，DSGA 模块在将输入特征投影为$(Q,K,V)$时，还会额外地投影出一个门控特征 $U$。也就是说，DSGA 模块会将输入特征投影为四元组$(Q,K,V,U)$。具体有：

$$Q=W_d^Q W_p^Q \mathrm{LN}(X),\quad K=W_d^K W_p^K \mathrm{LN}(X),\quad V=W_d^V W_p^V \mathrm{LN}(X),\quad U=W_d^U W_p^U \mathrm{LN}(X),$$

其中，LN()为层归一化，$W_p$ 为1×1逐点卷积，$W_d$ 为3×3 深度卷积。

**动态 Top-$k$ 稀疏注意力分数。**本文注意到，现有的大多数基于 Transformer 图像恢复模型(例如 IPT [7]与 Restormer [8]等)中，一般采用标准的点积缩放自注意力机制：

$$\mathrm{Att}(Q,K,V)=\mathrm{Softmax}\left(\frac{Q\cdot K^{\mathrm{T}}}{\alpha}\right)\cdot V.$$

需要注意的是，这样的简单自注意力模式是基于密集的全连接操作，而本文提出了 DTSAS 来替代它，从而有效地避免无效信息参与特征交互过程。具体而言，首先利用查询与键计算像素对之间的相似度，再利用一个二值掩码屏蔽取值较小的注意力分数。使用 Top-$k$ 操作实现了对前 $k$ 个贡献度大的分数的自适应选择，这里 $k\in[0,1]$ 是一个可调节参数，能够控制稀疏程度的大小。因此，这样仅使用注意力分数每一行的前 $k$ 个最大值激活，而剩余的注意力分数均替换为 0。因此这种选择方式能够使注意力的计算模式从密集转化为稀疏，具体由如下的公式导出：

$$\mathrm{Att}(Q,K,V)=\mathrm{DTSAS}(Q,K)\cdot V=\sum_{n=1}^{N}W_n\mathrm{ReLU}^2\left(T_{k_n}\left(\frac{Q\cdot K^{\mathrm{T}}}{\alpha}\right)\right)\cdot V,$$

其中，$W_n$ 为可学习的分配权重，$T_{k_n}$ 为 Top-$k_n$ 选择算子：

$$\left[T_{k_n}(A)\right]_{ij}=\begin{cases}A_{ij}, & A_{ij}\geq(\tau_i)_n,\\ 0, & A_{ij}<(\tau_i)_n.\end{cases}$$

对于一个输入特征 $A=\left[A_{ij}\right]_{C\times C}$，Top-$k$ 选择算子只保留每一行特征中最大的 $k$ 个分量，将不相关的特征融合结果丢弃。$(\tau_i)n$ 代表第 $i$ 行中的第 $k_n$ 个最大分量，也就是第 $i$ 行的阈值。动态 Top-$k$ 稀疏注意力分数的结构如图 1(c)所示。

**判别式频域门控模块。**并非所有的高频信息与低频信息都有利于清晰图像的恢复。因此我们开发了一种能够自适应地确定应该保留那些频率信息的判别式频域门控模块 DFGB。但如何有效地确定哪些频率信息十分重要。具体而言，在 JPEG 压缩算法的启发下，我们通过引入一个可学习的量化矩阵 $\mathcal{M}$ 并通过 JPEG 压缩的逆方法来学习它，以确定保留哪些频率信息。利用判别式频域门控模块对门控特征进行特征变换：

$$\begin{aligned}&\text{Step1:}\quad U_1=\mathcal{F}_{\mathrm{R}}\left(\mathcal{P}(U)\right),\\&\text{Step2:}\quad U_2=\mathcal{F}_{\mathrm{R}}^{-1}\left(\mathcal{M}\odot U_1\right),\\&\text{Step3:}\quad U_{\mathrm{out}}=\mathrm{GELU}\left(\mathcal{P}^{-1}\left(U_2\right)\right).\end{aligned}$$

其中，$\mathcal{P}$ 与 $\mathcal{P}^{-1}$ 代表 JPEG 压缩算法中的块展开与块折叠算子；$\mathcal{F}_R$ 与 $\mathcal{F}_R^{-1}$ 分别代表实部 Fourier 变换及其逆变换；$\mathcal{M}$ 为为可学习的量化矩阵；GELU()为 GELU 激活函数。判别式频域门控模块的具体结构如图 1(d)所示。

之后，将注意力特征 $\mathrm{Att}(Q,K,V)$ 与激活的门控特征 $\mathrm{DFGB}(X)$ 进行逐点乘法实现空间特征融合：

$$\mathcal{G}\left(\mathrm{Att}(Q,K,V)\right)=\mathrm{DFGB}(U)\odot\mathrm{Att}(Q,K,V).$$

