

基于 FAKD 的 SwinIR 超分辨率模型知识蒸馏

摘要

针对 Transformer-based 图像超分辨率模型存在的计算复杂度高、推理速度慢等问题，本文使用迁移学习的方法，将应用在卷积神经网络的特征亲和性知识蒸馏（FAKD）方法迁移到 SwinIR 中，对其进行轻量化优化。该方法通过提取教师网络（原始 SwinIR）特征图的二阶统计信息，构建特征亲和矩阵损失函数，引导学生网络（轻量化 SwinIR）学习教师的结构知识。本文在图像超分辨率常用的数据集上进行测试，结果表明经过蒸馏后的轻量化模型在提升计算效率的同时，仍保持较高的超分精度。

关键词：图像超分辨率、SwinIR、FAKD、知识蒸馏

1 引言

单图像超分辨率（Single Image Super-Resolution, SISR）作为计算机视觉领域的重要任务，旨在从低分辨率（Low-Resolution, LR）图像中恢复高分辨率（High-Resolution, HR）图像，广泛应用于监控安防、医学影像、卫星遥感等领域。迄今为止，研究者已提出大量单图像超分辨率方法，主要包括基于插值的方法^[1]、基于稀疏表示的方法^[2]以及基于深度学习的方法^[3]。

随着深度学习技术的演进，基于 Swin Transformer 的 SwinIR^[4]模型融合了卷积神经网络（CNN）和 Transformer 的优点。它凭借窗口注意力机制对局部与全局特征的高效捕捉能力，以及层次化特征融合策略，在 Set5、Set14 等主流 benchmarks 上实现了当前最优的超分精度。然而，SwinIR 为追求高精度设计的复杂网络结构，使其在移动端、嵌入式设备等计算资源受限的边缘场景中面临推理延迟高、部署成本大的瓶颈，亟需高效的轻量化方案。

为获取轻量化模型，近年来研究者已提出多种网络模型压缩方案，主要包括模型剪枝、轻量化网络设计以及知识蒸馏（Knowledge Distillation, KD）方法。其中，模型剪枝与轻量化网络设计方法不仅需要精细的人工设计，还可能导致模型性能下降。相比之下，知识蒸馏方法在不改变网络结构的前提下，展现出优于其他模型压缩方法的优势。

传统知识蒸馏最初是为图像识别任务提出的，其遵循“教师-学生”范式：利用性能强大的教师网络输出的软标签（soft labels），对结构精简的学生网络（tiny student network）的训练过程进行监督。迄今为止，研究者已提出多种知识蒸馏方法。例如，Romero^[5]等人提出 FitNet，用于提取中间层特征图中隐含的知识并传递给学生网络；Sergey^[6]等人提出注意力迁移方法，通过从中间层特征计算注意力图，促使学生网络生成与教师网络相似的注意力图；考虑到层间相关性的重要性，Yim^[7]等人提出“Flow of Solution Procedure, FSP”，提取网络解决问题的相关信息，并将其作为监督信号指导学生网络训练。

为实现高效的图像超分辨率，本文使用一种基于特征亲和性的知识蒸馏（Feature Affinity-based Knowledge Distillation, FAKD）框架^[8]，该框架从特征的二阶统计信息入手，通过构建特征亲和矩阵量化不同通道间的相关性，进而将教师网络的结构知识高效迁移给学生网络。这种方法对 RCAN、SAN 等卷积超分辨率网络进行知识蒸馏取得了不错的效果，本文将其迁移到 Transformer-based 超分辨率网络。

本文聚焦于 SwinIR 模型的轻量化，研究 FAKD 方法在 Transformer 架构超分模型中的蒸馏机制，具体研究内容包括：

1. 针对 SwinIR 的网络特性设计轻量化学生网络，在压缩参数量与计算量的同时保留核心注意力机制。
2. 结合 SwinIR 的模型特点选择蒸馏损失的特征层，并使用 FAKD 方法进行知识蒸馏。
3. 在标准超分数据集上对模型进行测评分析，并与无蒸馏轻量化模型进行对比分析。

