

用于高质量图像去模糊的高效基于频域的变压器

孔令顺 1、董江新 1^â、葛建军 2、李明强 2 和潘金山 1^â

1 南京理工大学计算机科学与工程学院

2 中国电子科技集团公司信息科学研究院

Abstract

我们提出了一种有效且高效的方法来探索变换器在频域中的特性，以实现高质量图像去模糊。我们的方法受到卷积定理的启发，即空间域中两个信号的相关或卷积等效于它们在频域中的逐元素乘积。这启发我们开发一种高效的基于频域的自注意力力求解器 (FSAS)，以通过逐元素乘积运算而不是空间域中的矩阵乘法来估计缩放的点积注意力。此外，我们注意到在 Transformers 中简单地使用朴素的前馈网络 (FFN) 不会产生良好的去模糊结果。为了克服这个问题，我们提出了一种简单而有效的基于频域的判别式 FFN (DFFN)，我们在 FFN 中引入了一种基于联合图像专家组 (JPEG) 压缩算法的门控机制，以区别性地确定哪个是低频和高频——应保留特征的频率信息以恢复潜在的清晰图像。我们将所提出的 FSAS 和 DDFN 制定为基于编码器和解码器架构的非对称网络，其中 FSAS 仅用于解码器模块以实现更好的图像去模糊。实验结果表明，所提出的方法优于最先进的方法。

一、简介

图像去模糊旨在从模糊的图像中恢复高质量的图像。由于开发了具有大规模训练数据集的各种有效深度模型，该问题已取得重大进展。

大多数最先进的图像去模糊方法主要基于深度卷积神经网络 (CNN)。这些方法的主要成功归功于开发各种网络架构设计，例如，多

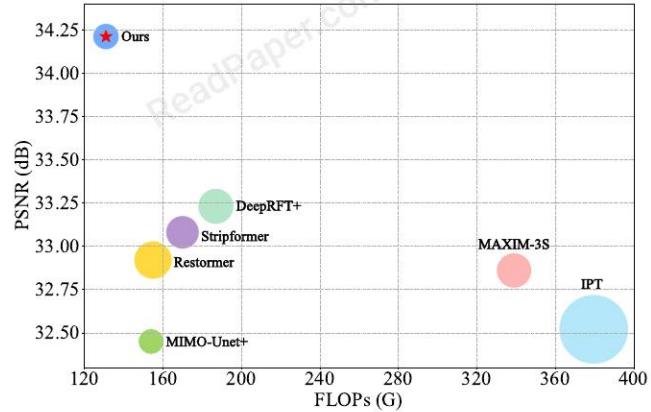


图 1. GoPro 数据集 [16] 上提出的方法与最先进方法在精度、浮点操作 (FLOP) 和网络参数方面的比较。圆圈大小表示网络参数的个数。

规模 [4, 16, 22] 或多阶段 [31, 32] 网络架构、生成对抗学习 [11, 12]、受物理模型启发的网络结构 [6–8, 17, 33] 等。作为这些网络中的基本操作，卷积操作是一种空间不变的局部操作，它不对图像内容的空间变化属性进行建模。他们中的大多数使用更大更深的模型来弥补卷积的局限性。然而，如 [17, 33] 所示，简单地增加深度模型的容量并不总是会带来更好的性能。

与对局部连通性建模的卷积运算不同，Transformer 能够通过计算一个标记与所有其他标记的相关性来对全局上下文建模。它们已被证明是许多高级视觉任务的有效方法，并且也有很大的潜力成为深度 CNN 模型的替代品。在图像去模糊中，基于 Transformers [27, 30] 的方法也比基于 CNN 的方法取得了更好的性能。然而，Transformers 中缩放点积注意力的计算导致了令牌数量的二次空间和时间复杂度。虽然使用 aller 和更少的标记可以减少空间和

*Corresponding author: Jiangxin Dong and Jinshan Pan.

时间复杂度，这种策略不能很好地对特征的远程信息进行建模，并且在处理高分辨率图像时通常会导致明显的伪影，从而限制了性能的提高。

为了缓解这个问题，大多数方法使用下采样策略来降低特征的空间分辨率[26]。然而，降低特征的空间分辨率会导致信息丢失，从而影响图像去模糊。有几种方法通过根据特征数量计算缩放的点积注意力来降低计算成本 [29, 30]。虽然计算成本降低了，但空间信息没有得到很好的探索，这可能会影响去模糊性能。在本文中，我们开发了一种有效且高效的方法来探索用于高质量图像去模糊的 Transformer 的属性。我们注意到，缩放点积注意力计算实际上是估计查询中一个标记与键中所有标记的相关性。这个过程可以在重新排列标记的排列时通过卷积运算来实现。基于这一观察和空间域中的卷积等于频域中的逐点乘法的卷积定理，我们开发了一种有效的基于频域的自注意求解器 (FSAS) 来估计通过逐元素乘积运算而不是矩阵乘法来缩放点积注意力。因此，对于每个特征通道，空间和时间复杂度可以降低到 $O(N) O(N \log N)$ ，其中 N 是像素数。此外，我们注意到仅使用 [30]