最后利用输出投影与残差连接得到最终的变换结果：

$$Y=W_p\mathcal{G}\left(\mathrm{Att}(Q,K,V)\right)+X,$$

其中，$W_p$ 为1×1逐点卷积。

## 2.2. 编码解码网络 DSGformer

编码解码网络 DSGformer 主要由四部分组成：头部(Head)，编码器(Encoder)，解码器(Decoder)与尾部(Tail)。DSGformer 将输入退化图像 B 输入头部后，经过编码解码处理，再由尾部输出最终的恢复图像。

每一部分的具体细节如下：

**头部。**模糊图像 $B\in\mathbb{R}^{H\times W\times 3}$ 首先输入头部 $\mathcal{H}$，即经过一个 3×3 的卷积层，由此得到浅层特征 $H$，将架构头部的参数记为 $\Theta_{head}$，该过程可以表示为：

$$H=\mathcal{H}\left(B,\Theta_{head}\right)\in\mathbb{R}^{H\times W\times C},$$

其中，$H\times W$ 代表特征的空间维数，$C$ 代表通道数。

**编码器。**浅层特征 $H$ 作为编码器 $\mathcal{E}$ 的输入，通过四层的编码器 $\mathcal{E}$ 得到四个中间特征 $E_1,E_2,E_3,E_4$：

$$\left(E_1,E_2,E_3,E_4\right)=\mathcal{E}\left(H,\Theta_{enc}\right),$$

其中，$\Theta_{enc}$ 为编码器的参数，四个中间特征的维数分别为：

$$E_1\in\mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 2C},\quad E_2\in\mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 4C},\quad E_3,E_4\in\mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 8C}.$$

**解码器。**解码器将四个中间特征作为输入，将它们解码为一个深度特征 $D$：

$$D=\mathcal{D}\left(E_1,E_2,E_3,E_4,\Theta_{dec}\right)\in\mathbb{R}^{H\times W\times 2C},$$

其中，$\Theta_{dec}$ 为解码器参数。

**尾部。**利用一个 3×3 的卷积层作为尾部将深度特征 $D$ 转化为残差图 $T$，再与输入图像 $B$ 进行残差连接，得到恢复图像：

$$T=\mathcal{T}\left(D,\Theta_{tail}\right)\in\mathbb{R}^{H\times W\times 3},\quad \hat{S}=T+B,$$

其中，$\Theta_{tail}$ 为尾部参数。

给定一幅模糊图像 $B$，目标架构通过学习残差的形式得到最终的恢复图像 $S$：

$$\hat{S}=\mathcal{N}\left(B\right)+B,$$

其中，$\mathcal{N}\left(\cdot\right)$ 代表 DSGformer，由头部，编码器，解码器与尾部等四部分按照函数复合的形式组成。具体地，$\mathcal{N}\left(B\right)$ 通过如下方式定义：

$$\mathcal{N}\left(B\right)=\mathcal{T}\left(\mathcal{D}\left(\mathcal{E}\left(\mathcal{H}\left(B,\Theta_{head}\right),\Theta_{enc}\right),\Theta_{dec}\right),\Theta_{tail}\right).$$

因此，参数集 $\left\{\Theta_{head},\Theta_{enc},\Theta_{dec},\Theta_{tail}\right\}$ 为 DSGformer 的网络参数。

## 3. 数值实验

### 3.1. 实验设置

本文在 PyTorch 深度学习环境下搭建、训练与测试 DSGformer。训练过程在 1 块 NVIDIA RTX 3060 12GB GPU 上进行。使用 AdamW 优化器对参数集进行迭代更新。损失函数采用 $L_1$ 损失。另外，提出的模型在 GoPro 数据集上训练 3000 轮，学习率初始化为 2e−4。根据不同的学习率调度方式，将 3000 训练轮次分为两个阶段：前 10 轮为线性预热阶段，后 2990 轮利用余弦退火算法将学习率从 2e−4 衰减到 1e−6。在训练 DSGformer 时，批量大小设置为 8。与先进的图像恢复方法类似，DSGformer 在每次接收图像数据后，随机从图像中裁剪出一个尺寸为 256 图像块，再对图像块进行随机的几何增强。

### 3.2. 数据集

本文的方法在 GoPro 数据集[3]上进行训练与测试。GoPro 数据集包含了 2103 对训练图像以及 1111 对测试图像。

### 3.3. 与先进方法的对比

**对比方法。**为了评估神经网络的性能，将目标方法与最新的先进算法进行比较，例如基于 CNN 的方法：MPRNet [9]、ConvIR [5]。我们还与先进的基于 ViT 的算法进行了比较，包括：Restormer [8]、Stripformer [10]、以及 MB-TaylorFormer-V2 [11]。

