SRFormer: 用于单图像超分辨率的置换自注意力

周宇鹏 [1] 李震 [1] 郭春乐 [1] 柏松 [2] 程明明 [1] 侯淇彬 [1*]

[1] 天津市媒体计算工程研究中心, 南开大学 [2] 字节跳动公司, 新加坡

https://github.com/HVision-NKU/SRFormer

摘要

先前的研究表明，对于基于 *Transformer* 的图像超分辨率模型 *(*如 *SwinIR)*，增加其窗口大小可以显着提高模型性能，但是也带来了更大的计算开销。在本文中，我们提出了一种简单而新颖的方法 *SRFormer*，它可以在引入更少的额外计算条件下，享受大窗口自注意力的好处。我们的 *SRFormer* 的核心是置换自注意力 *(PSA)*，它在自注意的通道和空间信息之间取得了适当的平衡。*PSA* 可以简单且容易地被应用于现有的基于窗口自注意力的超分辨率网络。没有任何花哨的东西，我们的 *SRFormer* 在使用更少的参数和计算的条件下，在 *Urban100* 数据集上的 *PSNR* 得分为 *33.86dB*，比 *SwinIR* 高 *0.46dB*。我们希望我们简单而有效的方法能够成为一种可供今后超分辨率模型设计的研究提供参考的工具。

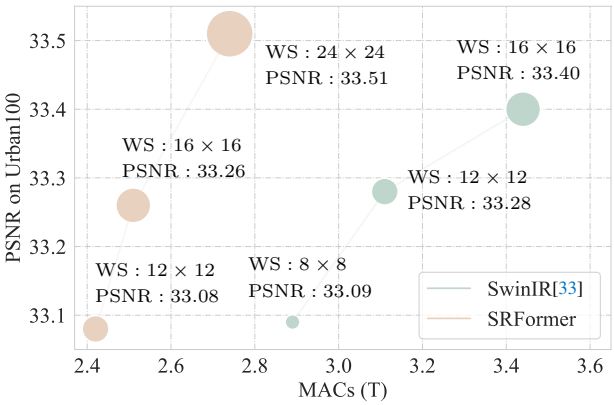

图 1: 用不同的窗口大小 (WS) 训练 200k 次迭代时，SwinIR 和我们的 SRFormer 的性能比较。窗口大小为 24 × 24 的 SRFormer 甚至以更少的计算量获得了更高的 PSNR 分数。

1. 引言

单图像超分辨率 (**S**uper-**R**esolution, SR) 旨在将图像从退化的低分辨率版本恢复成原高分辨率版本。探索高效的超分辨算法一直是计算机视觉领域的研究热点，具有多种应用 [23, 59, 2]。从几个先驱性的工作开始 [9, 24, 75, 28, 48, 34]，基于 CNN 的方法在很长一段时间内是图像超分辨率的主流。得益于残差学习 [51, 28, 34, 24, 73, 30]，密集连接 [63, 79, 54]，或通道注意力 [78, 66]，这些方法设计的网络结构为超分辨率的模型发展做出了巨大贡献。

尽管基于 CNN 的超分辨率模型取得了成功，但最近的工作 [5, 33, 72, 77] 表明基于 Transformer 的模型表现得更好。他们观察到，与卷积相比，利用自注意力建立成对关系的能力能够更有效地产生高质量超分辨率图像。其中一个典型的工作是 SwinIR [33]，它将 Swin Transformer [37] 引入图像超分辨率，在各种基准上相比最先进的基于 CNN 的模型获得了很大的提升。后来，各种工作，如 SwinFIR [72], ELAN [77] 和 HAT [6]，进一步改进了 SwinIR，并使用 Transformer 为 SR 任务设计不同的网络架构。

上述方法表明，在 SwinIR 中适当扩大移位窗口自注意力的窗口可以获得明显的性能增益 (参见图 1)。然而，随着窗口尺寸的增大，计算负担也是一个重要问题。此外，基于 Transformer 的方法使用用了自注意力，并且与之前基于 CNN 的方法 [78, 79, 24] 相比，需要更

*Corresponding author.

大通道数的网络。为了探索高效有效的超分辨率算法，一个直接的想法是: 如果我们在减少通道数量的同时增加窗口大小，性能将如何变化?

受上述想法的启发，本文提出置换自注意力 (**P**ermuted **S**elf-**A**ttention, PSA)，一种在大窗口 (例如 24×24) 中构建成对关系的有效方法。PSA 的目的是使更多的像素参与注意力图的计算，同时不引入额外的计算负担。为此，本文提出缩小键值矩阵的通道维度并采用置换操作将部分空间信息传递到信道维度中。这样，尽管减少了通道数，但没有空间信息的损失，并且每个注意力头也被允许保持适当数量的通道来生成表达多样的注意力图 [56]。此外，本文还对原始的前馈网络 (**F**eed-**F**orward **N**etwork, FFN) 进行了改进。我们发现，在两个线性层之间添加了深度可分离卷积有助于高频成分的恢复。

基于所提出的 PSA，我们为图像超分辨率任务构建了一个新的网络，称为 SRFormer。我们在五个广泛使用的数据集上评估了 SRFormer。得益于提出的 PSA，我们的 SRFormer 可以明显提高在几乎所有五个数据集上的性能。值得注意的是，对于 $\times 2$ 超分辨率任务，我们的 SRFormer 仅在 DIV2K 数据集 [34] 上训练，在具有挑战性的 Urban100 数据集 [18] 上取得了 33.86 的 PSNR 分数。我们的结果远高于最近的 SwinIR(33.40) 和 ELAN(33.44)。在 $\times 3$ 和 $\times 4$ 超分辨率任务上也可以观察到类似的现象。此外，我们还设计了轻量版的 SRFormer 进行实验。与先前的轻量级 SR 模型相比，所提出方法在所有基准上都取得了更好的性能。

综上所述，我们的贡献可以概括如下:

- 本文提出了一种新的基于置换自注意力的图像超分辨率算法，通过将空间信息转移到通道维度来获得大窗口的自注意力。利用它，我们第一个在超分辨率任务中中以可接受的时间复杂度实现了 24x24 大窗口注意力机制。

- 基于提出的 PSA 和从频率角度改进 FFN (ConvFFN)，我们构建了一个新的基于 Transformer 的超分辨率网络 SRFormer。我们的 SRFormer 在经典的、轻量级的和真实世界的图像超分辨率任务中获得了最佳性能。

2. 相关工作

在本节中，我们简要回顾了图像超分辨率的相关文献。首先介绍基于 CNN 的方法，然后是近流行的基于 Transformer 的模型。

2.1. 基于 CNN 的图像超分辨率模型

自 SRCNN [9] 首次将 CNN 引入图像 SR 以来，涌现了大量基于 CNN 的 SR 模型。DRCN [25] 和 DRRN [51] 引入循环卷积网络，在不增加参数的情况下增加网络的深度。一些早期基于 CNN 的方法 [52, 9, 25, 51] 试图将低分辨率 (**L**ow-**R**esolution, LR) 的插值结果作为输入，这导致特征提取的计算成本很高。为了加速 SR 推理过程，FSRCNN [10] 在 LR 尺度上提取特征，并在网络末端进行上采样操作。上采样模块 pixel shuffle[48] 在后来的工作中中被广泛使用 [77, 78, 33]。LapSRN [27] 和 DBPN [17] 在特征提取过程中进行上采样，以学习 LR 和 HR 之间的相关性。还有一些工作 [28, 63, 76, 61] 使用 GAN [14] 在重建中生成逼真的纹理。MemNet [52]、RDN [79] 和 HAN [45] 有效地聚合了中间特征，以增强重建图像的质量。非局部注意力 [60] 也在 SR 中被广泛探索，以更好地对长程依赖性进行建模，包括 CS-NL [44], NLSA [43], SAN [7], IGNN [81] 等等。