2 相关理论

2.1 SwinIR 网络结构

SwinIR 是一种基于 Transformer 架构的图像超分辨率重建算法，它包括浅层特征提取单元、深层特征提取单元、图像重建单元三部分。此方法不仅可以提升图像分辨率，还可以有效处理图像去噪、图像增强等多种图像恢复任务。本文只讨论其图像超分辨重建的功能。SwinIR 的结构如图 1 所示。

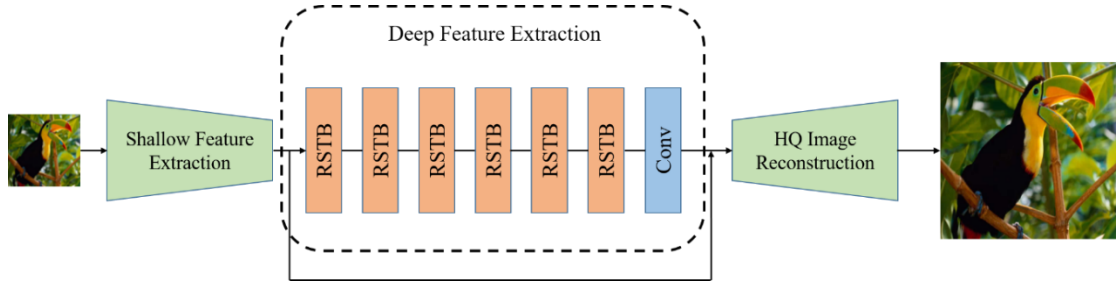


图 1 SwinIR 网络结构

浅层特征提取单元对图像进行卷积操作来提取低频信息，这些低频信息直接传输到图像重建单元。深度特征提取单元由残差 RSTB 模块和一层卷积组成，RSTB 模块专注于图像高频信息的提取，卷积层能够将图像低频信息与高频信息相融合。图像重建单元使用亚像素卷积操作，结合图像的低频信息与高频信息，得到重建后的高分辨率图像。

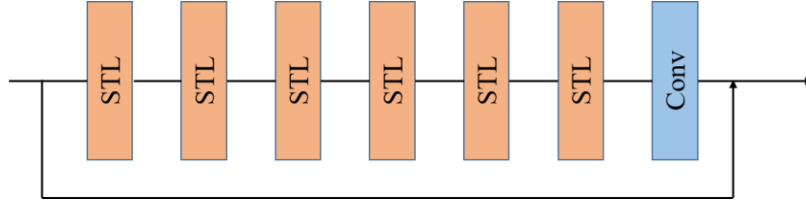


图 2 RSTB 结构

RSTB 模块的结构如图 2 所示，该模块由多个 STL 层和一个卷积层组成，STL 模块结构如图 3 所示。STL 架构的注意力机制在标准注意力机制的基础上，进行了两个创新，主要包括局部注意力机制和移动窗口机制。首先，将维度为 $H \times W \times C$ 的特征图作为输入，通过网格划分方式切分成多个互不重叠的 $M \times M$ 局部窗口，并将特征图的大小调整为 $\frac{HW}{M^2} \times M^2 \times C$ 。然后对每个窗口计算 query、key、value 矩阵，计算方式如公式(1)。

$$Q = XP_Q, K = XP_K, V = XP_V \quad (1)$$

其中 P_Q 、 P_K 、 P_V 是不同窗口共享的投影矩阵，注意力矩阵由公式(2)。

$$Attention(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d} + B}\right)V \quad (2)$$

其中 B 是可训练的相对位置偏置参数。然后对每一个特征图并行执行 h 次独立的注意力函数，并将每一次输出的特征进行拼接，从而构建出多头注意力机制 (MSA)。将 MSA、归一化层 (LayerNorm)、MLP 连接起来组成 STL 模块。