**Table 1.** Comparison of quantitative evaluations for different methods on the GoPro dataset
**表 1.** 不同方法在 GoPro 数据集上的性能评估结果

| 去模糊方法 | 平均 PSNR | 平均 SSIM | 模型参数量 (M) |
|---|---|---|---|
| MPRNet (CVPR 2021) | 32.66 | 0.959 | 20.1 |
| Restormer (CVPR 2022) | 32.92 | 0.961 | 26.1 |
| Stripformer (ECCV 2022) | 33.08 | 0.962 | 19.7 |
| ConvIR (TPAMI 2024) | 33.28 | 0.963 | 14.83 |
| MB-TaylorFormer-V2 (TPAMI 2025) | 33.24 | 0.963 | 7.29 |
| DSGformer (本文方法) | 33.72 | 0.966 | 15.57 |

**在 GoPro 数据集上的评估结果。**表 1 给出了不同方法在 GoPro 数据集上的定量比较结果。与基于 CNN 架构的 ConvIR 相比，PSNR 能够提高 0.44 dB。与先进的 ViT 架构的算法 MB-TaylorFormer-V2 相比，目标方法能够取得 0.48 dB 的 PSNR 性能增益。另外，图 2 展示了一个具体的视觉对比实例。可以看出我们提出的 DSGformer 则同时能够恢复更多的细节信息，得到与真实图像最接近的结果。

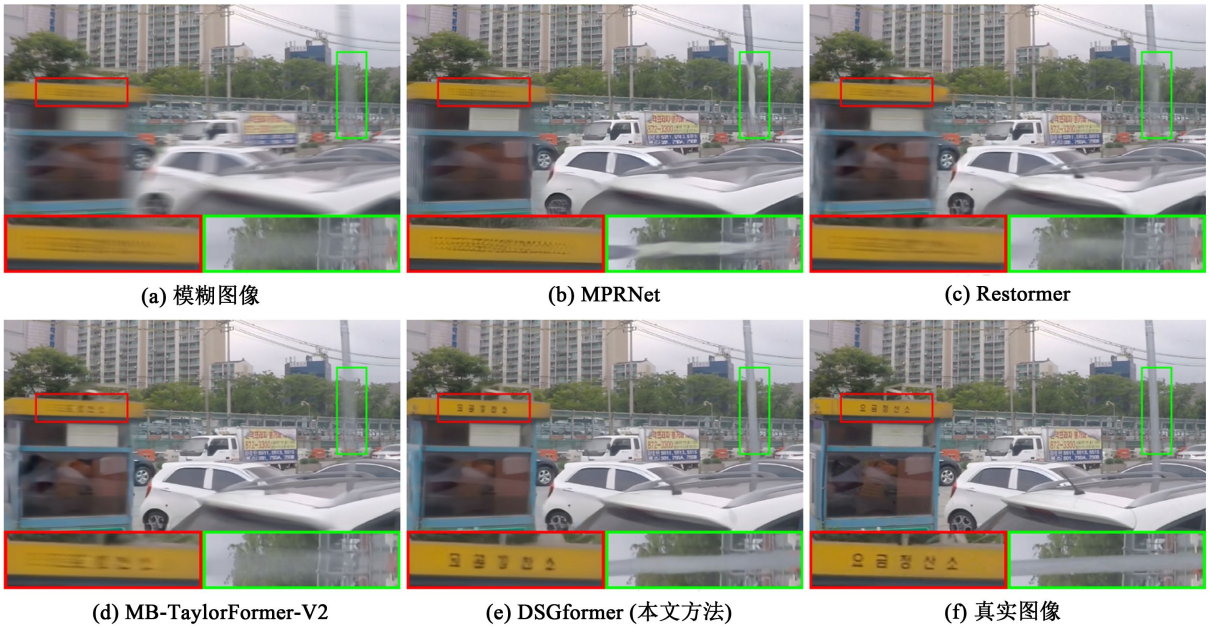


**Figure 2.** The visual comparison examples from the GoPro dataset
**图 2.** 一个来自 GoPro 数据集的去模糊视觉对比实例

### 3.4. 消融实验

本次消融实验在真实数据集 RealBlur-J [12]上进行。

**DSGA 模块的有效性。**为了研究深度稀疏门控注意力的有效性，本节将其替换为多头转置自注意力[8] (Multi-Deconv Head Transposed Attention, MDTA)。表的结果可以看出，当门控模块固定时，使用 DSGA 会比使用 MDTA 带来更好的性能增益。这也说明使用深度稀疏门控注意力能够有效地消除负相关与冗余令牌，得到紧凑且相关性强的注意力计算结果，能够有效地促进清晰图像重建。

