arXiv:2304.01994v2 [cs.CV] 5 Apr 2023

挥别低分辨率：用于图像超分辨率的扩散-小波方法

Brian B. Moser[1,2], Stanislav Frolov[1,2],Federico Raue[1],Sebastian Palacio[1],Andreas Dengel[1,2] 德国人工智能研究中心（DFKI），德国[2] 凯撒斯劳滕-兰道莱茵兰-普法尔茨工业大学，德国

first.second@dfki.de

摘要

本文提出了一种用于单图像超分辨率（*SISR*）的新型扩散-小波（*DiWa*）方法。该方法结合了去噪扩散概率模型（*DDPMs*）和离散小波变换（*DWT*）的优势。通过使 *DDPMs* 在 *DWT*域中运行，我们的 *DDPM* 模型能够在小波频谱上为超分辨率图像有效生成高频信息，从而在图像空间中实现高质量、细节丰富的重建。在定量评估方面，无论是在面部超分辨率（*8*倍缩放）还是通用超分辨率（*4*倍缩放）基准测试中，我们的方法在峰值信噪比、结构相似性指数和学习感知图像块相似度方面均优于当前最先进的基于扩散的单图像超分辨率方法 *SR3* 和 *SRDiff*。同时，使用 *DWT* 使我们的模型能够采用比对比模型更少的参数：与 *SR3* 相比为 *92M*，而非 *550M*；与 *SRDiff* 相比为 *9.3M*，而非 *12M*。此外，在经典通用超分辨率数据集上，我们的方法优于其他当前最先进的生成方法，同时节省了推理时间。最后，我们的工作展现了该方法在多种应用中的潜力。

1 引言

单图像超分辨率（SISR）是计算机视觉领域长期存在且适定性不佳的挑战，因为对于任意给定的低分辨率（LR）图像，都可能存在许多有效的高分辨率（HR）图像。基于回归的方法（如卷积神经网络（CNN））在放大倍率较低时或许有效，但往往无法重现高倍率所需的高频细节。生成模型，以及近年来的去噪扩散概率模型（DDPM），已被证明是应对此类情况的有效工具 [49, 10, 7]。此外，与基于回归的方法相比，DDPM生成的重建结果在主观感知上质量更佳 [41]。要进一步推进DDPM的发展，需要对高频细节进行更精细的预测 [33]。另一个迫切需求是提升可及性，因为DDPM的计算要求很高 [12]。

本文提出一种新颖的扩散-小波方法（DiWa），通过引入离散小波变换（DWT）来解决这两个问题，并发挥DDPM的能力。我们修改了DDPM流程，使其不再在图像空间中运作，而是在DWT域中运作，从而更好地学习高频细节。采用DWT的动机有两方面：首先，结合DWT与DDPM能够提升图像质量，使模型得以捕捉并保留那些直接处理时可能丢失或失真的关键特征。DWT提供了一种替代表示方式，将高频细节明确分离到不同子带中。因此，这些表示更为稀疏，也更易于网络学习[29, 15]。这一特性也已应用于基于扩散的音频合成，并取得了令人瞩目的成果 [20]。

其次，根据奈奎斯特定理，DWT 会将图像的空间尺寸减半 [13]，，从而加快去噪函数（卷积神经网络）的推理速度；当模型在去噪扩散概率模型推理过程中被多次应用时，这一优势尤为明显。Phung 等人在近期的一项研究中 [35]，采用了类似方法，通过 DiffusionGAN [50], 进行图像生成，展现了该方法的加速潜力。然而，DiffusionGAN 与传统的去噪扩散概率模型不同，它使用生成对抗网络近似中间步骤，以减少图像生成所需的时间步数。

总而言之，DWT 为当前超分辨率领域的方法带来了切实的优势。我们的工作主要贡献如下：

- 首次将去噪扩散概率模型与 DWT 融合应用于超分辨率，通过迭代式高频细化利用降维后的频率表示，在性能和参数方面获益。
- 在通用超分辨率任务中，我们的方法优于当前最先进的扩散模型SR3 [41]（用于人脸超分辨率）和SRDiff [23]，也优于其他生成式方法，包括SFTGAN[47], 、SRGAN [22], 、ESRGAN [48], 、NatSR [42], 和SPSR [28], 。性能的提升使我们的方法适合作为图像分类 [53]等其他应用的预处理步骤。

- 与图像空间相比，基于频率的表示的空间面积缩小了4倍，因此所需的感受野更小，去噪函数的参数数量也得以减少。这使得无法使用大规模计算资源的研究人员也能更轻松地采用该方法：与SR3相比，我们的方法仅需9,200万参数，而非5.5亿；与SRDiff相比，仅需930万参数，而非1,200万。

2 背景

将离散小波变换（DWT）与去噪扩散概率模型（DDPMs）融合，既利用了 DDPMs 的生成能力，也发挥了 DWT 在表示方面的优势。这使推理速度更快、学习目标更稀疏，因此结果也更清晰。本节介绍二维离散小波变换和 DDPMs 的定义。

2.1 离散小波变换

离散小波变换（DWT）是一种用于分析和表示信号的技术，能够揭示原始数据中不明显的重要细节。DWT 用于多种图像处理应用，包括图像去噪、压缩和超分辨率 [45, 2, 25]。

给定一个信号 $\mathbf{x}\,[n] \in \mathbb{R}^N$，一维离散小波变换（1D-DWT）首先应用半带高通滤波器 $h\,[n]$，然后应用低通滤波器 $l\,[n]$。滤波器的构造取决于所选的小波。一个广为人知的选择是 Haar（“db1”）小波 [14]。

每幅图像都可以表示为一个索引为 $[n, m]$ 的二维信号，其中 n 和 m 分别表示列和行。因此，设$\mathbf{x} \in \mathbb{R}^{w\times h\times c}$为一幅彩色图像。二维离散小波变换将图像细节分解到四个子带中：平均（A）、垂直（V）、水平（H）和对角（D）信息：

$$\check{\mathbf{x}} = \text{2D-DWT}\,(\mathbf{x}) = \mathbf{x}_A \odot \mathbf{x}_V \odot \mathbf{x}_H \odot \mathbf{x}_D, \quad (1)$$

其中 $\odot$ 是按通道拼接的算子，$\mathbf{x}_A$ 、$\mathbf{x}_V$ 、$\mathbf{x}_H$ 和 $\mathbf{x}_D$ 分别表示$\mathbf{x}$的平均、垂直、水平和对角细节，其中 $\mathbf{x}_A$，$\mathbf{x}_V$，$\mathbf{x}_H$，$\mathbf{x}_D \in \mathbb{R}^{\frac{w}{2}\times\frac{h}{2}\times c}$。