的前馈网络 (FFN) 不会产生良好的去模糊结果。为了为潜在的清晰图像恢复生成更好的特征，我们开发了一种简单而有效的基于判别频域的 FFN (DFFN)。我们的 DFFN 受到联合图像专家组 (JPEG) 压缩算法的启发。它在 FFN 中引入了一种门控机制，以有区别地确定应保留哪些低频和高频信息以用于潜在的清晰图像恢复。我们将提出的 FSAS 和 DFFN 制定为基于编码器和解码器架构的端到端可训练网络，以解决图像去模糊问题。然而，我们发现由于浅层特征通常包含模糊效果，将缩放点积注意力应用于浅层特征并不能有效地探索全局清晰内容。由于深层特征通常比浅层特征更清晰，我们开发了一种非对称网络架构，其中 FSAS 仅用于解码器模块以实现更好的图像去模糊。我们分析 Transformer 在频域中的探索特性能够促进模糊去除。实验结果表明，所提出的方法在准确性和效率方面对最先进的方法产生了有利的结果

(图1)。

主要贡献总结如下：• 我们开发了一种高效的基于频域的自注意力求解器来估计缩放的点积注意力。我们的分析表明，使用基于频域的求解器可以降低空间和时间的复杂性，并且更加有效和高效。• 我们提出了一种基于 JPEG 压缩算法的简单而有效的基于频域的判别式 FFN，以判别式地确定应保留哪些低频和高频信息

用于潜在的清晰图像恢复。• 我们开发了一种基于编码器和解码器网络的非对称网络架构，其中基于频域的自注意力求解器仅用于解码器模块，以实现更好的图像去模糊。• 我们分析了变换器在频域中的探索特性能够促进模糊去除，并表明我们的方法有利于巧妙地对抗最先进的方法。

二、相关工作

基于深度 CNN 的图像去模糊方法。近年来，由于不同的深度 CNN 模型 [3, 4, 9, 16, 22, 31, 32] 的发展，我们见证了图像去模糊方面的重大进步。在 [16] 中，Nah 等人。提出了一种基于多尺度框架的深度 CNN 来直接从模糊图像中估计清晰图像。为了更好地利用多尺度框架中每个尺度的信息，Tao 等人。[22] 开发一个有效规模的递归网络。高等。[9] 提出了一种选择性网络参数共享方法来改进 [16, 22]。

由于使用更多的尺度并不能显着提高性能，Zhang 等人。[32] 开发基于多补丁策略的有效网络。去模糊过程是逐步实现的。为了更好地探索不同阶段的特征，Zamir 等人。[31] 提出了一种跨阶段特征融合以获得更好的性能。为了降低基于多尺度框架的方法的计算成本，Cho 等人。[4] 提出了一个多输入和多输出网络。陈等。[3] 分析基线模块并简化它们以获得更好的图像恢复。正如 [30] 中所证明的，卷积运算是空间不变的，并且不能有效地对图像去模糊的全局上下文进行建模。变压器及其在图像去模糊中的应用。由于 Transformer [25] 可以对全局上下文进行建模，并在许多高级视觉任务（例如图像分类 [14]、对象检测 [1, 34] 和语义分割 [28, 35] 中取得了显着进步），它已被开发用于解决图像超分辨率 [13]、图像去模糊 [24, 30] 和图像去噪 [2, 27]。为了降低 Transformer 的计算成本，Zamir 等人。[30] 通过计算缩放的点积注意力提出了一个有效的 Transformer 模型

在特征深度域。该方法可以有效地探索沿通道维度的不同特征的信息。然而，对于图像恢复至关重要的空间信息尚未得到充分探索。蔡等人。[24] 通过构建带内和带间标记来代替全局注意力来简化自我注意力的计算。王等。[27] 提出了一种基于 UNet 的 Transformer，它使用非重叠的基于窗口的自注意力来进行单幅图像去模糊。虽然使用分裂策略降低了计算成本，但粗分裂并没有完全探索每个补丁的信息。此外，这些方法中的缩放点积注意力通常需要空间和时间复杂度为二次的复杂矩阵乘法。

与这些方法不同，我们开发了一种基于 Transformer 的高效方法，该方法探索频域的特性，以避免缩放点积注意力的复杂矩阵乘法。

3. 拟议方法

我们的目标是提出一种有效且高效的方法来探索用于高质量图像去模糊的 Transformer 的属性。为此，我们首先开发了一种高效的基于频域的自注意力求解器来估计缩放的点积注意力。为了改进基于频域的求解器估计的特征，我们进一步开发了一个基于频域的判别前馈网络。我们将上述这些方法制定为基于编码器和解码器架构的端到端可训练网络，以解决图像去模糊问题，其中使用基于频域的自注意力求解器来估计缩放的点积注意力在解码器模块中以获得更好的特征表示。图 2(a) 显示了所提出方法的概述。下面，我们介绍每个组件的详细信息。

3.1. Frequency domain-based self-attention solver

已有的视觉变换器通常在给定空间分辨率为 h 像素和 c 通道的输入特征 x 的情况下，通过 W_q , W_k 和 W_v 到 x 的线性变换计算特征 f_q , f_k 和 f_v ，然后对特征 f_q , f_k 和 f_v 应用展开函数提取图像块 $\{q_i\}_{i=1}^n$, $\{k_i\}_{i=1}^n$, $\{v_i\}_{i=1}^n$ ，其中 n 表示提取的块数。通过对提取的补丁应用重塑操作，可以通过以下方式获得查询 Q 、键 K 和值 V :

$$Q = R(\{q_i\}_{i=1}^n), \quad K = R(\{k_i\}_{i=1}^n), \quad V = R(\{v_i\}_{i=1}^n),$$