为了使窗口之间实现信息交互，引入了移动窗口机制。通过对特征图像进行

$\lfloor \frac{M}{2} \rfloor, \lfloor \frac{M}{2} \rfloor$ 的偏移，再对特征图像进行窗口划分，实现对特征图像的全覆盖，减少了窗口之间的割裂感和边界伪影的生成，提高了全局建模能力。

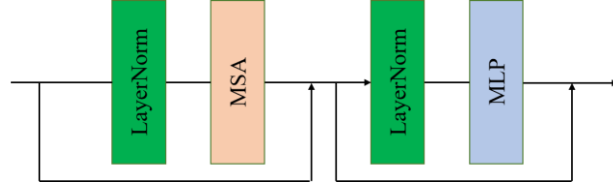


图 3 STL 结构

2.2 FAKD 知识蒸馏

基于特征亲和性的知识蒸馏框架（FAKD）的流程如图 4 所示。退化的低分辨率（LR）图像会同时输入教师网络（T）和学生网络（S）进行传播。其中，教师模型是性能强大但结构复杂冗余的网络，而学生模型则是轻量级网络。在该框架中，两者采用相同的网络架构，仅超参数（如网络深度）存在差异。如图 4 所示，教师网络和学生网络分别由 m 个和 n 个残差块组成（且 $m > n$ ）。为了将教师模型的知识有效迁移到学生模型中，模型强制学生网络的中间特征图模仿教师网络的特征图。此外，教师网络的输出图像和真实标签图像（ground-true images）还会分别通过教师监督（TS）和数据监督（DS）两种方式，对学生网络的训练过程进行监督。

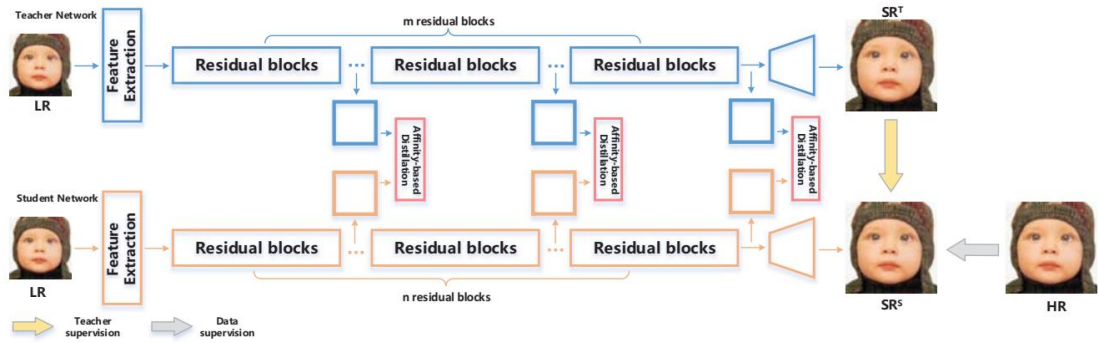


图 4 FAKD 算法流程图^[8]

为了使学生网络的中间特征图模仿教师网络的特征图，模型构建了特征图的空间亲和矩阵 A （affinity matrix）。构建流程如图 5 所示，给定一批特征图 $F \in \mathbb{R}^{b \times C \times W \times H}$ ，首先将其重塑为三维张量 $F \in \mathbb{R}^{b \times C \times WH}$ ，这三个维度分别对应实例维度（instance）、通道维度（channel）和空间维度（spatial dimension）。在特征图中每个像素均被视为一个 C 维向量（即蓝色列向量），且需对每一列进行归一化处理。