二维逆离散小波变换是二维离散小波变换的逆运算：

$$\text{2D-IDWT}\,(\check{\mathbf{x}}) = \mathbf{x}. \quad (2)$$

2.2 去噪扩散概率模型

去噪扩散概率模型（DDPM）的核心思想是根据像素值周围的邻居逐步更新像素值 [16]。这一过程可以理解为随着时间推移，对图像中的信息进行“扩散”，以平滑噪声并降低数据的整体变化幅度。去噪扩散概率模型的应用通常分为两个阶段：前向扩散和反向过程。

扩散过程逐步向输入添加噪声，而反向过程旨在通过迭代去除所添加的噪声。通过迭代细化，卷积神经网络更容易跟踪微小扰动并加以纠正，而不是预测一次较大的变换 [16, 34]。

设 $\mathcal{D} = \{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^N$ 为一个条件分布 $p(\mathbf{y}\,|\mathbf{x})$ 未知的数据集。对于超分辨率任务，$\mathbf{x}_i$ 是低分辨率图像，而$\mathbf{y}_i$ 是对应的高分辨率图像；$\mathbf{x}_i$ 与$\mathbf{y}_i$ 之间存在未知的退化关系。

2.2.1 高斯扩散过程

为实现迭代细化，马尔可夫扩散过程 q 会向输入图像$\mathbf{z}_0 \leftarrow$ x</style}中逐步添加高斯噪声，共进行 T 个时间步 [34, 43]:

$$q(\mathbf{z}_{1:T} \mid \mathbf{z}_0) = \prod_{t=1}^{T} q(\mathbf{z}_t \mid \mathbf{z}_{t-1}), \quad (3)$$

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t \mid \sqrt{\alpha_t}\,\mathbf{z}_{t-1}, (1-\alpha_t)\mathbf{I}) \quad (4)$$

其中 $0 < \alpha_{1:T} < 1$ 是每个时间步添加噪声方差的超参数。该公式的一个主要优点是，可以将其简化为单步计算：

$$q(\mathbf{z}_t \mid \mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t \mid \sqrt{\gamma_t}\,\mathbf{z}_0, (1-\gamma_t)\mathbf{I}), \quad (5)$$

其中 $\gamma_t = \prod_{i=1}^{t} \alpha_i$。这种边缘化使得可以针对任意 $t \in \{1, ..., T\}$进行单时间步训练。

2.2.2 条件反向过程

反向马尔可夫过程 p 从高斯噪声$\mathbf{z}_T$开始，并以$\mathbf{x}$ 为条件进行推理，方法是

$$p_\theta(\mathbf{z}_{0:T}|\mathbf{x}) = p(\mathbf{z}_T) \prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) \quad (6)$$

$$p(\mathbf{z}_T) = \mathcal{N}(\mathbf{z}_T \mid \mathbf{0}, \mathbf{I}) \quad (7)$$

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{x}) = \mathcal{N}(\mathbf{z}_{t-1} \mid \mu_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t), \sigma_t^2\mathbf{I}). \quad (8)$$

均值 μ_θ 取决于参数化的去噪函数 f_θ，该函数可以预测添加的噪声 ε，也可以预测底层图像$\mathbf{z}_0$。遵循 Ho 等人 [16], 的标准方法，本文重点预测噪声。因此，均值为

$$\mu_\theta(\mathbf{x}, \mathbf{z}_t, \gamma_t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta\left(\mathbf{x}, \mathbf{z}_t, \gamma_t\right) \right). \quad (9)$$

按照 Saharia 等人[41]，将$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t,\mathbf{x})$ 的方差设为 $(1-\alpha_t)$，可得到以下细化步骤：

$$\mathbf{z}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\gamma_t}} f_\theta\left(\mathbf{x}, \mathbf{z}_t, \gamma_t\right) \right) + \sqrt{1-\alpha_t}\varepsilon_t, \quad (10)$$

where $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

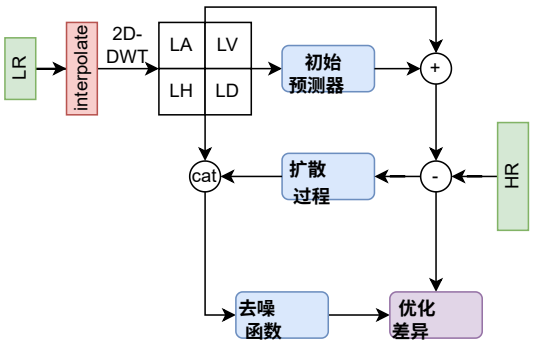


图 1：训练概览。扩散过程将初始预测器与对应高分辨率图像之间的差值作为输入。训练后的反向过程学习对含噪残差图像进行去噪，并以初始预测器重建结果与对应高分辨率图像之间的差值作为优化目标。

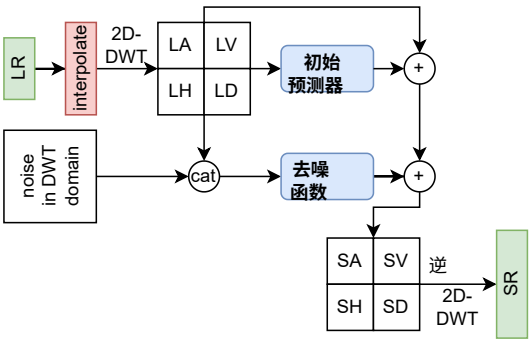


图 2：推理概览。首先使用二维离散小波变换将图像分解为子带，并由初始预测器处理。随后，去噪函数在以子带为条件并结合噪声的情况下，添加并计算剩余细节。最后，通过逆二维离散小波变换函数将结果返回像素域。

算法1 DiWa - 训练

需要： f_θ：去噪网络，g_θ：初始预测器，
$\mathcal{D} = \{(\mathbf{x}, \mathbf{y})\}$：低分辨率图像及对应的高分辨率图像对，
$\alpha_{1:T}$：噪声调度。

重复
  $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$
  $\check{\mathbf{x}} \leftarrow$ 2D-DWT (x)
  $\check{\mathbf{y}} \leftarrow$ 2D-DWT (y)
  $t \sim$ 均匀分布$(\{1, \cdots, T\})$ 和 $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\gamma_t \leftarrow \prod_{i=1}^{t} \alpha_i$
  $\mathbf{z}_t \leftarrow \sqrt{\gamma_t}(\check{\mathbf{y}} - g_\theta(\check{\mathbf{x}})) + \sqrt{1-\gamma_t}\varepsilon_t$
  执行梯度步：
    $\nabla_\theta \| \varepsilon_t - f_\theta(\check{\mathbf{x}}, \mathbf{z}_t, \gamma_t) \|$
直到收敛