(1) 其中 R 表示重塑函数，它确保 $\{K, Q, V\} \in \mathbb{R}^{N \times C \times H_p \times W_p}$ ， H_p 和 W_p 表示提取的块的高度和宽度。基于获得的查询 Q 、键 K 和值 V ，缩放点积注意力通过以下方式实现:

$$V_{att} = \text{softmax} \left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{CH_pW_p}} \right) V. \quad (2)$$

注意力图的计算涉及到 QK 的矩阵多重化，其空间复杂度和时间复杂度分别为 $O(n^2)$ 和 $O(n^2c)$ 。如果图像分辨率和提取的斑块数量较大，则是负担不起的。虽然使用下采样操作来降低图像分辨率或非重叠方法提取更少的斑块可以缓解这一问题，但这些策略会导致信息丢失，并限制对每个斑块内和跨斑块的细节建模的能力[29]。

我们注意到 QK 的每个元素都是通过内积得到的:

$$(\mathbf{Q}\mathbf{K}^\top)_{ij} = \langle \mathbf{q}_i, \mathbf{k}_j \rangle, \quad (3)$$

其中 q_i 和 k_j 是由 f_q 和 f_k 得到的 i -th 和 j -th 块的向量化形式。在 (3) 的基础上，如果分别对 q_i 和所有的曲面片 k_j 应用重塑函数，则可以通过卷积运算得到 qk 的所有第 i 列元素，即 $q_i \hat{=} \sum_j q_i k_j$ ，其中 q_i 和 k_j 表示 q_i 和 k_j 的重塑结果； $\hat{=}$ 表示卷积运算。根据卷积定理，两个信号在空间域中的相关或卷积等效于它们在频域中的元素乘积。因此，一个自然的问题是，我们是否可以在频域中通过元素乘积运算而不是在空域中计算 QK 的矩阵倍数来直接匹配注意力图？

为此，我们开发了一种有效的基于频域的自注意力求解器。具体来说，我们首先通过 1×1 逐点卷积和 3×3 深度卷积获得 F_q 、 F_k 和 F_v 。然后，我们将快速傅立叶变换 (FFT) 应用于估计的特征 F_q 和 F_k ，并通过以下方式估计 F_q 和 F_k 在频域中的相关性:

$$A = \mathcal{F}^{-1} \left(\mathcal{F}(F_q) \overline{\mathcal{F}(F_k)} \right), \quad (4)$$

其中 $\mathcal{F}(\cdot)$ 表示 FFT， $\mathcal{F} \times 1(\cdot)$ 表示逆 FFT， $\mathcal{F}(\cdot)$ 表示共轭转置运算。最后，我们通过以下方式估计聚合特征:

$$V_{att} = \mathcal{L}(A) F_v, \quad (5)$$

其中层范数 $\mathcal{L}(\cdot)$ 用于对 A 进行归一化。最后，我们通过以下方式生成 FSAS 的输出特征:

$$X_{att} = X + \text{Conv}_{1 \times 1}(V_{att}), \quad (6)$$

其中 $\text{Conv}_{1 \times 1}(\cdot)$ 表示过滤器大小为 1×1 像素的卷积。所提出的 FSAS 的详细网络架构如图 2(b) 所示。

3.2. Discriminative frequency domain-based FFN

FFN 用于通过缩放点积注意力来改进特征。因此，开发一个

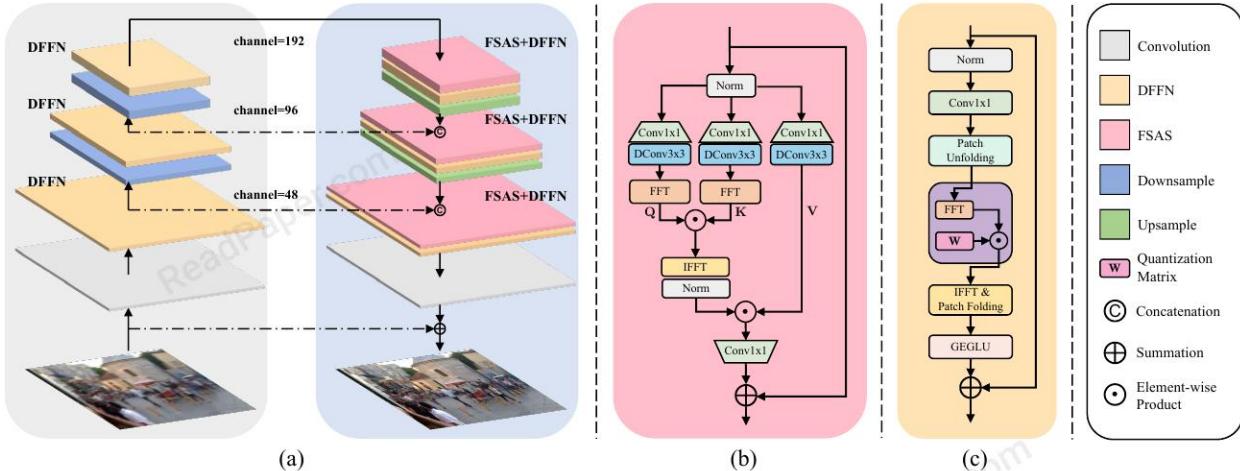


图 2. 网络架构。 (a) 所提出的非对称编码器-解码器网络，在编码器模块中仅包含 DFFN，在解码器模块中包含 FSAS 和 DFFN，用于图像去模糊。(b) 拟议的 FSAS 模块。(c) 提议的 DFFN 模块。