2.2. Vision Transformers

Transformer 最近在一系列任务中展现出了巨大的潜力，包含图像分类 [11, 55, 67, 58, 68]，目标检测 [4, 50, 12]，语义分割 [65, 80, 49]，图像恢复 [69, 33, 5] 等等。其中，最典型的工作应该是视觉 Transformer (ViT) [11]，它证明 Transformer 可以在特征编码方面优于卷积神经网络。Transformer 在 low-level 视觉中的应用主要包括两类: 生成 [22, 29, 71, 8] 和恢复。进一步地，恢复任务也可以分为两类: 视频恢复 [38, 36, 47, 35, 13] 和图像恢复 [5, 69, 64, 16]。

图像超分辨率作为图像恢复中的一项重要任务，需要保留输入的结构信息，这对基于 Transformer 的模型设计提出了巨大挑战。IPT [5] 是一个基于 Transformer 编码器和解码器结构的大型预训练模型，已被应用于超分辨率、去噪和除雨。基于 Swin Transformer 编码器 [37], SwinIR [33] 在特征提取中对 8×8 局部窗口应用自注意力，并获得了极其强大的性能。ELAN [77] 简化了 SwinIR 的架构，并使用在不同窗口大小中计算的自

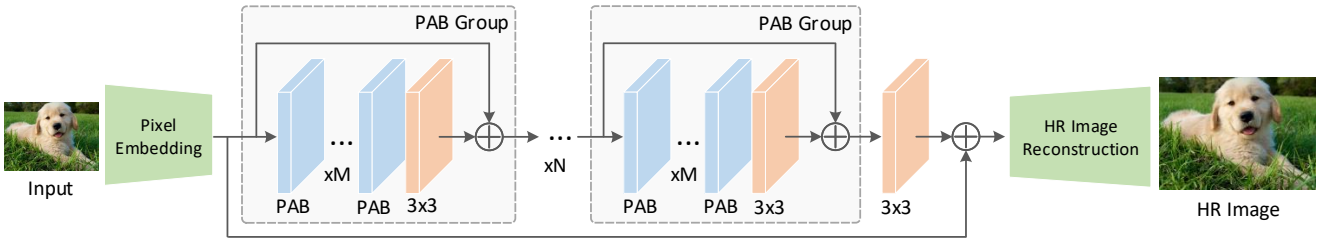

图 2: SRFormer 的整体架构。像素嵌入模块 (Pixel Embedding) 是一个 3×3 卷积，将输入图像映射到特征空间。HR 图像重建模块 (HR Image Reconstruction) 包含一个 3×3 卷积和一个像素混洗操作来重建高分辨率图像。中间的特征编码部分有 N 组 PAB，后接 3×3 卷积。

注意力来收集远程像素之间的相关性。

本文的 SRFormer 也是基于 Transformer 的。与前面提到的直接利用自注意力来构建模型的方法不同，我们的 SRFormer 主要针对自注意力本身。本文的目的是研究如何在一个大窗口中计算自注意力，以在不增加参数和计算成本的情况下提高 SR 模型的性能。

3. 方法

3.1. 整体架构

SRFormer 的整体架构如图 2所示，由三部分组成: 像素嵌入层 G_P、特征编码器 G_E 和高分辨率图像重建层 G_R。同先前的工作 [33, 77] 一样，像素嵌入层 G_P 是一个单独的 3×3 卷积，它将低分辨率 RGB 图像 $I \in \mathbb{R}^{H \times W \times 3}$ 转换为特征嵌入 $F_P \in \mathbb{R}^{H \times W \times C}$。$F_P$ 被送入具有层次结构的特征编码器 G_E。该编码器由 N 个置换自注意力组组成，每个组包含 M 个置换自注意力块以及一个 3×3 卷积。在特征编码器的末尾添加了一个 3×3 卷积，结果为 F_E。将 F_E 和 F_P 的和输入到 G_R 中进行高分辨率图像重建，G_R 包含 3×3 卷积和亚像素卷积层 [48]。计算高分辨率重建图像和真实 HR 图像之间的 L1 损失以优化 SRFormer。

3.2. 置换自注意力块

SRFormer 的核心是置换自注意力块 (**P**ermuted self-**A**ttention **B**lock, PAB)，它由 PSA 层和卷积前馈网络 (ConvFFN) 组成。

置换自注意力。如图 3(b) 所示，给定输入特征 $\mathrm{X}_{in} \in \mathbb{R}^{H \times W \times C}$ 和一个 token 缩减因子 r，我们首先将 X_{in} 分割成 N 个不重叠的正方形窗口 $\mathrm{X} \in \mathbb{R}^{NS^2 \times C}$，其中 S 是每个窗口的边长。然后，我们使用 3 个线性层 L_Q L_K L_V 来得到 Q、K 和 V:

$$\mathrm{Q}, \mathrm{K}, \mathrm{V} = L_Q(\mathrm{X}), L_K(\mathrm{X}), L_V(\mathrm{X}). \tag{1}$$

其中，Q 保持与 X 相同的通道维数，而 L_K 和 L_V 将通道数压缩为 C/r^2，得到 $\mathrm{K} \in \mathbb{R}^{NS^2 \times C/r^2}$ 和 $\mathrm{V} \in \mathbb{R}^{NS^2 \times C/r^2}$。之后，为了使更多的 token 参与自注意力的计算，同时避免计算成本的增加，我们提出将 K 和 V 中的空间 token 置换到通道维度上，以获得置换后的 token $\mathrm{K}_p \in \mathbb{R}^{NS^2/r^2 \times C}$ 和 $\mathrm{V}_p \in \mathbb{R}^{NS^2/r^2 \times C}$

我们使用 Q 与缩减后的 K_p 和 V_p 来执行自注意力操作。K_p 和 V_p 窗口大小为 $\frac{S}{r} \times \frac{S}{r}$，但它们的通道维度仍然为 C，以保证每个注意力头生成的注意力图的表现力 [56]。所提出的 PSA 的计算可以形式化为:

$$\mathrm{PSA}(\mathrm{Q}, \mathrm{K}_p, \mathrm{V}_p) = \mathrm{Softmax}\left(\frac{\mathrm{Q}\mathrm{K}_p^T}{\sqrt{d_k}} + \mathrm{B}\right)\mathrm{V}_p, \tag{2}$$