算法2 DiWa - T 个细化步骤中的推理

需要： f_θ：去噪网络，g_θ：初始预测器，
$\mathbf{x}$：模糊输入图像，$\alpha_{1:T}$：噪声调度。
$\check{\mathbf{x}} \leftarrow$ 2D-离散小波变换(x)
$\mathbf{x}_{\text{init}} \leftarrow g_\theta(\check{\mathbf{x}})$
$\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
对于 $t = T, \ldots, 1$ **执行**
  $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{z}_{t-1} \leftarrow \mu_\theta(\check{\mathbf{x}}, \mathbf{z}_t, \gamma_t) + \sqrt{1-\alpha_t}\varepsilon_t$
循环结束
返回 二维逆离散小波变换 $(\mathbf{x}_{\text{init}} + \mathbf{z}_0)$

3 方法

我们提出的扩散-小波方法（DiWa）将条件扩散模型与小波分解相结合，从而实现更稀疏、更易学习的学习目标，进而加快推理并获得更精细的重建。因此，我们的扩散模型在小波频谱上运行，而非在原始图像空间中运行。它利用小波域的独特属性，挖掘高频信息并提升最终重建的质量。

本节首先介绍我们如何使用二维离散小波变换域。随后，我们描述如何集成并优化初始预测器，该预测器用于生成最终重建的初步估计。我们在算法 1 （训练）和算法 2 （推理）中给出了我们方法的完整算法。该方法由三个主要部分组成：二维离散小波变换表示、初始预测器和扩散。

3.1 DWT域

设 $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}$ 为一对低分辨率-高分辨率图像。在应用扩散过程之前，我们使用 $\check{\mathbf{x}}_i = \text{2D-DWT}(\mathbf{x}_i)$ 和 Haar（"db1"） [14] 小波将 $\mathbf{x}_i$ 变换到小波域。该分解将输入拆分为四个子带，分别表示平均图像（LA）以及水平方向、垂直方向和对角方向上的高频细节（LH、LV、LD），如公式 1 所述。同时，子带的空间区域比原始图像小四倍。接下来，将高斯噪声 $\varepsilon \sim \mathcal{N}(0, \text{I})$添加到这些子带中，去噪函数 f_θ 学习在该域中运行。我们还针对目标 $\check{\mathbf{y}}_i$，在小波域中优化去噪函数的参数 θ。在最终推理采样时，我们的方法返回逆变换二维逆离散小波变换。

通过在小波域中进行处理，我们使去噪函数能够专注于孤立的高频细节，这些细节在直接处理数据时往往会丢失。 f_θ 此外，利用离散小波变换可以缩短推理时间，因为根据奈奎斯特定理，空间尺寸减半 [13]。

3.2 初始预测器

由于 DWT 子带的方差各不相同，去噪函数必须适应不同的分布。尤其是，平均子带类似于原始图像的下采样版本，因此所含信息比其余高频子带的稀疏表示更丰富。我们通过仅学习残差来解决适应问题，从而使目标分布更具可比性。因此，我们的方法会对确定性初始预测器的输出进行细化 g_θ 该预测器能够针对给定的低分辨率输入提供多样且合理的超分辨率估计。这进一步推动了高频预测范式的发展，通过学习传统超分辨率模型未能捕捉到的稀疏且缺失的细节。该初始预测器的设计受到 Whang 等人工作的启发 [49] ，但在小波域中运行。

由于我们的方法在二维离散小波变换域中运行，为简洁起见，我们采用 Guo 等人提出的 DWSR 网络 [13]作为初始预测器 g_θ。该网络是一种用于超分辨率的简单十层小波卷积神经网络，可预测四个高分辨率小波系数。

3.3 优化

对去噪函数 f_θ 进行优化，以去除添加的噪声。因此，我们最小化目标函数

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})} \mathop{\mathbb{E}}_{t} \left\| \varepsilon_t - f_\theta\left(\check{\mathbf{x}}, \mathbf{z}_t, \gamma_t\right) \right\|_1 \tag{11}$$

$$\mathbf{z}_t = \sqrt{\gamma_t}\left(\check{\mathbf{y}} - g_\theta\left(\check{\mathbf{x}}\right)\right) + \sqrt{1-\gamma_t}\varepsilon_t \tag{12}$$

与 Whang 等人 [49]的方法类似，初始预测器是目标函数的一部分。因此，无需额外的辅助损失函数来训练初始预测器，因为梯度 $\nabla_\theta f_\theta$ 会经由 f_θ 流入 $\nabla_\theta g_\theta$。

参数化去噪函数 f_θ 只需对残差子带进行建模，这比重建整幅图像更容易学习。因此，初始预测器提高了优化过程的整体效率，并减少了收敛到满意解所需的迭代次数。推理和训练过程如图 1和图 2所示。

4 实验

我们评估了所提出的方法在通用超分辨率任务以及更具挑战性的人脸超分辨率任务中的表现。我们的实验旨在展示该方法的有效性，并将其性能与3 [41] 用于人脸超分辨率的方法和 SRDiff [23] 以及其他当前最先进的生成式方法 [27, 28, 24, 42, 22, 54, 47, 48] （用于通用超分辨率）进行比较。我们展示了视觉示例，并给出了这两项任务的定量和定性结果。总体而言，我们在人脸超分辨率和通用超分辨率任务中均取得了高质量的结果，并优于所比较的方法在峰值信噪比、结构相似性指数和学习感知图像块相似度等标准指标上的表现。

4.1 数据集

在人脸超分辨率任务中，我们将 SRDiff 与我们的方法进行比较。在通用超分辨率任务中，我们将我们的方法与 SR3 进行基准比较。与文献中的常见做法一致，我们采用双三次插值和抗锯齿处理生成低分辨率—高分辨率图像对，这会丢弃高频信息 [31]。

人脸超分辨率： 我们使用来自 Flickr 的 Flickr 高质量人脸数据集（FFHQ） [18], 50K 张高质量人脸图像进行训练。评估时，我们使用了 CelebA-HQ [17], , 该数据集包含 30K 张人脸图像。我们遵循了 Saharia 等人提出的两种 8 倍缩放设置 [41]。我们调整了所有图像的大小，以匹配这些情况$16 \times 16 \rightarrow 128 \times 128$ 和 $64 \times 64 \rightarrow 512 \times 512$。