归一化后，每一列向量的长度均为 1，使用内积计算归一化后的特征图的余弦相似度，即得到空间亲和矩阵 A 。亲和矩阵的计算公式如公式(3)和公式(4)所示。

$$\tilde{F}_{[i,:]} = \frac{F_{[i,:]}}{\|F_{[i,:]} \|_2} \quad (3)$$

$$A = \tilde{F}^T \cdot \tilde{F} \quad (4)$$

其中， $\tilde{F}$ 表示归一化后的特征图。生成的空间亲和矩阵维度为 $b \times HW \times HW$ ，矩阵中的每个元素均表示两个像素之间的空间相关性。FAKD 使用特征亲和性的蒸馏损失促使学生网络生成与教师网络相似的亲和矩阵，如公式(5)所示。

$$L_{AD} = \frac{1}{|A|} \sum_{l=1}^{l'} \|A_l^S - A_l^T\|_1 \quad (5)$$

其中， A_l^T 和 A_l^S 分别表示从教师网络和学生网络第 l 层特征图中提取的亲和矩阵； l' 表示所选的特征提取层数； $|A|$ 表示亲和矩阵中元素的数量。

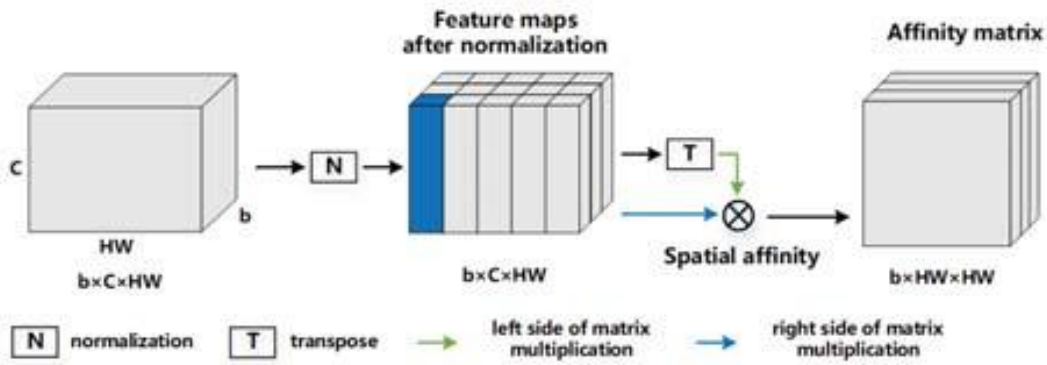


图 5 亲和矩阵构建流程^[8]

除了使学生网络的中间特征图模仿教师网络的特征图，模型还使用教师监督（TS）与数据监督（DS）提升蒸馏性能。其中，教师监督（TS）用于将学生网络的输出与教师网络的输出进行对比，数据监督（DS）用于将学生网络的输出与真实标签图像（ground-true images）进行对比，具体如公式 6 和公式 7 所示，整体的损失函数的计算公式如公式 8 所示。

$$L_{TS} = \sum_{i=1}^N \|I_{SR}^{S(i)} - I_{SR}^{T(i)}\|_1 \quad (6)$$

$$L_{DS} = \sum_{i=1}^N \|I_{SR}^{S(i)} - I_{HR}^{(i)}\|_1 \quad (7)$$

$$L(\theta) = \alpha L_{DS} + \beta L_{TS} + \gamma L_{AD} \quad (8)$$

其中, I_{SR}^S 、 I_{SR}^T 、 I_{HR} 分别表示学生网络输出的图像、教师网络输出的图像以及真实标签图像集中的图像; α 、 β 和 γ 是惩罚系数, 用于平衡损失函数中不同部分的权重。

3 系统实现

3.1 模型设计

本文采用“教师-学生”双网络架构, 教师网络为原始的 SwinIR 模型, 其包括浅层特征提取单元、深层特征提取单元、图像重建单元, 深层特征提取单元包括 6 个 RSTB 模块。学生网络基于教师网络进行轻量化设计, 根据 FAKD 的学生网络设计原则, SwinIR 模型的学生网络将深层特征提取单元设置为 4 个 RSTB 模块。同时, 本文将教师网络的第 2 个 RSTB 模块、第 4 个 RSTB 模块和深层特征提取单元输出的特征图分别对应学生网络的第 1 个 RSTB 模块、第 3 个 RSTB 模块和深层特征提取单元输出的特征图, 以此来计算特征亲和损失。具体结构如图 6 所示。