因为 Q 的窗口大小与 K_p 的窗口大小不匹配，可以通过插值 [37] 中定义的位置嵌入来获得对齐的相对位置嵌入 B。$\sqrt{d_k}$ 是在 [11] 中定义的标量。通过将通道分成多个组，上式可以很容易地转换为多头版本。PSA 将空间信息传输到通道维度，确保了以下两个关键设计原则：i) 我们不像在 [65, 58] 中那样首先对 token 进行降采样，而是允许每个 token 独立参与自注意力计算，这将使我们获得更多有表征能力的注意力图。我们将在 3.3节中讨论 PSA 的更多变体，并在实验部分展示其效果。ii) 与图 3(a) 中的原始自注意力相比，PSA 可以在大窗口 (例如 24×24) 设置下执行，使用的计算量甚至比 SwinIR 的 8×8 窗口更少，同时获得更好的性能。

ConvFFN. 先前的工作表明，自注意力可以被视为一个低通滤波器 [46, 57]。为了更好地恢复高频信息，通常在每组 Transformer 的末尾添加 3×3 卷积，就像在 SwinIR

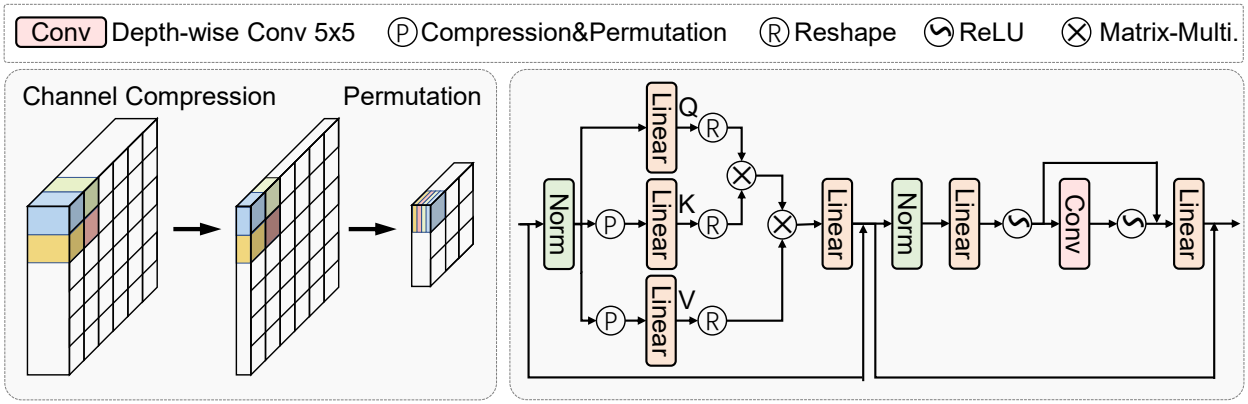

图 3: 我们提出的置换自注意力块的结构示意图。左侧部分是我们提出的减少通道参数并转移空域信息到通道维度来避免空域信息损失的操作。

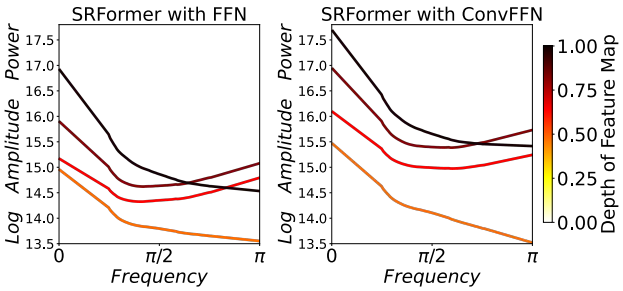

图 4: SRFormer 中 FFN 和 ConvFFN 产生的中间特征图的功率谱。颜色越深的线条对应越深层的特征。

[33] 中所做的那样。而与 SwinIR 不同的是，在 PAB 中，FFN 块的两个线性层之间添加了一个局部深度可分离卷积分支，以协助编码更多的细节。我们将这样修改后的块称为 ConvFFN。如图 4所示，我们在实验中发现，这样的操作增加计算量几乎可以忽略，但可以补偿由自注意力造成的高频信息损失。我们简单地计算了分别由 FFN 和 ConvFFN 得到特征图的功率谱，经对比，ConvFFN 明显地增加了高频信息，因而如表 1所示，获得了更好的结果。

3.3. 大窗口自注意力的变体

为了给大窗口自注意力的设计提供指导，并证明 PSA 的优势，本文提出另外两种大窗口自注意力的变体。定量的对比和分析记录在实验部分。

Token 缩减。引入大窗口自注意力并避免计算成本增加的第一种方法是减少 token 的数量，正如在 [65] 中所做的那样。设 r 和 S 分别为缩减因子和窗口大小。给定输入 $\mathrm{X} \in \mathbb{R}^{NS^2 \times C}$，使用卷积核大小为 $r \times r$、步长为 r 的深度可分离卷积来将 K 和 V 的每一个窗口的 token 数量缩减为 $(\frac{S}{r})^2$，有 $\mathrm{Q}_r \in \mathbb{R}^{NS^2 \times C}$ 和 $\mathrm{K}_r, \mathrm{V}_r \in \mathbb{R}^{NS^2/r^2 \times C}$。$\mathrm{Q}_r$ 和 K_r 用于计算注意力图 $\mathrm{A} \in \mathbb{R}^{S^2 \times S^2/r^2}$。A 与 V_r 的矩阵乘得到与 X 形状相同的输出。

Token 采样。实现大窗口自注意力的第二种方法是根据给定的采样比 T，从键 K 和值 V 的每个窗口中随机采样 T^2 $(0 \leq T \leq S)$ 个 token。给定输入 $\mathrm{X} \in \mathbb{R}^{NS^2 \times C}$，Q 与 X 形状相同，而 K 和 V 被缩减到 $NT^2 \times C$。在 T 是定值的前提下，计算成本与窗口尺寸成线性关系。然而，该方法对 token 的部分随机选择会丢失场景内容的结构信息，不利于图像超分辨率。我们将在实验部分展示更多的定量结果。

4. 实验

在本节中，我们在在经典的、轻量级的和真实世界的图像 SR 任务上进行了实验，将 SRFormer 与现有的最先进方法进行了比较，并对 SRFormer 进行了消融分析。

4.1. 实验设置

数据集与评价指标。训练数据集的选择与被对比的模型保持一致。在经典图像 SR 任务中，我们使用 DIV2K [34] 和 DF2K (DIV2K [34] + Flickr2K [53]) 分别训练两个版本的 SRFormer。在轻量级图像 SR 任务中，我们使用 DIV2K [34] 来训练 SRFormer -light。在真实世界 SR 任务中，我们使用 DF2K 和 OST [62] 来训练。模型在 5 个基准数据集上进行了测试，包括 Set5 [3], Set14 [70], BSD100 [41], Urban100 [18], and Manga109 [42]。为了进一步提高性能，引入了自集成策略，对应的模型为