通用超分辨率： 我们使用来自 DIV2K [1] 的 800 张 2K 分辨率高质量图像进行训练，并使用数据集 Set5[4], Set 14 [52], BSDS100 [30], 和 General100 [9] 进行评估。此外，我们使用 DIV2K验证集将我们的方法与 SRDiff 进行比较。与文献中的常见做法一致，我们遵循标准流程，从 DIV2K 中提取子图像 $48 \times 48 \rightarrow 192 \times 192$，用于 4 倍缩放训练 [8, 33]。测试时，我们保留图像的原始尺寸，这是单图像超分辨率中的标准做法。

4.2 训练细节

我们实验所用的代码和训练好的模型可在 GitHub 上找到[1], , 该项目补充了非官方的超分辨率实现3[2]。

扩散专用： 为在训练期间实现细粒度的扩散过程，我们将时间步设为 2,000。为加快推理速度，我们像 Saharia 等人那样，在评估时将时间步减少至 500。 [41]。我们未评估其他时间步，以免影响可比性。我们为人脸超分辨率训练了 100 万次迭代，为通用超分辨率训练了 10 万次迭代。线性噪声调度的端点分别为 $1-\alpha_0 = 10^{-6}$ 和 $1-\alpha_T = 10^{-2}$。

正则化： 我们的所有实验均以 50% 的概率进行水平翻转。此外，除$64 \times 64 \rightarrow 512 \times 512$ 人脸超分辨率外，我们在所有实验中都使用 Dropout [44] （比例为 10%）。

架构： 与 SR3 和 SRDiff 类似，我们采用 U-Net 架构 [39] 作为去噪函数。与 SR3 采用的方法不同，我们使用 Ho 等人 [16] 提出的残差块[15]，而非 Big- GAN [5] 中使用的残差块。与 SRDiff 不同，我们不预先使用预训练的低分辨率编码器将低分辨率图像转换为特征表示，而是使用初始预测器。

[1] https://github.com/Brian-Moser/diwa [2] https://github.com/Janspiry/图像-超分辨率-via-迭代-细化

	Pulse [32]	FSRGAN [6]	SR3 [41]	DiWa （我们的方法）
峰值信噪比 ↑	16.88	23.01	23.04	**23.34**
结构相似性指数 ↑	0.44	0.62	0.65	**0.67**

表1：人脸超分辨率（CelebA-HQ）上的峰值信噪比和结构相似性指数比较。我们的方法在这两项指标上均优于 SR3，同时参数量约少 4.58 亿，训练迭代所用的图像也更少。数据由 Saharia 等人提供 [41]。

优化器： 训练时，我们使用 AdamW [26] 优化器，其权重衰减为 10^{-4}，而非 Adam [19]。与其他基于扩散的方法不同，为节省额外计算，我们不对模型参数 [34] 使用指数移动平均（EMA）。尽管指数移动平均在提升去噪扩散概率模型的质量方面非常有效，我们的方法在未使用 EMA 的情况下仍优于 SR3 和 SRDiff。

人脸超分辨率细节： 在我们的 $16 \times 16 \rightarrow 128 \times 128$ 人脸超分辨率实验中，我们采用了比 SR3 更小的配置。我们将通道维度从 128 降至 64，并将 ResNet 块的数量从 3 个减至 2 个，因此参数总量约为 92M，而非 550M。对于$64 \times 64 \rightarrow 512 \times 512$ 人脸超分辨率，我们采用了与 SR3 相同的架构设置（625M 参数），以便对视觉示例进行公平的主观比较。学习率设为 1×10^{-4}。此外，批大小也从 256 降至 4，这是在 A100 GPU 上运行实验所必需的。

通用超分辨率细节： 在与 SRDiff 的比较中，我们采用了规模较小但具有可比性的架构，通道大小为 $[1, 2, 2, 4]$, 48，包含两个 ResNet 块，模型参数量为 9.3M，而 SRDiff 为 12M。此外，较小的模型规模使我们能够使用更大的批大小 256 进行训练。学习率设为 2×10^{-5}。

4.3 结果

本节展示了我们提出的方法在面部超分辨率和通用超分辨率任务上的定量与定性结果。我们使用峰值信噪比（PSNR）、结构相似性指数（SSIM）以及基于 AlexNet 的学习感知图像块相似度（LPIPS）等标准指标，将本方法的表现与当前最先进的扩散方法 SR3 和 SRDiff 进行比较 [33, 53, 21]。较高的 PSNR 和 SSIM 值表示图像质量较好；相反，较低的 LPIPS 值表示感知质量较好。



（a）低分辨率　（b）超分辨率　（c）高分辨率

图3：对比一张低分辨率、超分辨率和高分辨率图像（CelebA-HQ），展示了我们提出的方法在 $16 \times 16 \rightarrow 128 \times 128$ 设置下的效果。低分辨率图像存在明显的信息损失，尤其是嘴部前方手指的呈现。我们提出的方法能够重建出细节丰富的图像，尤其是头发细节。然而，高分辨率图像表明，我们的方法无法重建该手指。此外，高分辨率图像中的边缘更加清晰，眼睛的细节也更加锐利。



(a) 低分辨率　(b) 超分辨率　(c) 高分辨率

图4：$64 \times 64 \rightarrow 512 \times 512$ 实验中三张不同人脸图像（CelebA-HQ）的低分辨率、高分辨率和超分辨率（我们的方法）示例结果。最后一行显示，我们的模型生成了连续的皮肤纹理，但与真实值中的痣和痘痘等细微细节不符。

4.3.1 人脸超分辨率

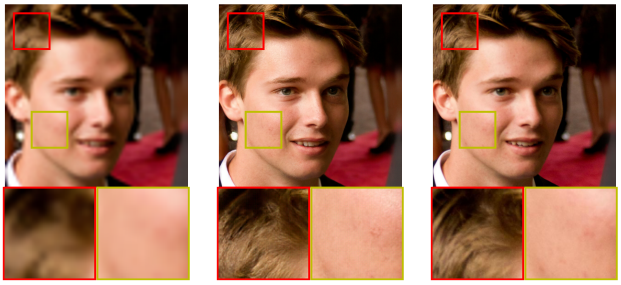
表1显示，我们的方法在 CelebA-HQ 和 $16 \times 16 \rightarrow 128 \times 128$ 设置上优于 SR3 及其他生成方法。尽管我们的模型处理的样本更少（受批大小限制），且参数规模显著更小（9200万，而非5.5亿），但它