有效的 FFN 生成有助于潜在清晰图像重建的特征。由于并非所有的低频信息和高频信息都有助于潜在的清晰图像恢复，因此我们开发了一种可以自适应地确定应保留哪些频率信息的 DFFN。然而，如何有效地确定哪个频率信息很重要。受 JPEG 压缩算法的启发，我们引入了一个可学习的量化矩阵 W ，并通过 JPEG 压缩的逆向方法对其进行学习，以确定应保留哪些频率信息。拟议的 DFFN 可以通过以下方式制定：

$$X_1 = \text{Conv } 1 \times 1 (L(X_{\text{att}}))$$

$$\begin{aligned} X_1^f &= \mathcal{F}(\mathcal{P}(X_1)) \\ X_2 &= \mathcal{F}^{-1}(\mathbf{W}X_1^f) \\ X_{\text{out}} &= \mathcal{G}(\mathcal{P}^{-1}(X_2)) + X_{\text{att}}, \end{aligned} \quad (7)$$

其中 $\mathcal{P}(\cdot)$ 和 $\mathcal{P}^{-1}(\cdot)$ 表示 JPEG 压缩方法中的补丁展开和折叠操作； \mathcal{G} 用 [19] 表示 GEGLU 函数。所提出的 DFFN 的详细网络架构如图 2(c) 所示。

3.3. 非对称编码器-解码器网络

我们将所提出的 FSAS 和 DFFN 嵌入到基于编码器和解码器架构的网络中。我们注意到大多数现有方法通常在编码器和解码器模块中使用对称架构。例如，如果 FSAS 和 DFFN 用于编码器模块，则它们也用于解码器模块。我们注意到编码器模块提取的特征是浅层特征，与解码器模块的深层特征相比，通常包含模糊效果。然而，模糊通常会改变两个相似块与清晰特征的相似性。因此，在编码器模块中使用 FSAS 可能无法正确估计相似度，从而影响图像恢复。为了克服这个问题，我们将 FSAS 嵌入到解码器模块中，这导致了非对称架构以获得更好的图像。

去模糊。图 2(a) 显示了所提出的非对称编码器-解码器网络的网络架构。

最后，给定一个模糊图像 B ，通过非对称编码器-解码器网络估计恢复图像 I ：

$$I = \mathcal{N}(B) + B, \quad (8)$$

其中 \mathcal{N} 表示非对称编码器-解码器网络。

四、实验结果

在本节中，我们评估我们的方法并将其与使用公共基准数据集的最先进方法进行比较。

4.1. 数据集和参数设置

数据集。我们在常用的图像去模糊数据集上评估我们的方法，包括 GoPro 数据集 [16]、HIDE 数据集 [20] 和 RealBlur 数据集 [18]。我们遵循现有方法的协议进行公平比较。

参数设置。我们使用与 [4] 相同的损失函数来约束网络，并使用具有默认参数的 Adam [10] 优化器对其进行训练。学习率的初始值为 10×3 ，并在 600,000 次迭代后使用余弦退火策略进行更新。学习率的最小值是 10×7 。补丁大小根据经验设置为 256×256 像素，批量大小设置为 16。我们在训练期间采用与 [30] 相同的数据增强方法。基于 JPEG 压缩方法，量化矩阵估计的补丁大小根据经验设置为 8×8 。在实现中，量化矩阵的张量形式设置为 $[\text{batch_size}, \text{channel_num}, 8, 8]$ ，其中 batch_size 和 channel_num 是批次和特征的数量。通过在训练期间求解损失函数，量化矩阵与其他参数共同学习。相似地，我们在计算自注意力时也使用 8×8 像素的补丁大小 (4)。由于页数限制，我们在补充材料中包含了更多的实验结果。培训代码和模型可在 <https://github.com/kkkls/FFTformer> 获得。