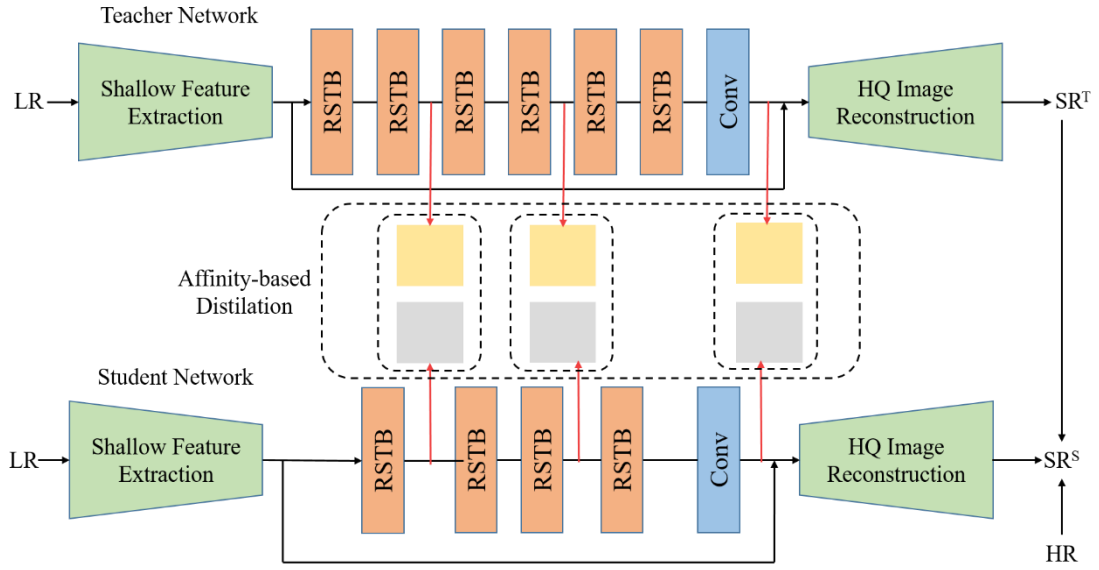


图 6 蒸馏结构示意图

3.2 训练设置

训练环境如表 1 所示, 本文使用 DIV2K 数据集作为训练集, Set5、Set14、BSD100、Urban100、Mango109 数据集作为测试集, 低分辨率图像通过双三次下采样生成。学生网络使用 Adam 优化器进行训练, 其中优化器参数设置为 $\beta_1 =$

0.9、 $\beta_2 = 0.99$ 、 $\epsilon = 10^{-8}$ 。训练轮数为 200epoch，初始学习率为 10^{-8} ，训练达到 150epoch 时，学习率衰减至当前值的一半。

表 1 训练环境

CPU	AMD EPYC 7542
GPU	RTX-4090-24G
Pytorch	2.3.0
Python	3.10
CUDA	11.8.0

4 测试评估

测试评估分为定量评估和定性评估。其中定量评估图像重建质量使用峰值信噪比（PSNR）、结构相似性（SSIM）进行评估，分别比较 2 倍、3 倍、4 倍超分尺度的重建效果；轻量化指标使用乘加操作数（MACs）和参数量（Params）。定性评估展示超分辨率重建的结果图像，进行定性评估。