图 5：使用我们的方法在人脸超分辨率任务中得到的中间去噪结果（$64 \times 64 \rightarrow 512 \times 512$）。左上角的图像是低分辨率输入。第一行中间的图像是初始预测器的估计结果。其余图像展示了去噪函数在迭代应用过程中生成的中间去噪估计结果，顺序为从左到右、从上到下。网格右下角是去噪函数的最终预测结果。

在峰值信噪比方面比 SR3 高 0.3 dB，在结构相似性指数方面高 0.02。图3展示了我们方法的重建图像示例。这表明，当低分辨率输入图像的信息不足（例如低分辨率图像中的手指）且需要推断或合成额外细节时，使用离散小波变换是有益的。

图4 和图5 展示了我们提出的方法在 $64 \times 64 \rightarrow 512 \times 512$ 场景中的表现。图4展示了三个重建示例，并突显了我们方法的优势与局限。与 $16 \times 16 \rightarrow 128 \times 128$ 重建结果相比，我们的方法能生成更逼真的高质量高分辨率结果，并能更高效地利用低分辨率图像中的额外信息。如图所示，它能生成平滑的皮肤纹理，效果与 SR3 相似，但难以保留痣、粉刺或穿孔饰品等细小细节。



（a）低分辨率　（b）超分辨率　（c）高分辨率

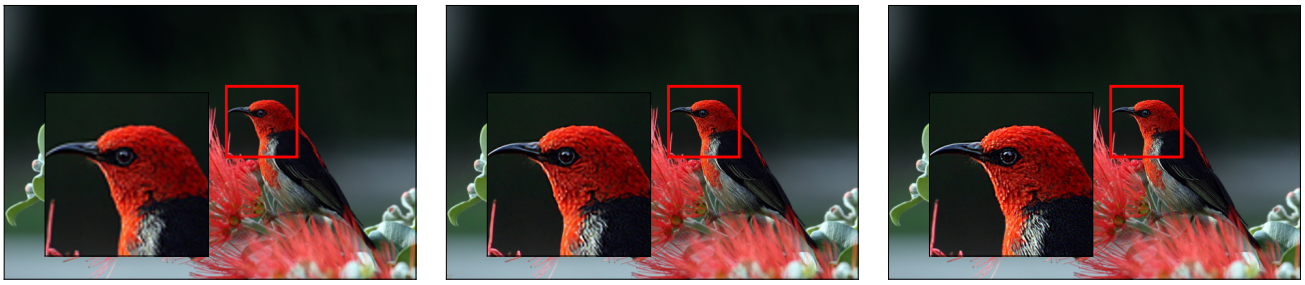
图 6：低分辨率、超分辨率和高分辨率人脸图像中放大区域的并排对比（$64 \times 64 \rightarrow 512 \times 512$，CelebA-HQ）。我们的方法生成的超分辨率图像展现了更清晰的头发细节，可见更多发丝和纹理。然而，与高分辨率图像相比，超分辨率图像在皮肤结构方面仍有不足，尤其是对痘痘和皮肤纹理的呈现。

图5 展示了中间去噪结果以及 $64 \times 64 \rightarrow 512 \times 512$ 场景下初始预测器的重建结果。结果表明，初始预测器提供了有力的基线预测，但对于高分辨率图像而言，还需要更细致的细节。因此，这说明高频细节是在扩散过程中生成的。此外，与真实值相比，初始预测中的颜色更为饱和，反差也更强。遗憾的是，由于作者未在论文中提供 $64 \times 64 \rightarrow 512 \times 512$ 的定量结果（峰值信噪比或结构相似性指数），因此无法直接与 SR3 进行比较。由于按照完全相同的规格训练 SR3 模型对硬件要求很高，我们无法复现其结果。

图 6 展示了一张人脸图像及其两个放大区域的对比。图 6。观察头发可以发现，我们的超分辨率图像比高分辨率图像呈现出更精细的细节。超分辨率图像中的发丝更加清晰突出。与图 4类似，在观察皮肤结构时，高分辨率图像对毛孔和痘痘的呈现效果更好。

4.3.2 通用超分辨率

表2 展示了我们提出的方法在DIV2K验证集上进行4倍缩放的结果。与人脸超分辨率不同，我们没有将测试图像调整为固定大小。我们根据峰值信噪比、结构相似性指数和学习感知图像块相似度，将我们的方法与SRDiff进行比较。请注意，表2 包含了基于回归的方法，其峰值信噪比和结构相似性指数高于生成方法 [41]。训练10万步后，我们的方法在学习感知图像块相似度方面优于所有受测的当前最先进生成方法（RankSRGAN [54], ESRGAN [48], 和SRFlow [27]），并且



(a) 低分辨率　(b) 超分辨率　(c) 高分辨率

图 7:　低分辨率、超分辨率和高分辨率图像的并排比较（4倍缩放），选自 DIV2K 验证集　t.

方法	峰值信噪比 ↑	结构相似性指数 ↑	学习感知图像块相似度 ↓
双三次插值	26.70	0.77	0.409
EDSR [24]	28.98	0.83	0.270
RRDB [48]	29.44	0.84	0.253
RankSRGAN [54]	26.55	0.75	0.128
ESRGAN [48]	26.22	0.75	0.124
SRFlow [27]	27.09	0.76	0.120
SRDiff [23]	27.41	**0.79**	0.136
DiWa（我们的方法）	**28.09**	0.78	**0.104**

表2：DIV2K 验证集上通用图像 4×超分辨率的结果。我们的方法在峰值信噪比和学习感知图像块相似度方面优于所有生成式方法，在结构相似性指数方面取得第二好的结果。请注意，EDSR 和 RRDB 属于基于回归的方法，其峰值信噪比和结构相似性指数得分通常高于生成式方法 [41]。

在学习感知图像块相似度和峰值信噪比方面优于SRDiff。就结构相似性指数而言，我们的方法也取得了具有竞争力的0.78分，比SRDiff的最佳结果低0.01。

表3总结了我们的方法与当前最先进的生成式方法在Set5、Set14、BSDS100和General100上的比较。表3除Set5外，我们的方法在所有数据集上的学习感知图像块相似度（LPIPS）方面均优于所比较的方法；在Set5上，我们取得了第二好的结果。两项实验中较低的LPIPS值表明，与其他生成式方法相比，我们的方法在感知相似性判断方面更有效。