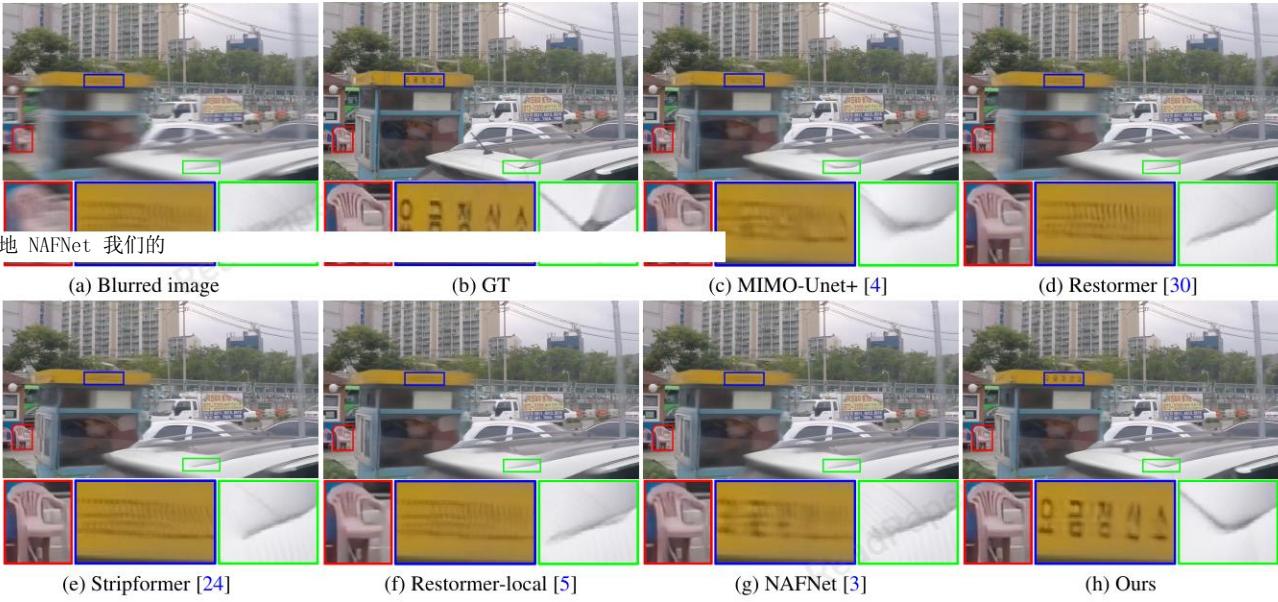


图 3. GoPro 数据集 [16] 的去模糊结果。 (c)–(g) 中的去模糊结果仍然包含显着的模糊效果。所提出的方法生成更清晰的图像。例如，人物和界限更加清晰。

表 1. GoPro 数据集 [16] 的定量评估。平均运行时间是在大小为 256×256 像素的图像上测试的。

Methods	PSNRs	SSIMs	Parameters (M)	Avg. runtime
DeblurGAN-v2 [12]	29.55	0.9340	60.9	0.04s
SRN [22]	30.26	0.9342	6.8	0.07s
DMPHN [32]	31.20	0.9453	21.7	0.21s
SAPHIN [21]	31.85	0.9480	23.0	-
MIMO-Unet+ [4]	32.45	0.9567	16.1	0.02s
MPRNet [31]	32.66	0.9589	20.1	0.09s
IPT [2]	-	-	114	0.50s
DeepRFT+ [15]	33.23	0.9632	23.0	0.09s
Restormer [30]	32.92	0.9611	26.1	0.08s
Uformer-B [27]	33.06	0.9670	50.9	0.07s
Stripformer [24]	33.08	0.9624	19.7	0.04s
MPRNet-local [5]	33.31	0.9637	20.1	0.11s
Restormer-local [5]	33.57	0.9656	26.1	0.42s
NAFNet [3]	33.71	0.9668	67.9	0.04s
Ours	34.21	0.9692	16.6	0.13s

4.2. Comparisons with the state of the arts

我们将我们的方法与最先进的方法进行比较，并使用 PSNR 和 SSIM 来评估恢复图像的质量。

GoPro 数据集的评估。我们首先通过 [16] 在常用的 GoPro 数据集上评估我们的方法。为了公平比较，我们遵循该数据集的协议并重新训练或调整未在该数据集上训练的深度学习方法。表 1 表示定量评价结果。我们的方法生成具有最高 PSNR 和 SSIM 值的结果。与最先进的基于 CNN 的方法 NAFNet [3] 相比，我们方法的 PSNR 增益至少比 NAFNet 高 0.5dB，而所提出的模型参数数量是 NAFNet 的四分之一。此外，与基于 Transformer 的方法 [24, 27, 30] 相比，我们的方法模型参数更少，但性能更好。由于 FFT 实现在 Py 中没有得到很好的优化-

表 2. 根据 PSNR 和 SSIM 对 RealBlur 数据集 [18] 的定量评估。

Methods	Realblur-R		Realblur-J	
	PSNRs	SSIMs	PSNRs	SSIMs
DeblurGAN-v2 [12]	36.44	0.9347	29.69	0.8703
SRN [22]	38.65	0.9652	31.38	0.9091
MIMO-Unet+ [4]	-	-	31.92	0.9190
BANet [23]	39.55	0.9710	32.00	0.9230
DeepRFT+ [15]	39.84	0.9721	32.19	0.9305
Stripformer [24]	39.84	0.9737	32.48	0.9290
Ours	40.11	0.9732	32.62	0.9326

Torch，在 PyTorch 中使用 FFT 需要更多的运行时间。然而，我们方法的运行时间仍然与最先进的方法相比具有竞争力。此外，我们的方法比基于原始自我注意的 IPT [2] 至少快 4×。图 3 显示了 GoPro 数据集上提出的方法和评估方法的视觉比较。正如 [30] 所证明的，基于 CNN 的方法 [3, 4] 不能有效地探索非局部信息以恢复潜在的清晰图像。因此，方法 [3, 4] 的去模糊结果仍然包含显着的模糊效果，如图 3 (c) 和 (g) 所示。基于 Transformer 的方法 [5, 24, 30] 能够为图像去模糊建模全局上下文。然而，一些主要结构，例如人物和椅子，并没有很好地恢复（见图 3 (d) – (f)）。

与现有的基于空间域的基于 Transformer 的方法相比，我们开发了一种高效的基于频域的 Transformer，其中所提出的 DFFN 能够有区别地估计有用的频率信息，用于潜在的清晰图像恢复。因此，去模糊后的结果包含清晰的结构，字符更加清晰，如图 3 (h) 所示。