方法	窗口大小	参数量	MACs	SET5 [3]		SET14 [70]		B100 [41]		Urban100 [18]		Manga109 [42]	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR [33]	8×8	11.75M	2868G	38.24	0.9615	33.94	0.9212	32.39	0.9023	33.09	0.9373	39.34	0.9784
	12×12	11.82M	3107G	38.30	0.9617	34.04	0.9220	32.42	0.9026	33.28	0.9381	39.44	0.9788
	16×16	11.91M	3441G	38.32	0.9618	34.00	0.9212	32.44	0.9030	33.40	0.9394	39.53	0.9791
SRFormer w/o **ConvFFN**	12×12	9.97M	2381G	38.23	0.9615	34.00	0.9216	32.37	0.9023	32.99	0.9367	39.30	0.9786
	16×16	9.99M	2465G	38.25	0.9616	33.98	0.9209	32.38	0.9022	33.09	0.9371	39.42	0.9789
	24×24	10.06M	2703G	38.30	0.9618	34.08	0.9225	32.43	0.9030	33.38	0.9397	39.44	0.9786
SRFormer	12×12	10.31M	2419G	38.22	0.9614	34.08	0.9220	32.38	0.9025	33.08	0.9372	39.13	0.9780
	16×16	10.33M	2502G	38.31	0.9617	34.10	0.9217	32.43	0.9026	33.26	0.9385	39.36	0.9785
	24×24	10.40M	2741G	38.33	0.9618	34.13	0.9228	32.44	0.9030	33.51	0.9405	39.49	0.9788

表 1: 关于窗口大小的消融实验。我们分别展示了原始 SwinIR、没有 ConvFFN 的 SRFormer，以及完整的 SRFormer 的结果。请注意，窗口大小为 24×24 的 SRFormer 的参数和 MACs 比窗口大小为 8×8 的 SwinIR 要少，而更大的窗口带来了更好的性能。

ConvFFN	Urban100 [18]		Manga109 [42]	
	PSNR	SSIM	PSNR	SSIM
w/o Depth-wise Conv	33.38	0.9397	39.44	0.9786
3×3 Depth-wise Conv	33.42	0.9398	39.34	0.9787
5×5 Depth-wise Conv	**33.51**	**0.9405**	**39.49**	**0.9788**

表 2: 在 $\times 2$ SR 任务上关于 ConvFFN 的消融研究。从 Urban100 和 Manga109 的结果中，我们可以看到使用 5×5 深度可分离卷积产生了最好的结果。这表明局部细节对于基于 Transformer 的模型也是必不可少的。

方法	参数量	MACs	S	r	PSNR	SSIM
SwinIR [33]	11.75M	2868G	8	-	33.09	0.9373
Token 缩减	11,78M	2471G	16	2	33.09	0.9372
Token 缩减	11.85M	2709G	24	2	33.24	0.9387
Token 采样	11.91M	2465G	16	2	32.38	0.9312
Token 采样	12.18M	2703G	24	2	32.34	0.9305
PSA	9.99M	2465G	16	2	33.09	0.9371
PSA	10.06M	2703G	24	2	33.38	0.9397

表 3: 在 $\times 2$ SR 任务上，SwinIR [33]、PSA 和两个变体在 Urban100 [18] 上的性能比较。这里报告的时在 DIV2K 上训练 200k 次迭代内的最佳模型的结果。对于 token 采样，$r = S/T$。PSA 的性能优于另外两种变体。

SRFormer+。模型的评价指标是在 YCbCr 空间的 Y 通道上计算的的 PSNR 和 SSIM。

实现细节。在经典的图像 SR 任务中，我们将 PAB 组数量、每组 PAB 数量、通道数量和注意力头数分别设置为 6、6、180 和 6。在 DIV2K [34] 上训练时，patch 大小、窗口大小 S 和缩减因子 r 分别设置为 48×48、24 和 2；而在 DF2K [34, 53] 上训练时，这些参数分别设置为 64×64、22 和 2。对于轻量级图像 SR 任务，PAB 组数、每组 PAB 数量、通道数量、窗口大小 S、缩减因子 r 和注意力头数分别为 4、6、60、16、2 和 6，而 patch 大小是 64×64。按 90°、180° 或 270° 随机旋转图像，并随机水平翻转图像以增强数据。采用 $\beta_1 = 0.9$ 和 $\beta_2 = 0.99$ 的 Adam [26] 优化器来训练模型 50 000 次迭代。初始学习率设置为 2×10^{-4}，并在 $\{250\text{k}, 400\text{k}, 450\text{k}, 475\text{k}\}$ 轮迭代开始时减少为之前的一半。

4.2. 消融实验

PSA 中窗口尺寸的影响。置换自注意力使得放大窗口尺寸更加高效。为了研究不同窗口大小对模型性能的影响，我们进行了三组实验，报告在在表 1中。第一组实验是对于原始 SwinIR[33] 分别设置 8×8、12×12 和 16×16 的窗口大小。第二组实验是对使用 PSA，而未使用 ConvFFN 的 SRFormer，分别设置 12×12、16×16 和 24×24，以观察性能差异。第三组实验是对完整的 SRFormer 分别设置 12×12, 16×16, 和 24×24 的窗口大小来观察性能变化。结果表明，对于上述三组实验，更大的窗口大小都能获得更好的性能。此外，SRFormer (24×24 窗口大小) 的参数和 MACs 比原来的 SwinIR (窗口大小) 还要少。为了平衡性能和 MACs，我们将 SRFormer 和 SRFormer-light 的窗口大小分别设置 24×24 和 16×16。