使用我们的方法得到的重建图像示例见图7。结合定量测量结果可以看出，我们的模型能够细化输入图像，同时保留其纹理（见放大区域）。

在收敛速度方面，SRDiff 使用 100k 次迭代对其编码器结构进行预训练，并结合 300k 次迭代训练去噪函数。相比之下，我们仅使用少量的 100k 次迭代，且未进行编码器预训练。此外，我们的参数比 SRDiff 少 2.7M 。因此，我们的模型在不牺牲泛化能力的情况下，展现出更快的收敛速度。

我们对通用超分辨率的观察结果支持Wang等人 [46] 关于卷积神经网络能够利用高频细节进行图像分类的发现。他们发现，分类准确率较高的卷积神经网络会利用输入图像中更多的高频分量。由于LPIPS是通过卷积神经网络分类器计算的，有效地生成高频细节有助于超分辨率任务获得较低的LPIPS值。因此，我们的方法能够有效保留生成的高分辨率图像中的高频信息，从而提升卷积神经网络分类器的性能，这一点也得到了较低LPIPS值的印证。

我们还发现，部分重建结果的左下角含有少量高斯噪声；在固定空间尺寸的人脸超分辨率设置中，这一现象并不明显，但在图7中也能观察到。我们推测，这很可能源于网络架构的位置编码。移除位置编码或增加训练迭代次数或许可以缓解这一问题。

4.3.3 消融研究

我们进行消融研究，以探究初始预测器和二维离散小波变换的影响。我们在通用超分辨率任务（$48 \times 48 \rightarrow 192 \times 192$）上评估了不同变体，即 Set5，并在 DIV2K 上以 64 的批大小对每种变体训练 10 万次迭代。其余所有超参数均与其他实验相同。

表4 展示了消融研究的结果。结果表明，初始预测器和二维离散小波变换这两个组件各自都能显著提升最终重建质量，峰值信噪比和结构相似性指数的测量结果也证实了这一点。两者结合时，峰值信噪比表现最佳，但与仅使用初始预测器相比，结构相似性指数略有下降。

值得注意的是，我们的方法目标并非完美复现高分辨率图像，而是提高

方法（生成式）	Set5 [4]	Set14 [52]	BSDS100 [30]	General100 [9]
双三次插值	0.3407	0.4393	0.5249	0.3528
SFTGAN [47]	0.0890	0.1481	0.1769	0.1030
SRGAN [22]	0.0882	0.1663	0.1980	0.1055
ESRGAN [48]	0.0748	0.1329	0.1614	0.0879
NatSR [42]	0.0939	0.1758	0.2114	0.1117
SPSR [28]	**0.0644**	0.1318	0.1611	0.0863
DiWa（我们的方法）	0.0747	**0.1143**	**0.1398**	**0.0783**

表3：与当前最先进生成式超分辨率方法在 x4 倍放大下的学习感知图像块相似度比较。

方法	峰值信噪比 ↑	结构相似性指数 ↑
SR3（基线）	22.74	0.6363
SR3 + 2D-离散小波变换	26.94	0.7150
SR3 +初始化预测器	27.02	**0.7540**
SR3 + 2D-DWT + 初始化预测器	**27.37**	0.7220

表4：我们的方法在通用超分辨率任务上的消融实验，在 Set5 上进行评估，并在 DIV2K 上训练 100k 次迭代（$48 \times 48 \rightarrow 192 \times 192$）。两个组件都对基线模型有积极作用，而将它们融合则结合了各自的优势，从而取得最高的峰值信噪比。

尽可能提高 SR 图像的分辨率，同时保留高频细节。由于 SR 问题是不适定的，这一目标也是当前研究的共同方向。该比较凸显了生成式超分辨率中的权衡，而我们提出的方法在保留精细细节与图像整体自然度之间取得了平衡。

5 未来工作

未来的工作中，我们的方法仍面临一些挑战，这些挑战在 SR3 中也很明显，有待进一步研究，例如如何保留精细皮肤纹理细节（如痣、痘痘和穿孔饰品）。这些局限性部分源于病态问题的定义，在将我们的方法应用于真实世界场景时应予以解决。尽管如此，将这一方法应用于多级离散小波变换（类似于 MWCNN [25]）将会很有意思。结合潜在扩散 [38, 37, 36, 40], , 这将拓宽参与类似 SR3 中实验的途径。此外，我们预计使用指数移动平均（EMA）还能带来进一步改进，但这会增加训练时间。在架构方面，进一步探索初始预测器是未来研究的一个有趣方向 [25, 55, 51]。

6 结论

在这项工作中，我们提出了一种新颖的扩散-小波（DiWa）图像超分辨率方法，结合了条件扩散模型和小波分解的优势。我们在两个人脸超分辨率任务（8 倍放大）上将该方法与 SR3 进行了评估。对于通用超分辨率（4 倍放大），我们将该方法与 SRDiff 以及非扩散式生成方法进行了比较。实验结果表明，在这两项任务中，我们的方法在峰值信噪比（PSNR）、结构相似性指数（SSIM）和学习感知图像块相似度（LPIPS）方面均优于当前最先进的生成式技术。

此外，我们的方法依托小波域，从而提高了推理速度。由于小波域的空间尺寸仅为图像空间的四分之一，我们的方法能够减少内存消耗并加快处理速度。此外，离散小波变换缩小了去噪函数所需的感受野，进一步提升了推理速度和收敛速度。

这些优化使我们的方法以仅 92M 个参数取得当前最先进的结果；相比之下，SR3 需要 550M 个参数，SRDiff 需要 12M 个参数。因此，我们的方法不仅有效，而且轻量，特别适合难以获得高性能硬件的研究人员。它既能生成高质量的图像重建结果，也提供了一种切实可行的方法，可轻松复现并集成到现有的扩散流程中。

这项工作的影响超越了单图像超分辨率领域。凭借更快的推理速度和收敛速度、更低的内存消耗以及更小的参数规模，我们的方法非常适合在各种应用中广泛采用，包括实时超分辨率，以及图像分类、医学成像、多图像超分辨率、卫星图像和文本到图像生成等下游任务的预处理步骤 [11, 3, 33]。此外，我们的方法在学习感知图像块相似度指标上的低评估值表明，该方法可作为图像分类的预处理步骤，具有应用价值。

致谢

本工作得到德国联邦教育与研究部资助的 XAINES 项目（资助编号 01IW20005）以及卡尔蔡司基金会通过 SustainableEmbedded AI 项目（P2021-02-009）提供的资助。