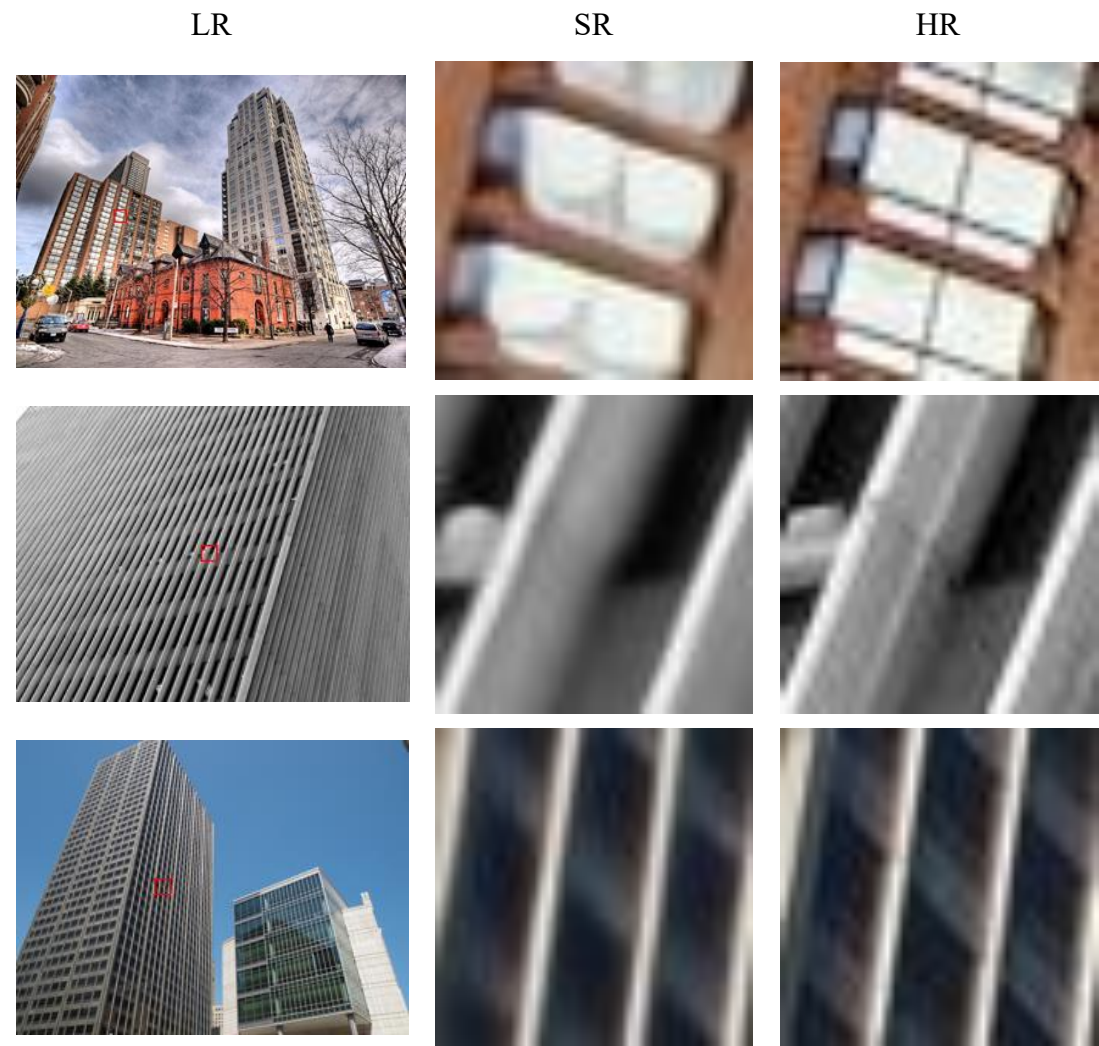


图 7 可视化结果展示

4.1 定量评估

2 倍、3 倍、4 倍尺度的超分辨率重建指标如表 2、表 3、表 4 所示。轻量化模型与原始模型的运算量和参数量对比如表 5 所示。可以看出，与原模型和训练的轻量化模型对比峰值信噪比和结构相似性仅有微小的损失，但是其在 manga109 数据集上的效果较差。运算量与参数量与原始模型对比有着明显的下降。