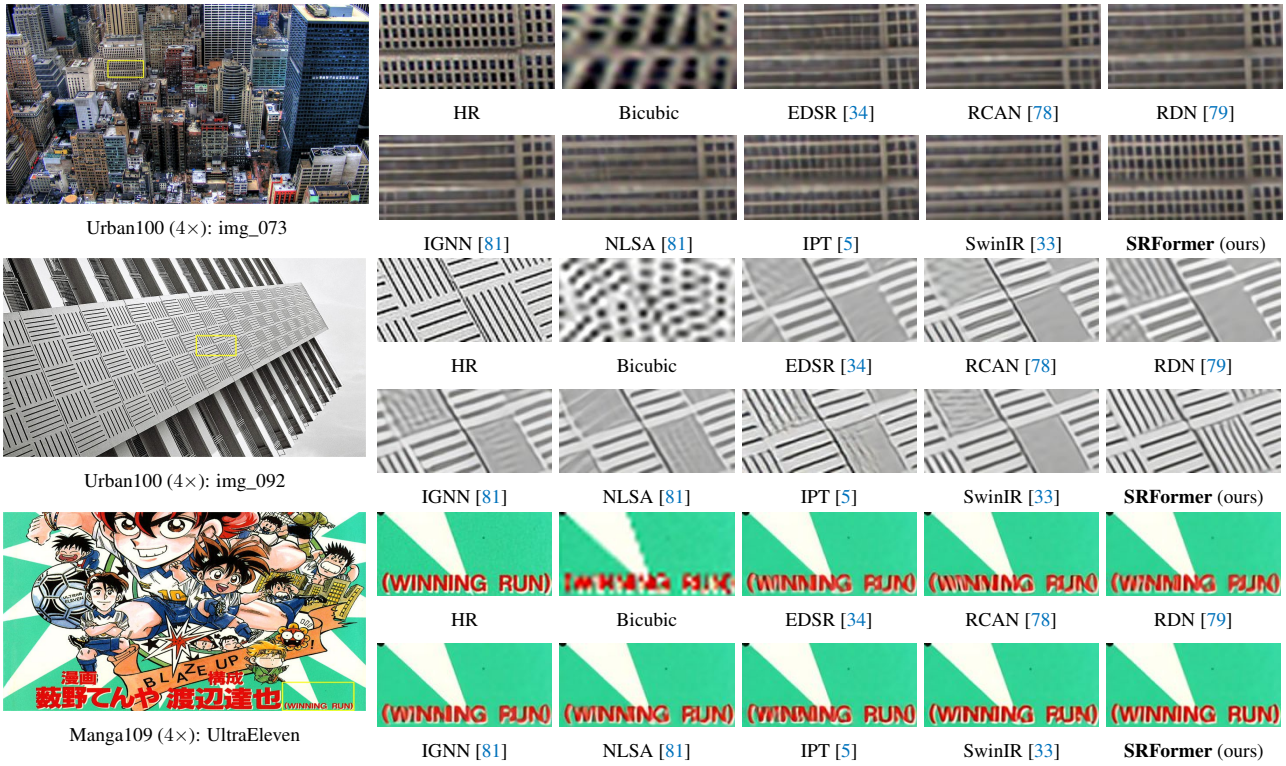

图 5: 在 ×4 SR 任务上与最近最先进的**经典图像 SR** 方法进行定性比较。

ConvFFN 的卷积核大小的影响。我们在 3.2节中引入了 ConvFFN，旨在不增加太多的计算的前提下编码更多的局部信息。为了探索哪种卷积核大小可以带来最好的性能提升，我们分别尝试使用 3×3 和 5×5 的深度可分离卷积，并在表 2中报告结果。考虑到深度可分离卷积对参数量和 MACs 影响不大，我们没有在表中列出它们。显然，5×5 的深度可分离卷积可以得到最好的结果，因此最终被用在 ConvFFN 中。

大窗口自注意力变体。在 3.3节中，我们引入了另外两种大窗口自注意力机制变体。我们将结果总结在表 3中。虽然 token 缩减可以在使用大窗口时轻微改善 SwinIR，但是参数数量没有减少且性能增益低于我们的方法。本文认为，直接对键和值进行下采样操作可能会导致空间信息损失。对于 token 采样，性能甚至比原始的 SwinIR 更差，这可能是因为删除一些 token 严重破坏了图像内容结构。

4.3. 经典图像超分辨率

对于经典图像 SR 任务，我们将 SRFormer 与一系列基于 CNN 和基于 Transformer 的最先进的 SR 方法进行比较：RCAN [78], RDN [79], SAN [7], IGNN [81], HAN [45], NLSA [43], IPT [5], SwinIR [33], EDT [31], 以及 ELAN [77].

定量比较。经典图像超分辨率任务上不同方法的定量比较如表 4所示。为了比较的公平性，SRFormer 的参数量和 MACs 都低于 SwinIR [33]，详见补充材料。可以清楚地看到，SRFormer 在几乎所有五个基准数据集上的所有比例因子上都取得了最佳性能。由于在大窗口内计算自注意力可以让更多的信息聚集在大区域上，因此 SRFormer 在高分辨率测试集 (如 Urban100 和 Manga109) 上的表现好得多。特别地，对于使用 DIV2K 的 ×2 SR 训练，我们的 SRFormer 在 Urban100 数据集上取得了 33.86dB 的 PSNR 分数，比 SwinIR 高 0.46dB，但使用的参数和计算量更少。对于引入集成策略的 SRFormer+，性能提升会更大。这些均证明 SRFormer 是更高效的。

定性比较。我们在图 5中展示了 SRFormer 与其他方法的定性比较。从图 5的前两个例子中，可以清楚地观察到 SRFormer 可以恢复清晰和详细的纹理和边缘。相比之下，其他模型只能恢复模糊或低质量的纹理。在第三个例子中，SRFormer 是唯一能还原每个字母的模型。定性比较表明 SRFormer 可以从低分辨率图像中恢复出

	方法	训练集	SET5 [3]		SET14 [70]		B100 [41]		Urban100 [18]		Manga109 [42]	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2 SR	**EDSR** [34]	DIV2K	38.11	0.9602	33.92	0.9195	32.32	0.9013	32.93	0.9351	39.10	0.9773
	RCAN [78]	DIV2K	38.27	0.9614	34.12	0.9216	32.41	0.9027	33.34	0.9384	39.44	0.9786
	SAN [7]	DIV2K	38.31	0.9620	34.07	0.9213	32.42	0.9028	33.10	0.9370	39.32	0.9792
	IGNN [81]	DIV2K	38.24	0.9613	34.07	0.9217	32.41	0.9025	33.23	0.9383	39.35	0.9786
	HAN [45]	DIV2K	38.27	0.9614	34.16	0.9217	32.41	0.9027	33.35	0.9385	39.46	0.9785
	NLSA [43]	DIV2K	38.34	0.9618	34.08	0.9231	32.43	0.9027	33.42	0.9394	39.59	0.9789
	SwinIR [33]	DIV2K	38.35	0.9620	34.14	0.9227	32.44	0.9030	33.40	0.9393	39.60	0.9792
	ELAN [77]	DIV2K	38.36	0.9620	34.20	0.9228	32.45	0.9030	33.44	0.9391	39.62	**0.9793**
	SRFormer (ours)	DIV2K	**38.45**	**0.9622**	**34.21**	**0.9236**	**32.51**	**0.9038**	**33.86**	**0.9426**	**39.69**	0.9786
	IPT [5]	ImageNet	38.37	-	34.43	-	32.48	-	33.76	-	-	-
	SwinIR [33]	DF2K	38.42	0.9623	34.46	0.9250	32.53	0.9041	33.81	0.9427	39.92	0.9797
	EDT [31]	DF2K	38.45	0.9624	34.57	0.9258	32.52	0.9041	33.80	0.9425	39.93	0.9800
	SRFormer (ours)	DF2K	38.51	0.9627	34.44	0.9253	32.57	0.9046	34.09	0.9449	40.07	0.9802
	SRFormer+ (ours)	DF2K	**38.58**	**0.9628**	**34.60**	**0.9262**	**32.61**	**0.9050**	**34.29**	**0.9457**	**40.19**	**0.9805**
×3 SR	**EDSR** [34]	DIV2K	34.65	0.9280	30.52	0.8462	29.25	0.8093	28.80	0.8653	34.17	0.9476
	RCAN [78]	DIV2K	34.74	0.9299	30.65	0.8482	29.32	0.8111	29.09	0.8702	34.44	0.9499
	SAN [7]	DIV2K	34.75	0.9300	30.59	0.8476	29.33	0.8112	28.93	0.8671	34.30	0.9494
	IGNN [81]	DIV2K	34.72	0.9298	30.66	0.8484	29.31	0.8105	29.03	0.8696	34.39	0.9496
	HAN [45]	DIV2K	34.75	0.9299	30.67	0.8483	29.32	0.8110	29.10	0.8705	34.48	0.9500
	NLSA [43]	DIV2K	34.85	0.9306	30.70	0.8485	29.34	0.8117	29.25	0.8726	34.57	0.9508
	SwinIR [33]	DIV2K	34.89	0.9312	30.77	0.8503	29.37	0.8124	29.29	0.8744	34.74	0.9518
	ELAN [77]	DIV2K	34.90	0.9313	30.80	0.8504	29.38	0.8124	29.32	0.8745	34.73	0.9517
	SRFormer (ours)	DIV2K	**34.94**	**0.9318**	**30.81**	**0.8518**	**29.41**	**0.8142**	**29.52**	**0.8786**	**34.78**	**0.9524**
	IPT [5]	ImageNet	34.81	-	30.85	-	29.38	-	29.49	-	-	-
	SwinIR [33]	DF2K	34.97	0.9318	30.93	0.8534	29.46	0.8145	29.75	0.8826	35.12	0.9537
	EDT [31]	DF2K	34.97	0.9316	30.89	0.8527	29.44	0.8142	29.72	0.8814	35.13	0.9534
	SRFormer (ours)	DF2K	35.02	0.9323	30.94	0.8540	29.48	0.8156	30.04	0.8865	35.26	0.9543
	SRFormer+ (ours)	DF2K	**35.08**	**0.9327**	**31.04**	**0.8551**	**29.53**	**0.8162**	**30.21**	**0.8884**	**35.45**	**0.9550**
×4 SR	**EDSR** [34]	DIV2K	32.46	0.8968	28.80	0.7876	27.71	0.7420	26.64	0.8033	31.02	0.9148
	RCAN [78]	DIV2K	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	31.22	0.9173
	SAN [7]	DIV2K	32.64	0.9003	28.92	0.7888	27.78	0.7436	26.79	0.8068	31.18	0.9169
	IGNN [81]	DIV2K	32.57	0.8998	28.85	0.7891	27.77	0.7434	26.84	0.8090	31.28	0.9182
	HAN [45]	DIV2K	32.64	0.9002	28.90	0.7890	27.80	0.7442	26.85	0.8094	31.42	0.9177
	NLSA [43]	DIV2K	32.59	0.9000	28.87	0.7891	27.78	0.7444	26.96	0.8109	31.27	0.9184
	SwinIR [33]	DIV2K	32.72	0.9021	28.94	0.7914	27.83	0.7459	27.07	0.8164	31.67	0.9226
	ELAN [77]	DIV2K	32.75	0.9022	28.96	0.7914	27.83	0.7459	27.13	0.8167	31.68	0.9226
	SRFormer (ours)	DIV2K	**32.81**	**0.9029**	**29.01**	**0.7919**	**27.85**	**0.7472**	**27.20**	**0.8189**	**31.75**	**0.9237**
	IPT [5]	ImageNet	32.64	-	29.01	-	27.82	-	27.26	-	-	-
	SwinIR [33]	DF2K	32.92	0.9044	29.09	0.7950	27.92	0.7489	27.45	0.8254	32.03	0.9260
	EDT [31]	DF2K	32.82	0.9031	29.09	0.7939	27.91	0.7483	27.46	0.8246	32.03	0.9254
	SRFormer (ours)	DF2K	32.93	0.9041	29.08	0.7953	27.94	0.7502	27.68	0.8311	32.21	0.9271
	SRFormer+ (ours)	DF2K	**33.09**	**0.9053**	**29.19**	**0.7965**	**28.00**	**0.7511**	**27.85**	**0.8338**	**32.44**	**0.9287**