参考文献

[1] Eirikur Agustsson 和 Radu Timofte。NTIRE 2017 单图像超分辨率挑战：数据集与研究。载于 IEEE计算机视觉与模式识别会议研讨会论文集，第 126–135 页，2017年。4[2] Ali N Akansu 和 Richard A Haddad。《多分辨率信号分解：变换、子带与小波》。学术出版社，2001年。2[3] Syed Muhammad Arsalan Bashir、Yi Wang、Mahrukh Khan 和 Yilong Niu。基于深度学习的单图像超分辨率综合综述。《PeerJ 计算机科学》，7:e621，2021年。8[4] Marco Bevilacqua、Aline Roumy、Christine Guillemot 和 Marie Line Alberi-Morel。基于非负邻域嵌入的低复杂度单图像超分辨率。2012年。4, 8[5] Andrew Brock、Jeff Donahue 和 Karen Simonyan。用于高保真自然图像合成的大规模生成对抗网络训练。arXiv 预印本 arXiv:1809.11096，2018年。4[6] Yu Chen、Ying Tai、Xiaoming Liu、Chunhua Shen 和 Jian Yang。FSRNet：利用人脸先验进行端到端人脸超分辨率学习。载于 IEEE计算机视觉与模式识别会议论文集，第 2492–2501 页，2018年。5[7] Hyungjin Chung、Eun Sun Lee 和 Jong Chul Ye。利用正则化反向扩散进行 MR 图像去噪与超分辨率。IEEE 医学影像汇刊，2022年。1[8] Chao Dong、Chen Change Loy、Kaiming He 和 Xiaoou Tang。利用深度卷积网络进行图像超分辨率。IEEE模式分析与机器智能汇刊，38(2):295–307，2015年。4[9] Chao Dong、Chen Change Loy 和 Xiaoou Tang。加速超分辨率卷积神经网络。载于《计算机视觉—ECCV 2016：第 14 届欧洲会议》，阿姆斯特丹，荷兰，2016年10月11日至14日，论文集，第二部分 14，第 391–407 页。施普林格，2016年。4, 8[10] Marcelo Dos Santos、Rayson Laroca、Rafael O Ribeiro、Jo˜ao Neves、Hugo Proenc¸a 和 David Menotti。利用随机微分方程进行人脸超分辨率。载于《2022年第35届 SIBGRAPI 图形学、模式与图像会议（SIBGRAPI）》，第 1 卷，第 216–221 页。IEEE，2022年。1[11] Stanislav Frolov、Tobias Hinz、Federico Raue、J¨orn Hees 和 Andreas Dengel。对抗式文本到图像合成：综述。arXiv 预印本 arXiv:2101.09983，2021年。8[12] Deep Ganguli、Danny Hernandez、Liane Lovitt、Amanda Askell、Yuntao Bai、Anna Chen、Tom Conerly、Nova Das-sarma、Dawn Drain、Nelson Elhage 等。可预测性与大型生成模型中的惊喜。在2022年ACM公平性、问责制与透明度会议中，第1747–1764页，2022年。 1[13] Tiantong Guo、Hojjat Seyed Mousavi、Tiep Huu Vu和Vishal Monga。用于图像超分辨率的深度小波预测。在IEEE计算机视觉与模式识别会议研讨会论文集，第104–113页，2017年。 1, 3, 4[14] Alfred Haar。正交函数系理论。《数学年刊》，71(1)：38–53，1911 年。 2, 3[15] Kaiming He、Xiangyu Zhang、Shaoqing Ren和Jian Sun。用于图像识别的深度残差学习。在IEEE计算机视觉与模式识别会议论文集，第770–778页，2016年。 1, 4[16] Jonathan Ho、Ajay Jain和Pieter Abbeel。去噪扩散概率模型。《神经信息处理系统进展》，33：6840–6851，2020年。 2, 4[17] Tero Karras、Timo Aila、Samuli Laine和Jaakko Lehtinen。用于提升质量、稳定性和多样性的GAN渐进式增长。arXiv预印本 arXiv:1710.10196，2017年。 4[18] Tero Karras、Samuli Laine和Timo Aila。基于风格的生成对抗网络生成器架构。在IEEE/CVF计算机视觉与模式识别会议论文集，第4401–4410页，2019年。 4[19] Diederik P Kingma和Jimmy Ba。Adam：一种随机优化方法。arXiv预印本 arXiv:1412 6980，2014年 5 。 .[20] Zhifeng Kong、Wei Ping、Jiaji Huang、Kexin Zhao和Bryan Catanzaro。DiffWave：一种用于音频合成的通用扩散模型。arXiv预印本 arXiv:2009.09761，2020年。 1[21] Alex Krizhevsky、Ilya Sutskever和Geoffrey E Hinton。使用深度卷积网络进行ImageNet分类。《美国计算机学会通讯》，60(6)：84–90，2017年。 5[22] Christian Ledig、Lucas Theis、Ferenc Husz´ar、Jose Caballero、Andrew Cunningham、Alejandro Acosta、Andrew Aitken、Alykhan Tejani、Johannes Totz、Zehan Wang等。使用生成对抗网络实现逼真单图像超分辨率。在IEEE计算机视觉与模式识别会议论文集，第4681 4690页，2017年 1, 4, 8 – 。 [23] Haoying Li、Yifan Yang、Meng Chang、Shiqi Chen、Huajun Feng、Zhihai Xu、Qi Li和Yueting Chen。SRDiff：使用扩散概率模型进行单图像超分辨率。《Neurocomputing》，479：47–59，2022年。 1, 4, 7[24] Bee Lim、Sanghyun Son、Heewon Kim、Seungjun Nah和Kyoung Mu Lee。用于单图像超分辨率的增强型深度残差网络。在IEEE计算机视觉与模式识别会议研讨会论文集，第136–144页，2017年。 4, 7[25] Pengju Liu、Hongzhi Zhang、Kai Zhang、Liang Lin和Wangmeng Zuo。用于图像复原的多级小波卷积神经网络。在IEEE计算机视觉与模式识别会议研讨会论文集，第773 782页，2018年 2, 8 – 。 [26] Ilya Loshchilov和Frank Hutter。解耦权重衰减正则化。arXiv预印本 arXiv:1711.05101，2017年。5