表 2 2 倍超分辨率重建结果

算法	Set5	Set14	Urban100	manga109	B100
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
原始 SwinIR	38.42/0.9623	34.46/0.9250	33.81/0.9427	39.92/0.9797	32.53/0.9041
训练的轻量化 SwinIR	38.14/0.9611	33.86/0.9206	32.76/0.9340	39.12/0.9783	32.31/0.9012
蒸馏的轻量化 SwinIR	36.74/0.9481	31.45/0.9073	30.71/0.8973	22.12/0.7065	30.43/0.8793

表 3 3 倍超分辨率重建结果

算法	Set5	Set14	Urban100	manga109	B100
	PSNR/SSIM	PSNR/SSIM	PSNRSSIM	PSNR/SSIM	PSNR/SSIM
原始 SwinIR	34.97/0.9318	30.93/0.8534	29.75/0.8826	35.12/0.9537	29.46/0.8145
训练的轻量化 SwinIR	34.62/0.9289	30.54/0.8436	28.66/0.8624	33.89/0.9478	29.20/0.8082
蒸馏的轻量化 SwinIR	32.61/0.9049	28.65/0.8145	27.09/0.8463	18.47/0.6176	27.34/0.7856

表 4 4 倍超分辨率重建结果

算法	Set5	Set14	Urban100	manga109	B100
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
原始 SwinIR	32.72/0.9021	28.94/0.7914	27.07/0.8146	31.67/0.9226	27.83/0.7459
训练的轻量化 SwinIR	32.44/0.8976	28.77/0.7858	26.47/0.7980	30.92/0.9151	27.69/0.7406
蒸馏的轻量化 SwinIR	30.52/0.8679	26.96/0.7490	24.96/0.7773	16.54/0.5913	26.34/0.7136

表 5 模型运算量与参数量对比

算法	MACs (G)	Params (M)
原始 SwimIR 模型	599.45	56.8
轻量化 SwinIR 模型	44.70	16.3

4.2 定性评估

图 7 展示了部分的重建结果，重建结果为 4 倍上采用倍率的图像超分辨率。可以看出使用 FAKD 蒸馏后的 SwinIR 模型能够很好地恢复图像的纹理信息。同时，其具备更低的运算量与参数量。

5 结论

本文将 FAKD 方法应用于 SwinIR 模型的知识蒸馏，通过蒸馏特征亲和矩阵的二阶统计信息，实现了模型的有效轻量化。实验结果表明，轻量化后的模型在保证计算效率提升的同时，仍保持了较高的超分精度，验证了 FAKD 方法在 Transformer-based 超分辨率模型轻量化中的可行性和优越性。

参考文献

- [1] Xiangjun Zhang and Xiaolin Wu, “Image interpolation by adaptive 2-d autoregressive modeling and soft-decision estimation,” TIP, 2008.
- [2] Weisheng Dong, Lei Zhang, Rastislav Lukac, and Guangming Shi, “Sparse representation based image interpolation with nonlocal autoregressive modeling,” TIP, 2013.
- [3] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang, “Second-order attention network for single image super-resolution,” in CVPR, 2019.
- [4] Liang J, Cao J, Sun G, et al. Swinir: Image restoration using swin transformer[C]Proceedings of the IEEE/CVF international conference on computer vision. 2021: 1833-1844.
- [5] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio, “Fitnets: Hints for thin deep nets,” arXiv preprint arXiv:1412.6550, 2014.
- [6] Sergey Zagoruyko and Nikos Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” arXiv preprint arXiv:1612.03928, 2016.
- [7] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim, “A gift from knowledge distillation: Fast optimization, network minimization and transfer learning,” in CVPR, 2017.
- [8] He, Zibin, et al. "Fakd: Feature-affinity based knowledge distillation for efficient image super-resolution." 2020 IEEE international conference on image processing (ICIP). IEEE, 2020.