对 RealBlur 数据集的评估。我们进一步评估

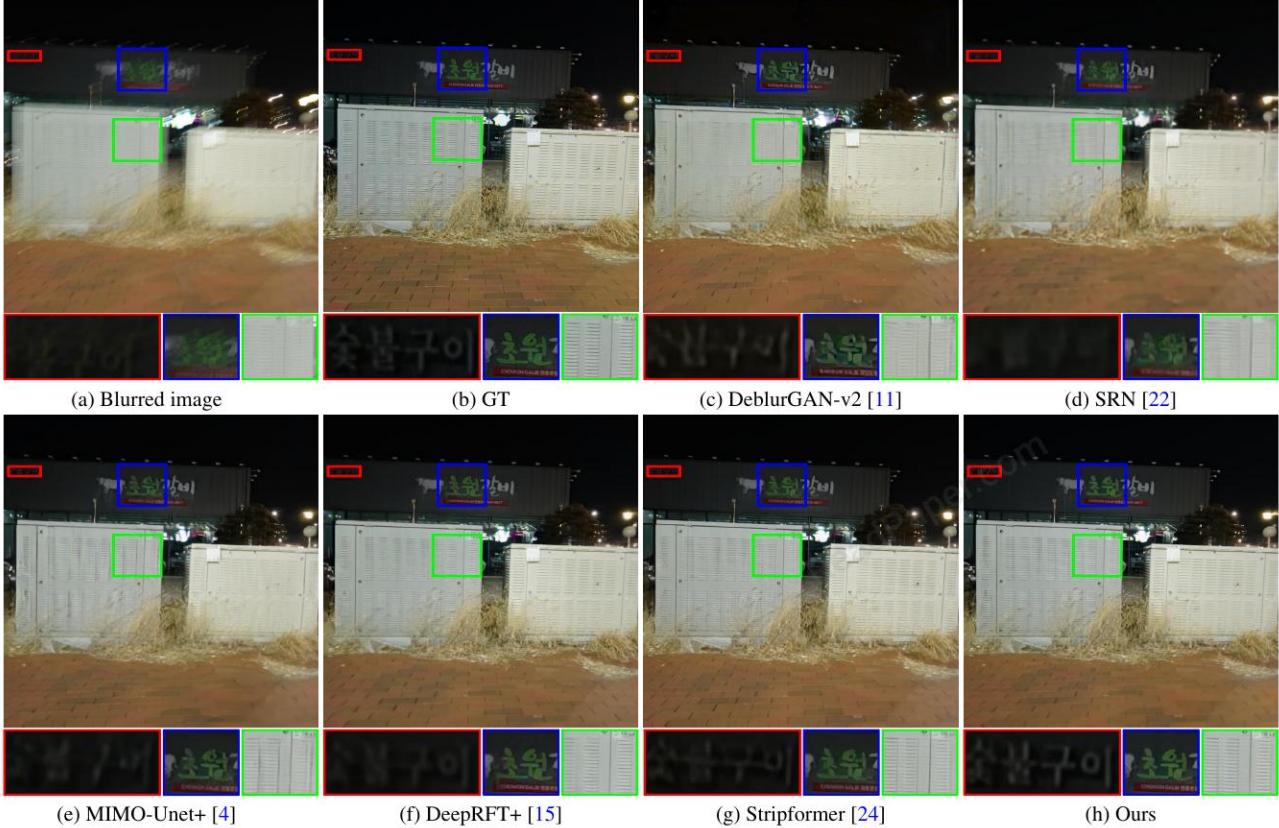


图 4. RealBlur 数据集 [18] 的去模糊结果。 (c)–(g) 中的字符或结构细节没有很好地恢复。所提出的方法生成的图像具有更清晰的特征和结构细节。

表 3. HIDE 数据集 [20] 的定量评估。我们使用在 GoPro 数据集 [16] 上训练的模型进行公平比较。

Methods	PSNRs	SSIMs	Parameters (M)
DeblurGAN-v2 [12]	26.61	0.8750	60.9
SRN [22]	28.36	0.9040	6.8
DMPHN [32]	29.09	0.9240	21.7
SAPHN [21]	29.98	0.9300	23.0
MIMO-Unet+ [4]	29.99	0.9304	16.1
MPRNet [31]	30.96	0.9397	20.1
Stripformer [24]	31.03	0.9395	19.7
MPRNet-local [5]	31.19	0.9418	20.1
Restormer [30]	31.22	0.9423	26.1
NAFNet [3]	31.31	0.9427	67.9
Restormer-local [5]	31.49	0.9447	26.1
Ours	31.62	0.9455	16.6

我们通过 [18] 在 RealBlur 数据集上使用我们的方法，并遵循该数据集的协议进行公平比较。[18] 的测试数据集包括来自原始图像的 RealBlur-R 测试集和来自 JPEG 图像的 RealBlur-J 测试集。表 2 总结了上述测试集的定量评估结果。所提出的方法生成具有更高 PSNR 值的结果。

图 4 显示了 RealBlur 数据集的视觉比较，其中我们的方法生成的结果具有更清晰的字符和更清晰的结构细节（图 4(h)）。

对 HIDE 数据集的评估。然后我们评估我们的

HIDE 数据集 [20] 上的方法，主要包含人类。与最先进的方法 [4, 31] 类似，我们直接使用评估方法的模型，这些模型在 GoPro 数据集上训练以进行测试。表 3 显示，所提出的方法生成的去模糊图像的质量优于评估方法，这表明我们的方法具有更好的泛化能力，因为模型没有在该数据集上进行训练。

我们在图 5 中展示了一些视觉比较。我们注意到评估的方法不能很好地恢复人类。相比之下，我们的方法生成更好的图像。例如，衣服的面和拉链就清晰多了。

五、分析与讨论

我们已经表明，在频域中探索 Transformers 的特性会产生优于最先进方法的有利结果。在本节中，我们对所提出的方法进行了更深入的分析，并展示了主要组件的效果。对于本节中的消融研究，我们使用 8 的批量大小在 GoPro 数据集上训练我们的方法和所有基线，以说明我们方法中每个组件的效果。