**DFGB 模块的有效性。**为了研究判别式频域门控模块的有效性，本节使用 GELU 激活与 SimpleGate 进行替换。表 2 的结果可以看出，当注意力模块固定时，使用 DFGB 模块能够获得最佳的性能收益，GELU 激活与 SG 都无法使模型达到最好的性能。这是因为 DFGB 模块在融合空间特征时，能够在频域进行判别学习，消除无用的高低频信息，增强空间特征的聚合能力。

**Table 2.** Effectiveness analysis of the proposed modules in the target model on the RealBlur-J dataset
**表 2.** 目标模型提出的模块在 RealBlur-J 数据集上的有效性分析

| 对比方法 | SimpleGate | GELU | ReLU | ReLU$^2$ | Softmax | PSNR |
|---|---|---|---|---|---|---|
| SimpleGate + ReLU | √ | × | √ | × | × | 30.29 |
| SimpleGate + ReLU$^2$ | √ | × | × | √ | × | 31.61 |
| SimpleGate + Softmax | √ | × | × | × | √ | 30.76 |
| GELU + Softmax | × | √ | × | × | √ | 30.89 |
| GELU + ReLU | × | √ | √ | × | × | 30.09 |
| GELU + ReLU$^2$ (本文方法) | × | √ | × | √ | × | 31.56 |

## 4. 结论

本文提出 DSGformer，一种基于 Top-*k* 稀疏选择和频域判别学习的视觉 Transformer，用于图像去模糊。研究发现，传统注意力机制中存在大量不相关或低相关性的图像块，仅少数关键令牌主导图像恢复。为此，DSGformer 采用 Top-*k* 稀疏选择和 ReLU$^2$ 稀疏激活来消除冗余注意力表示，并通过频域判别学习模块自适应保留有用特征。实验表明，该方法在去模糊任务中具有竞争优势。

## 参考文献


[1] Liu, J., Yan, M. and Zeng, T. (2021) Surface-Aware Blind Image Deblurring. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 1041-1055. https://doi.org/10.1109/tpami.2019.2941472

[2] Pan, J., Sun, D., Pfister, H. and Yang, M. (2018) Deblurring Images via Dark Channel Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 2315-2328. https://doi.org/10.1109/tpami.2017.2753804

[3] Nah, S., Kim, T.H. and Lee, K.M. (2017) Deep Multi-Scale Convolutional Neural Network for Dynamic Scene Deblurring. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 257-265. https://doi.org/10.1109/cvpr.2017.35

[4] Chen, X., Li, H., Li, M. and Pan, J. (2023) Learning a Sparse Transformer Network for Effective Image Deraining. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 5896-5905. https://doi.org/10.1109/cvpr52729.2023.00571

[5] Cui, Y., Ren, W., Cao, X., *et al*. (2024) Revitalizing Convolutional Network for Image Restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **46**, 9423-9438.

[6] Alexey, D., Lucas, B., Alexander, K., *et al*. (2021) An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *The 9th International Conference on Learning Representations* (*ICLR*), Austria, 3-7 May 2021, 1-22.

[7] Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., *et al*. (2021) Pre-Trained Image Processing Transformer. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 12294-12305. https://doi.org/10.1109/cvpr46437.2021.01212

[8] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S. and Yang, M. (2022) Restormer: Efficient Transformer for High-Resolution Image Restoration. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 5718-5729. https://doi.org/10.1109/cvpr52688.2022.00564

[9] Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M., *et al*. (2021) Multi-Stage Progressive Image Restoration. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 14816-14826. https://doi.org/10.1109/cvpr46437.2021.01458

[10] Tsai, F., Peng, Y., Lin, Y., Tsai, C. and Lin, C. (2022) Stripformer: Strip Transformer for Fast Image Deblurring. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M. and Hassner, T., Eds., *Lecture Notes in Computer Science*, Springer, 146-162. https://doi.org/10.1007/978-3-031-19800-7_9

[11] Jin, Z., Qiu, Y., Zhang, K., Li, H. and Luo, W. (2025) MB-Taylorformer V2: Improved Multi-Branch Linear Transformer Expanded by Taylor Formula for Image Restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **47**, 5990-6005. https://doi.org/10.1109/tpami.2025.3559891

[12] Rim, J., Lee, H., Won, J. and Cho, S. (2020) Real-World Blur Dataset for Learning and Benchmarking Deblurring Algorithms. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Lecture Notes in Computer Science*, Springer International Publishing, 184-201. https://doi.org/10.1007/978-3-030-58595-2_12