[27] Andreas Lugmayr、Martin Danelljan、Luc Van Gool和Radu Timofte。SRFlow：使用归一化流学习超分辨率空间。在欧洲计算机视觉会议中，第715–732页。施普林格，2020年。 4, 6, 7[28] Cheng Ma、Yongming Rao、Yean Cheng、Ce Chen、Jiwen Lu和Jie Zhou。具有梯度引导的结构保持超分辨率。在IEEE/CVF计算机视觉与模式识别会议论文集，第7769–7778页，2020年。 1, 4, 8[29] St´ephane Mallat。信号处理小波之旅。Elsevier，1999年。1[30] David Martin、Charless Fowlkes、Doron Tal和Jitendra Malik。人工分割的自然图像数据库及其在评估分割算法和测量生态统计中的应用。在第八届IEEE国际计算机视觉会议论文集。ICCV 2001，第2卷，第416–423页。IEEE，2001年。 4, 8[31] The Mathworks, Inc.，马萨诸塞州纳蒂克。MATLAB版本9.3.0.713579（R2017b），2017年。 4[32] Sachit Menon、Alexandru Damian、Shijia Hu、Nikhil Ravi和Cynthia Rudin。Pulse：通过生成模型潜在空间探索进行自监督照片上采样。在IEEE/CVF计算机视觉与模式识别会议论文集，第

2437 2445, 2020 5 –页。

[33] Brian B. Moser、Federico Raue、Stanislav Frolov、Sebastian Palacio、J¨orn Hees和Andreas Dengel。超分辨率指南：介绍与最新进展。《IEEE模式分析与机器智能汇刊》，第1–21页，2023年。 1, 4, 5, 8[34] Alexander Quinn Nichol和Prafulla Dhariwal。改进的去噪扩散概率模型。在国际机器学习会议中，第8162–8171页。PMLR，2021年。 2, 5[35] Hao Phung、Quan Dao和Anh Tran。小波扩散模型是快速且可扩展的图像生成器。arXiv预印本arXiv:2211.16152，2022年。 1[36] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu和Mark Chen。利用CLIP潜变量进行分层文本条件图像生成。arXiv预印本 arXiv:2204 06125，2022年 8 。 .[37] Aditya Ramesh、Mikhail Pavlov、Gabriel Goh、Scott Gray、Chelsea Voss、Alec Radford、Mark Chen和Ilya Sutskever。零样本文本生成图像。在国际机器学习会议中，第8821–8831页。PMLR，2021年。 8[38]Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser和Bj¨orn Ommer。使用潜在扩散模型进行高分辨率图像合成。在IEEE/CVF计算机视觉与模式识别会议论文集，第

10684 10695, 2022 8 –页。

[39] Olaf Ronneberger、Philipp Fischer和Thomas Brox。U-Net：用于生物医学图像分割的卷积网络。在医学图像计算与计算机辅助干预——MICCAI 2015：第18届国际会议，德国慕尼黑，2015年10月5日至9日，论文集，第III部分18，第234–241页。施普林格，2015年。 4[40]Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、S Sara Mahdavi、Rapha Gontijo Lopes 等。结合深度语言理解的逼真文本到图像扩散模型。arXiv 预印本 arXiv:2205.11487，2022年。8[41] Chitwan Saharia、Jonathan Ho、William Chan、Tim Salimans、David J Fleet 和 Mohammad Norouzi。通过迭代细化实现图像超分辨率。IEEE模式分析与机器智能汇刊，2022年。1、2、4、5、6、7[42] Jae Woong Soh、Gu Yong Park、Junho Jo 和 Nam Ik Cho。通过显式自然流形判别实现自然逼真的单幅图像超分辨率。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第8122–8131页，2019年。1、4、8[43] Jascha Sohl-Dickstein、Eric Weiss、Niru Maheswaranathan 和 Surya Ganguli。基于非平衡热力学的深度无监督学习。载于国际机器学习会议论文集，第2256–2265页。PMLR，2015年。2[44] Nitish Srivastava、Geoffrey Hinton、Alex Krizhevsky、Ilya Sutskever 和 Ruslan Salakhutdinov。Dropout：一种防止神经网络过拟合的简单方法。《机器学习研究杂志》，15(1):1929–1958，2014年。4[45] Michael Thompson。Rafael C. Gonzalez 和 Paul Wintz 所著《数字图像处理》。Leonardo，14(3):256–257，1981年。2[46] Haohan Wang、Xindi Wu、Zeyi Huang 和 Eric P Xing。高频分量有助于解释卷积神经网络的泛化能力。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第8684–8694页，2020年。7[47] Xintao Wang、Ke Yu、Chao Dong 和 Chen Change Loy。通过深度空间特征变换恢复图像超分辨率中的逼真纹理。载于IEEE计算机视觉与模式识别会议论文集，第606–615页，2018年。1、4、8[48] Xintao Wang、Ke Yu、Shixiang Wu、Jinjin Gu、Yihao Liu、Chao Dong、Yu Qiao 和 Chen Change Loy。ESRGAN：增强型超分辨率生成对抗网络。载于欧洲计算机视觉会议（ECCV）研讨会论文集，第0–0页，2018年。1、4、6、7、8[49] Jay Whang、Mauricio Delbracio、Hossein Talebi、Chitwan Saharia、Alexandros G Dimakis 和 Peyman Milanfar。通过随机细化实现去模糊。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第16293–16303页，2022年。1、4[50] Zhisheng Xiao、Karsten Kreis 和 Arash Vahdat。通过去噪扩散 GAN 应对生成式学习三难困境。arXiv 预印本 arXiv:2112.07804，2021年。1[51]Shengke Xue、Wenyuan Qiu、Fan Liu 和 Xinyu Jin。用于图像超分辨率的基于小波的残差注意力网络。Neurocomputing，382:116–126，2020年。8[52] Roman Zeyde、Michael Elad 和 Matan Protter。关于基于稀疏表示的单幅图像放大。载于国际曲线与曲面会议论文集，第711–730页。施普林格，2010年。4、8[53] Richard Zhang、Phillip Isola、Alexei A Efros、Eli Shechtman 和 Oliver Wang。《感知度量中的深度特征具有惊人的有效性》。

作为感知度量的深度特征。载于*IEEE*计算机视觉与模式识别会议论文集，第586–595页，2018年。1、5[54] Wenlong Zhang、Yihao Liu、Chao Dong 和 Yu Qiao。RankSRGAN：结合排序器的生成对抗网络，用于图像超分辨率。载于*IEEE/CVF*国际计算机视觉会议论文集，第3096–3105页，2019年。4、6、7[55] Wenbin Zou、Mingchao Jiang、Yunchen Zhang、Liang Chen、Zhiyong Lu 和 Yi Wu。SDWNet：一种结合小波变换、具有直扩张结构的网络，用于图像去模糊。载于*IEEE/CVF*国际计算机视觉会议论文集，第1895–1904页，2021年。8