FSAS 的影响。所提出的 FSAS 用于降低计算成本。根据 FFT 的性质，FSAS 的空间和时间复杂度为 $O(N)$ 并且



图 5. HIDE 数据集 [20] 的去模糊结果。 (c)–(g) 中的去模糊结果仍然包含显着的模糊效果。所提出的方法生成更清晰的图像。

表 4. 基于我们的方法和窗口方法 [14, 27] 的 Transformer 的内存和运行时间比较。测试图像的大小为 1280×720 像素。测试环境基于配备 NVIDIA GeForce RTX 3090 GPU 的机器。“GPU 内存”表示由 “`torch.cuda.max_memory_allocated()`” 函数计算的最大 GPU 内存消耗。

Window size	Window-based method [27]		Ours	
	Avg. runtime	GPU memory	Avg. runtime	GPU memory
8×8	53ms	6.3G	44ms	6.5G
16×16	56ms	7.1G	44ms	6.2G
32×32	89ms	12.0G	43ms	6.0G
64×64	-	Out of memory	42ms	5.9G
1280×720	-	Out of memory	42ms	5.9G

$O(N C \log N)$ ，远低于最初计算缩放点积注意力时的 $O(N^2)$ 和 $O(N^2 C)$ ，其中 C 是特征数。我们进一步检查了 FSAS 的空间和时间复杂度以及

Transformers 的基于窗口的策略 [14, 27]。尽管频域中的张量包含实部和虚部，但内存消耗不会增加，因为我们在实现中基于 FFT 的共轭对称性仅通过 FFT 存储一半的张量。表 4 显示使用建议的 FSAS 需要所有 GPU 内存，并且与基于窗口的策略 [27] 相比效率更高。此外，由于 FSAS 的内存与窗口大小无关，因此当窗口变大时，FSAS 的内存使用量不会增加。

此外，由于所提出的 FSAS 是在频域中执行的，因此人们可能想知道在空间域中估计的缩放点积注意力是否表现更好。为了回答这个问题，我们比较了 F-

表 5. GoPro 数据集 [16] 上提出的方法中每个组件的定量评估。

	FSAS	Swin attention	FFN	DFFN	PSNRs/SSIMs
w/ only FFN	✗	✗	✓	✗	33.19/0.9626
w/ only DFFN	✗	✗	✗	✓	33.55/0.9651
SA w/ SD	✗	✓	✗	✓	33.46/0.9645
FSAS+FFN	✓	✗	✓	✗	33.61/0.9654
FSAS+DFFN	✓	✗	✗	✓	33.73/0.9663

具有在空间域中执行的基线方法的 SAS (简称 SA w/ SD)。由于原始缩放点积注意力的空间复杂度为 $O(N^2)$ ，因此在使用与建议的 FSAS 相同的设置时训练 SA w/ SD 是负担不起的。我们使用 Swin

Transformer [14] 进行比较，因为它效率更高。表 5 显示了 GoPro 数据集上的定量评估结果。在空间域中计算缩放点积注意力的方法不会产生良好的去模糊结果，其 PSNR 值低 0.27 (请参见表 5 中“SA w/ SD”和“FSAS+DFFN”的比较)。主要原因是虽然使用移位窗口划分方法降低了计算成本，但它并没有充分挖掘跨不同窗口的有用信息。相比之下，所提出的 FSAS 的空间复杂度为 $O(N)$ ，并且不需要将移动窗口划分作为近似值，从而获得更好的去模糊结果。图 6(b) 进一步表明，使用移位窗口分区方法作为空间域中缩放点积注意力的近似值并不能有效地消除模糊。相比之下，拟议的 FSAS 生成更清晰的图像。此外，与仅使用 FFN (w/ only FFN) 的基线方法相比，在此基线中使用建议的 FSAS 会产生更好的结果，其中 PSNR



图 6. 空间域和频域中缩放点积注意力计算的有效性。在频域中使用 FSAS 计算缩放点积注意力会生成更清晰的图像。



图 7. 所提出的 FSAS 对图像去模糊的有效性。使用建议的 FSAS 生成更清晰的图像。

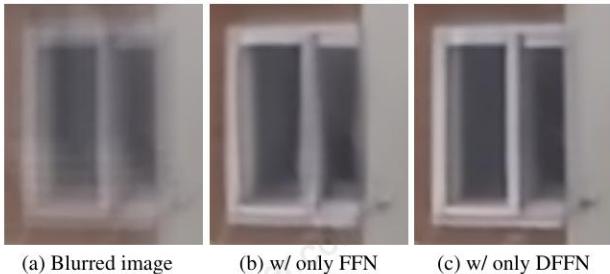


Figure 8. Effectiveness of the proposed DFFN on image deblurring.

值高出 0.42dB (参见表 5 中“仅 FFN”和“FSAS+FFN”的比较)。图 7(b) 和 (c) 中的视觉比较进一步表明，使用所提出的 FSAS 有助于很好地去除模糊，边界恢复得很好，如图 7(c) 所示。DFFN 的影响。所提出的 DFFN 用于有区别地估计潜在清晰图像恢复的有用频率信息。为了证明其对图像去模糊的有效性，我们将所提出的方法与两个基线进行比较。对于第一个基线，我们比较了仅使用 DFFN (简称 w/ only DFFN) 的建议方法和仅使用原始 FFN (简称 w/ only FFN) 的建议方法。对于第二个基线，我们将所提出的方法与在所提出的方法中用原始 FFN 替换 DFFN 的方法 (FSAS+FFN) 进行比较。表 5 中“w/ only DFFN”和“w/ only FFN”的比较表明，使用建议的 DFFN 会产生更好的结果，其中 PSNR 值高 0.36dB。