表 4: 在五个基准数据集上对 SRFormer 和最近最先进的**经典图像 SR** 方法进行了定量比较。为了比较的公平性，SRFormer 的参数和 MACs **比 SwinIR 低** (详情见补充材料)。最优性能以高亮显示，次优以下划线显示。

更好的高分辨率图像。

4.4. 轻量级图像超分辨率

为了证明 SRFormer 的可扩展性、效率和有效性，我们训练了 SRFormer-light，并将其与一系列最先进的轻量级 SR 方法进行了比较：EDSR-baseline [34], CARN [1], IMDN [19], LAPAR-A [32], LatticeNet [40], ESRT [39], SwinIR-light [33], 以及 ELAN [77]。

定量比较。轻量级图像 SR 模型的定量比较如表 5所示。与先前的工作 [40, 1] 一致，我们在所有尺度上汇报 MACs 时均使用在将低分辨率图像放大到 1280 × 720 分辨率的设置。可以看到 SRFormer-light 在所有五个基准数据集上的所有比例因子都取得了最佳性能。相比于 SwinIR-light[33]，SRFormer-light 在 Urban100 和 Manga109 数据集上的 PSNR 分数分别高了将近 0.20dB 和 0.25dB，并且参数量和 MACs 更少。结果表明，置换自注意力是一种简单、但是更加有效的编码空间信息的方法。

定性比较。在图 6中，我们将 SRFormer 与最先进的轻量级图像 SR 模型进行了定性比较。值得注意的是，对于图 6中的所有示例，SRFormer-light 是唯一可以恢复

	方法	训练集	参数量	MACs	SET5 [3]		SET14 [70]		B100 [41]		Urban100 [18]		Manga109 [42]	
					PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
×2 SR	**EDSR-baseline** [34]	DIV2K	1370K	316G	37.99	0.9604	33.57	0.9175	32.16	0.8994	31.98	0.9272	38.54	0.9769
	CARN [1]	DIV2K	1592K	222.8G	37.76	0.9590	33.52	0.9166	32.09	0.8978	31.92	0.9256	38.36	0.9765
	IMDN [19]	DIV2K	694K	158.8G	38.00	0.9605	33.63	0.9177	32.19	0.8996	32.17	0.9283	38.88	0.9774
	LAPAR-A [32]	DF2K	548K	171G	38.01	0.9605	33.62	0.9183	32.19	0.8999	32.10	0.9283	38.67	0.9772
	LatticeNet [40]	DIV2K	756K	169.5G	38.15	0.9610	33.78	0.9193	32.25	0.9005	32.43	0.9302	-	-
	ESRT [39]	DIV2K	751K	-	38.03	0.9600	33.75	0.9184	32.25	0.9001	32.58	0.9318	39.12	0.9774
	SwinIR-light [33]	DIV2K	910K	244G	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
	ELAN [77]	DIV2K	621K	203G	38.17	0.9611	**33.94**	0.9207	32.30	0.9012	32.76	0.9340	39.11	0.9782
	SRFormer-light	DIV2K	853K	236G	**38.23**	**0.9613**	**33.94**	**0.9209**	**32.36**	**0.9019**	**32.91**	**0.9353**	**39.28**	**0.9785**
×3 SR	**EDSR-baseline** [34]	DIV2K	1555K	160G	34.37	0.9270	30.28	0.8417	29.09	0.8052	28.15	0.8527	33.45	0.9439
	CARN [1]	DIV2K	1592K	118.8G	34.29	0.9255	30.29	0.8407	29.06	0.8034	28.06	0.8493	33.50	0.9440
	IMDN [19]	DIV2K	703K	71.5G	34.36	0.9270	30.32	0.8417	29.09	0.8046	28.17	0.8519	33.61	0.9445
	LAPAR-A [32]	DF2K	594K	114G	34.36	0.9267	30.34	0.8421	29.11	0.8054	28.15	0.8523	33.51	0.9441
	LatticeNet [40]	DIV2K	765K	76.3G	34.53	0.9281	30.39	0.8424	29.15	0.8059	28.33	0.8538	-	-
	ESRT [39]	DIV2K	751K	-	34.42	0.9268	30.43	0.8433	29.15	0.8063	28.46	0.8574	33.95	0.9455
	SwinIR-light [33]	DIV2K	918K	111G	34.62	0.9289	30.54	0.8463	29.20	0.8082	28.66	0.8624	33.98	0.9478
	ELAN [77]	DIV2K	629K	90.1G	34.61	0.9288	30.55	0.8463	29.21	0.8081	28.69	0.8624	34.00	0.9478
	SRFormer-light	DIV2K	861K	105G	**34.67**	**0.9296**	**30.57**	**0.8469**	**29.26**	**0.8099**	**28.81**	**0.8655**	**34.19**	**0.9489**
×4 SR	**EDSR-baseline** [34]	DIV2K	1518K	114G	32.09	0.8938	28.58	0.7813	27.57	0.7357	26.04	0.7849	30.35	0.9067
	CARN [1]	DIV2K	1592K	90.9G	32.13	0.8937	28.60	0.7806	27.58	0.7349	26.07	0.7837	30.47	0.9084
	IMDN [19]	DIV2K	715K	40.9G	32.21	0.8948	28.58	0.7811	27.56	0.7353	26.04	0.7838	30.45	0.9075
	LAPAR-A [32]	DF2K	659K	94G	32.15	0.8944	28.61	0.7818	27.61	0.7366	26.14	0.7871	30.42	0.9074
	LatticeNet [40]	DIV2K	777K	43.6G	32.30	0.8962	28.68	0.7830	27.62	0.7367	26.25	0.7873	-	-
	ESRT [39]	DIV2K	751K	-	32.19	0.8947	28.69	0.7833	27.69	0.7379	26.39	0.7962	30.75	0.9100
	SwinIR-light [33]	DIV2K	930K	63.6G	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
	ELAN [77]	DIV2K	640K	54.1G	32.43	0.8975	28.78	0.7858	27.69	0.7406	26.54	0.7982	30.92	0.9150
	SRFormer-light	DIV2K	873K	62.8G	**32.51**	**0.8988**	**28.82**	**0.7872**	**27.73**	**0.7422**	**26.67**	**0.8032**	**31.17**	**0.9165**

表 5: 在五个基准数据集上对 SRFormer-light 和最近最先进的**轻量级图像 SR** 方法进行定量比较。重点介绍了所有模型中性能最好的模型。

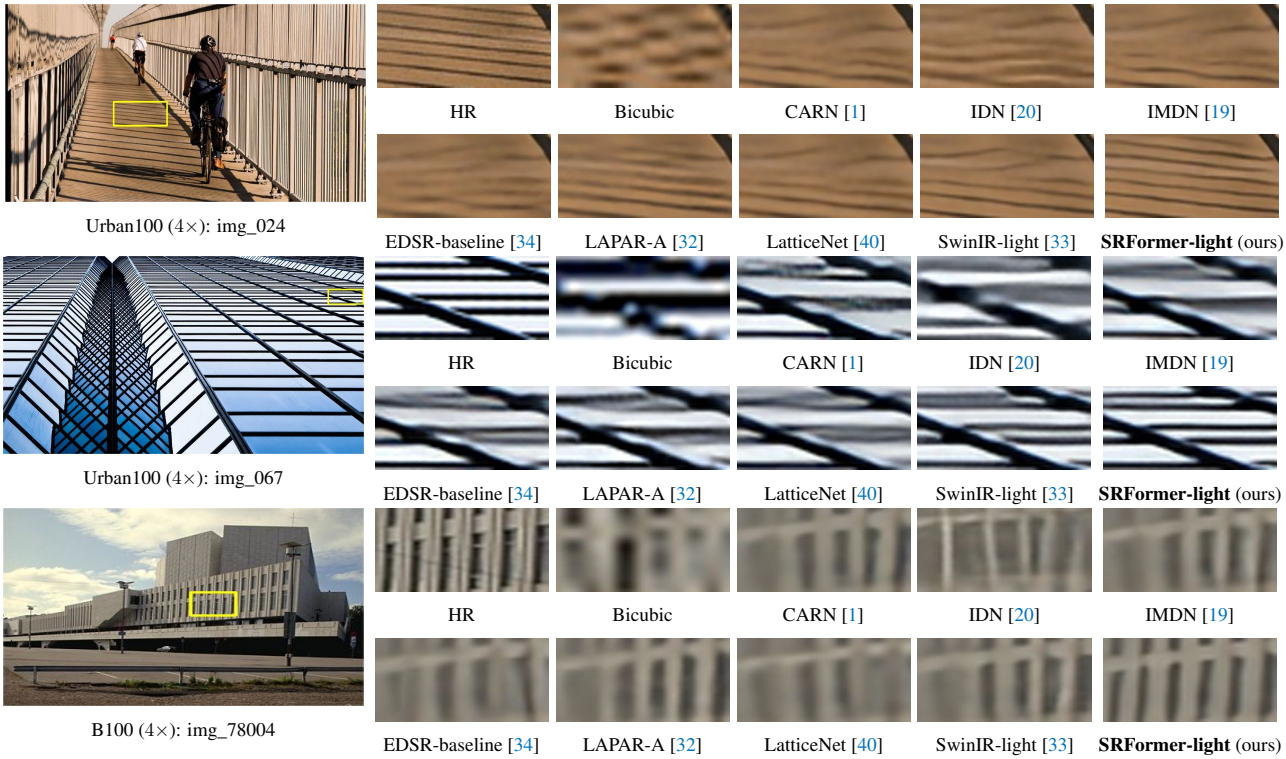

图 6: 在 ×4 SR 任务上，将 SRFormer-light 与最近最先进的**轻量级图像 SR** 方法进行定性比较。对于每个示例，我们的 SRFormer-light 可以比其他方法更好地恢复结构和细节。

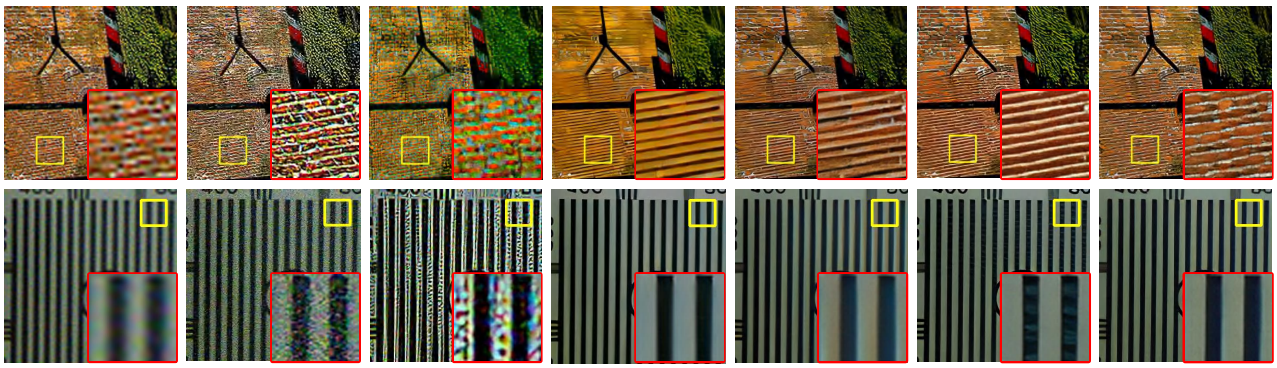

图 7: 在 ×4 **真实世界的图像 SR** 任务上与最近的最先进方法进行定性比较。

主要结构、并包含较少的模糊和伪影的模型。这有力地证明了 SRFormer 的轻量级版本在恢复边缘和纹理方面也比其他方法表现得更好。

4.5. 真实世界图像超分辨率

由于图像 SR 的最终目标是输入现实世界中的各种退化图像，并输出视觉上令人满意的的图像，我们与 SwinIR[33] 使用相同的退化模型 BSRGAN[74] 重新训练 SRFormer，并在图 7中展示结果。在处理真实世界的图像时，SRFormer 仍然可以产生更真实的结果、更舒适纹理、更少的伪影，这证明了我们方法的鲁棒性。