此外，表 5 中“FSAS+FFN”和“F-SAS+DFFN”的比较表明，使用建议的 DFFN 进一步提高了性能。图 8 显示了上述这些基线方法的可视化结果。使用建议的 DFFN 生成更好的去模糊图像，其中窗口是

表 6. GoPro 数据集上非对称编码器-解码器网络的定量评估

Methods	FSAS in enc&dec	FSAS in dec (Ours)
PSNRs	33.56	33.73
SSIMs	0.9653	0.9663



图 9. 非对称编码器-解码器网络对图像去模糊的有效性。

恢复良好如图 8(c) 所示。

非对称编解码器网络的影响。如3.3节所示，编码器模块提取的浅层特征通常包含影响FSAS估计的模糊效应。因此，我们将它嵌入到解码器模块中，从而形成一个非对称的编解码器网络，以更好地去模糊图像。为了检验这种网络设计的效果，我们比较了将FSAs放入编码器和解码器模块的网络（表6中的ENC和DEC中的FSAs）。表6显示，在解码器模块中使用FSAS会产生更好的结果，其中PSNR值至少高出0.17dB。图9(b)和(c)中的视觉比较进一步表明，在解码器模块中使用FSAS可以生成更清晰的图像。

六，结论

受卷积定理的启发，我们提出了一种有效且高效的方法来探索用于高质量图像去模糊的 Transformer 的属性。我们开发了一种高效的基于频域的自注意力求解器 (FSAS)，通过逐元素乘积运算而不是空间域中的矩阵乘法来估计缩放的点积注意力，我们展示了空间复杂度并且计算复杂度显着降低。我们进一步提出了一个 DFFN 来有区别地确定特征的哪些低频和高频信息应该被保留以用于潜在的清晰图像恢复。此外，我们开发了一个基于编码器和解码器架构的非对称网络，其中 FSAS 仅用于解码器模块以实现更好的图像去模糊。通过以端到端的方式训练我们的方法，我们表明它在准确性和效率方面优于最先进的方法。致谢。该工作得到了国家重点研发计划 (No. 2018AAA0102001)、国家自然科学基金 (Nos. U22B2049、62272233、61922043、U19B2040) 和基础研究基金的部分支持。中央高校 (编号：30920041109)。

References

- [1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2
- [2] 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*, 2021. 2, 5
- [3] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 2, 5, 6, 7
- [4] 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*, 2021. 1, 2, 4, 5, 6
- [5] Xiaojie Chu, Liangyu Chen, , Chengpeng Chen, and Xin Lu. Improving image restoration by revisiting global information aggregation. In *ECCV*, 2022. 5, 6, 7
- [6] Jiangxin Dong, Jinshan Pan, Jimmy S. Ren, Liang Lin, Jinhui Tang, and Ming-Hsuan Yang. Learning spatially variant linear representation models for joint filtering. *IEEE TPAMI*, 44(11):8355–8370, 2022. 1
- [7] Jiangxin Dong, Stefan Roth, and Bernt Schiele. Learning spatially-variant MAP models for non-blind image deblurring. In *CVPR*, 2021. 1
- [8] Jiangxin Dong, Stefan Roth, and Bernt Schiele. DWDN: deep wiener deconvolution network for non-blind image deblurring. *IEEE TPAMI*, 44(12):9960–9976, 2022. 1
- [9] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *CVPR*, 2019. 2
- [10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 4
- [11] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 1, 6
- [12] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, 2019. 1, 5, 6
- [13] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV Workshops*, 2021. 2
- [14] 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 *ICCV*, 2021. 2, 7
- [15] Xintian Mao, Yiming Liu, Wei Shen, Qingli Li, and Yan Wang. Deep residual fourier transformation for single image deblurring. *arXiv preprint arXiv:2111.11745*, 2021. 5, 6
- [16] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 1, 2, 4, 5, 6, 7
- [17] Jinshan Pan, Jiangxin Dong, Yang Liu, Jiawei Zhang, Jimmy S. J. Ren, Jinhui Tang, Yu-Wing Tai, and Ming-Hsuan Yang. Physics-based generative adversarial models for image restoration and beyond. *IEEE TPAMI*, 43(7):2449–2462, 2021. 1
- [18] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking de-blurring algorithms. In *ECCV*, 2020. 4, 5, 6
- [19] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 4
- [20] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *ICCV*, 2019. 4, 6, 7
- [21] Maitreya Suin, Kuldeep Purohit, and A. N. Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*, 2020. 5, 6
- [22] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018. 1, 2, 5, 6
- [23] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Banet: Blur-aware attention networks for dynamic scene deblurring. In *CVPR*, 2021. 5
- [24] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *ECCV*, 2022. 2, 3, 5, 6, 7
- [25] 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*, 2017. 2
- [26] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 2
- [27] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022. 1, 2, 3, 5, 7
- [28] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *CVPR*, 2022. 2
- [29] Weijian Xu, Yifan Xu, Tyler A. Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *ICCV*, 2021. 2, 3
- [30] 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 *CVPR*, 2022. 1, 2, 4, 5, 6, 7
- [31] 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*, 2021. 1, 2, 5, 6, 7
- [32] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, 2019. 1, 2, 5, 6
- [33] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson W. H. Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *CVPR*, 2018. 1
- [34] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. Learning generative vision transformer with energy-based latent space for saliency prediction. In *NeurIPS*, 2021. 2

- [35] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 2