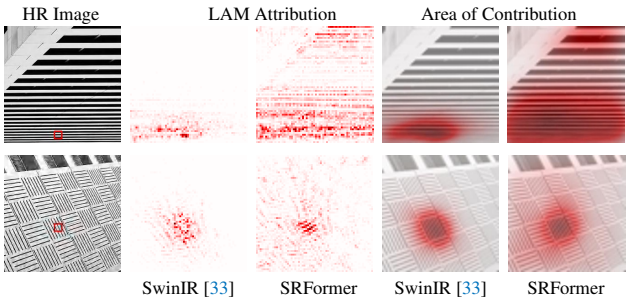

图 8: SwinIR [33] 和 SRFormer 在多个具有挑战性的例子上的 LAM 结果。我们可以看到 SRFormer 可以基于更加广泛的像素范围进行 SR 重建。

4.6. LAM 比较

为了观察用于 SR 重建的使用像素范围，我们使用 LAM [15] 将 SRFormer 与 SwinIR 进行比较，如图 8所示。基于极大的注意力窗口，SRFormer 使用比 SwinIR [33] 更大的像素范围来推断 SR 图像。实验结果与动机是强烈一致的，并从可解释性的角度证明了我们方法的优越性。

5. 结论

本文提出了一种高效的自注意力机制，PSA，可以有效地在大窗口中构建成对的相关性。基于 PSA，设计了一个简单有效的基于 Transformer 的单图像超分辨率模型 SRFormer。由于注意力窗口尺寸的增加以及高频信息增强，SRFormer 在经典的、轻量级的和真实的图像 SR 任务上都取得了最先进的性能。希望本文所提出的置换自注意力可以成为大窗口自注意力的一种范式，并成为一种可供今后超分辨率模型设计的研究提供参考的工具。

Acknowledgments. 本项目受到了国家自然科学基金 (No. 62276145, 62176130), 中央高校基本科研资金 (南开大学, 070-63223049, 070-63233089), 中国科协青年人才托举工程 (No. YESS20210377) 的支持。算力资源受到了南开大学超算中心的支持。

参考文献

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Eur. Conf. Comput. Vis.*, 2018.

[2] Saeed Anwar, Salman Khan, and Nick Barnes. A deep journey into super-resolution: A survey. *ACM Computing Surveys*, 2020.

[3] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *BMVC*, 2012.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-

end object detection with transformers. In *Eur. Conf. Comput. Vis.*, 2020.

[5] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv:2205.04437*, 2022.

[7] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[8] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Eur. Conf. Comput. Vis.*, 2014.

[10] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Eur. Conf. Comput. Vis.*, 2016.

[11] 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. In *Int. Conf. Learn. Represent.*, 2020.

[12] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. In *Adv. Neural Inform. Process. Syst.*, 2021.

[13] Zhicheng Geng, Luming Liang, Tianyu Ding, and Ilya Zharkov. Rstt: Real-time spatial temporal transformer for space-time video super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv:1406.2661*, 2014.

[15] Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[16] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[17] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[18] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015.

[19] Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *ACM Int. Conf. Multimedia*, 2019.

[20] Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[21] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2020.

[22] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. *arXiv:2102.07074*, 2021.

[23] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[24] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[25] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Deeply-recursive convolutional network for image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

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

[27] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[28] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative

adversarial network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[29] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *arXiv:2107.04589*, 2021.

[30] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Eur. Conf. Comput. Vis.*, 2018.

[31] Wenbo Li, Xin Lu, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *arXiv:2112.10175*, 2022.

[32] Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. Lapar: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. In *Adv. Neural Inform. Process. Syst.*, 2020.

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

[34] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2017.

[35] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[36] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *Int. Conf. Comput. Vis.*, 2021.

[37] 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 *Int. Conf. Comput. Vis.*, 2021.

[38] Liying Lu, Ruizheng Wu, Huaijia Lin, Jiangbo Lu, and Jiaya Jia. Video frame interpolation with transformer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[39] Zhisheng Lu, Hong Liu, Juncheng Li, and Linlin Zhang. Efficient transformer for single image super-resolution. *arXiv:2108.11084*, 2021.

[40] Xiaotong Luo, Yuan Xie, Yulun Zhang, Yanyun Qu, Cuihua Li, and Yun Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *Eur. Conf. Comput. Vis.*, 2020.

[41] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Int. Conf. Comput. Vis.*, 2001.

[42] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia Tools and Applications*, 2017.

[43] Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[44] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[45] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *Eur. Conf. Comput. Vis.*, 2020.

[46] Namuk Park and Songkuk Kim. How do vision transformers work? In *Int. Conf. Learn. Represent.*, 2022.

[47] Jingjing Ren, Qingqing Zheng, Yuanyuan Zhao, Xuemiao Xu, and Chen Li. Dlformer: Discrete latent transformer for video inpainting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[48] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[49] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Int. Conf. Comput. Vis.*, 2021.

[50] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *Int. Conf. Comput. Vis.*, 2021.

[51] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[52] Ying Tai, Jian Yang, Xiaoming Liu, and Chunyan Xu. Memnet: A persistent memory network for image restoration. In *Int. Conf. Comput. Vis.*, 2017.

[53] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2017.

[54] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *Int. Conf. Comput. Vis.*, 2017.

[55] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv:2012.12877*, 2020.

[56] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Int. Conf. Comput. Vis.*, 2021.

[57] Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv:2203.05962*, 2022.

[58] Wenhai 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 *Int. Conf. Comput. Vis.*, 2021.

[59] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019.

[60] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[61] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Int. Conf. Comput. Vis.*, 2021.

[62] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[63] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Eur. Conf. Comput. Vis. Worksh.*, 2018.

[64] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[65] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Adv. Neural Inform. Process. Syst.*, 2021.

[66] Yue Yang and Yong Qi. Image super-resolution via channel attention and spatial graph convolutional network. *Pattern Recognition*, 2021.

[67] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Int. Conf. Comput. Vis.*, 2021.

[68] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[69] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[70] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces*, 2010.

[71] Bowen Zhang, Shuyang Gu, Bo Zhang, Jianmin Bao, Dong Chen, Fang Wen, Yong Wang, and Baining Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022.

[72] Dafeng Zhang, Feiyu Huang, Shizhuo Liu, Xiaobing Wang, and Zhezhu Jin. Swinfir: Revisiting the swinir with fast fourier convolution and improved training for image super-resolution. *arXiv:2208.11247*, 2022.

[73] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[74] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Int. Conf. Comput. Vis.*, 2021.

[75] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Learning a single convolutional super-resolution network for multiple degradations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[76] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Int. Conf. Comput. Vis.*, 2019.

[77] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. *arXiv:2203.06697*, 2022.

[78] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Eur. Conf. Comput. Vis.*, 2018.

[79] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[80] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[81] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In *Adv. Neural Inform. Process. Syst.*, 2020.