DiT4SR: Taming Diffusion Transformer for Real-World Image Super-Resolution

Zheng-Peng Duan[1,2] * Jiawei Zhang[2] Xin Jin[1] Ziheng Zhang[1] Zheng Xiong[2]
Dongqing Zou[2,3] Jimmy S. Ren[2,4] Chunle Guo[1,5] Chongyi Li[1,5] †
[1]VCIP, CS, Nankai University [2]SenseTime Research [3]PBVR
[4]Hong Kong Metropolitan University [5]NKIARI, Shenzhen Futian
https://adam-duan.github.io/projects/dit4sr/

Figure 1. 我们的方法 DiT4SR 与其他最先进方法在两张真实低分辨率图像上的性能对比。得益于 SD3 强大的生成能力以及我们精心的设计（能够在输入的低分辨率信息与生成过程之间进行充分交互），我们的方法在保持保真度的同时生成了更加逼真的细节，尤其体现在 文本、结构和细节方面。

Abstract

大规模预训练扩散模型由于其丰富的生成先验，正逐渐在解决真实场景图像超分辨率（*Real-ISR*）问题中变得越来越受欢迎。 近年来，扩散 *Transformer*（*DiT*）的发展在图像生成任务中展现出远超传统基于 *UNet* 架构的性能，这也引出了一个问题：我们能否将先进的基于 *DiT* 的扩散模型应用于 *Real-ISR*？ 为此，我们提出了 *DiT4SR*，这是率先探索如何将大规模 *DiT* 模型用于 *Real-ISR* 的工作之一。 不同于像 *ControlNet* 那样直接注入从低分辨率（*LR*）图像中提取的嵌入，我们将 *LR* 嵌入整合进 *DiT* 原始的注意力机制中，从而实现了 *LR* 潜在表示与生成潜在表示之间的双向信息流动。 这两条信息流的充分交互使得 *LR* 流能够随扩散过程不断演化，在每一个扩散步骤中逐步产生更加精细的引导，从而与生成潜在表示更好地对齐。 此外，我们通过跨流卷积层将 *LR* 引导注入到生成潜在表示中，以弥补 *DiT* 在捕获局部信息方面的局限性。 这些简单而有

*This project is done during the internship at SenseTime Research.
†Corresponding author.

效的设计赋予了 *DiT* 模型在 *Real-ISR* 中的卓越性能，这一点也通过大量实验证明。

1. 介绍

Real-ISR [55, 69] 旨在从其低分辨率（LR）版本中恢复高分辨率（HR）图像，并能够处理各种类型的退化，例如压缩、模糊和噪声。 与传统的 ISR [19] 不同，Real-ISR 要求模型不仅能够去除复杂的退化，还需要生成感知上真实的细节，以提升视觉质量。 该任务的高度病态性对模型的先验知识提出了要求。若缺乏这类先验，模型容易生成模糊不清的结果，并导致恢复质量下降。 因此，研究者们将注意力转向了大规模预训练的文本到图像（T2I）模型 [11, 44]，尤其是 Stable Diffusion（SD），它们在数十亿张高质量的自然图像上进行训练，蕴含了丰富的真实世界先验知识。 现有方法主要基于 UNet 架构的 SD 模型，即 SD1 [43]、SD2 [43] 和 SDXL [39]，并将输入的 LR 图像作为条件，通过 ControlNet [70] 或类似的方式注入 LR 信息，从而生成对应的 HR 图像。

最近，扩散 Transformer（DiT） [38] 的发展使得基于 DiT 的架构逐渐流行。得益于 SD3 [21] 和 Flux [3] 的卓越性能，MM-DiT 已被证明是生成模型中一种有效的 Transformer 模块。 它包含两条可学习的流，分别用于视觉特征和文本标记，并通过注意力操作在两种模态之间实现双向信息流动。基于这种架构创新，以及扩展规模等技术，基于 DiT 的扩散模型在细节生成和图像质量方面表现出令人期待的性能。 基于这些优势，一个关键问题出现了：我们能否将先进的基于 *DiT* 的扩散模型应用于 *Real-ISR*?

一个直观的解决方案是采用 ControlNet [70] 作为 DiT 的控制器，将 LR 信息注入以引导 Real-ISR 的生成过程，如 Figure 2 (a) 所示[1]。 具体而言，若干 MM-DiT 模块被复制为一个可训练的副本，用于初始化 ControlNet。 LR 潜变量是通过将 LR 图像输入预训练的 VAE 编码器获得的，并作为输入使用。随后，ControlNet 中每个模块输出的隐藏状态通过一个可训练的卷积层直接加入到 SD3 的噪声流中。 然而，由于 ControlNet 最初是为基于 UNet 的架构设计的，直接添加 LR 信息忽视了 DiT 的独特特性，从而限制了信息交互，并可能导致性能受限。

为了充分利用 DiT 的优势，我们放弃了诸如 ControlNet 之类的传统技术路线，并提出了一种专为 DiT 设计的新控制架构，称为 DiT4SR。 不同于在额外创建的 DiT 模块中处理 LR 信息（如 Figure 2 (a) 所示），我们将 LR 流直接融入到 DiT 模块中，从而实现更高效的 LR 信息与扩散过程的交互。 考虑到 LR 潜变量和带噪潜变量本质上都是视觉信息的形式， 我们为 LR 流采用了与噪声流相似的设计。 在每个 MM-DiT 模块中，我们的方法复制了噪声流的模块来处理 LR 输入。

[1]SD3-ControlNet 的实现遵循 Diffusers [50] 提供的默认代码和配置。

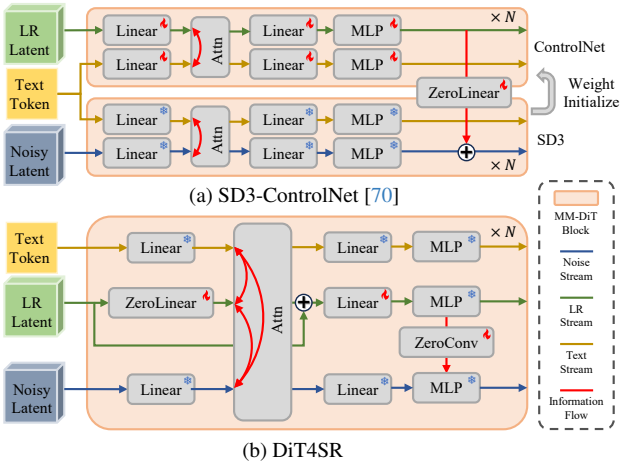


Figure 2. SD3-ControlNet 与我们提出的 DiT4SR 的网络结构对比。跨流的信息流以 红色 线标注，箭头表示方向。值得注意的是，我们的 DiT4SR 实现了双向信息交互，使得 LR 流与噪声流能够持续交互并共同演化，而 SD3-ControlNet 依赖于单向信息流，限制了交互能力。

这种 注意力中的 **LR** 融合 允许 LR 流与噪声流之间进行双向信息交互，如 Figure 2 (b) 所示。

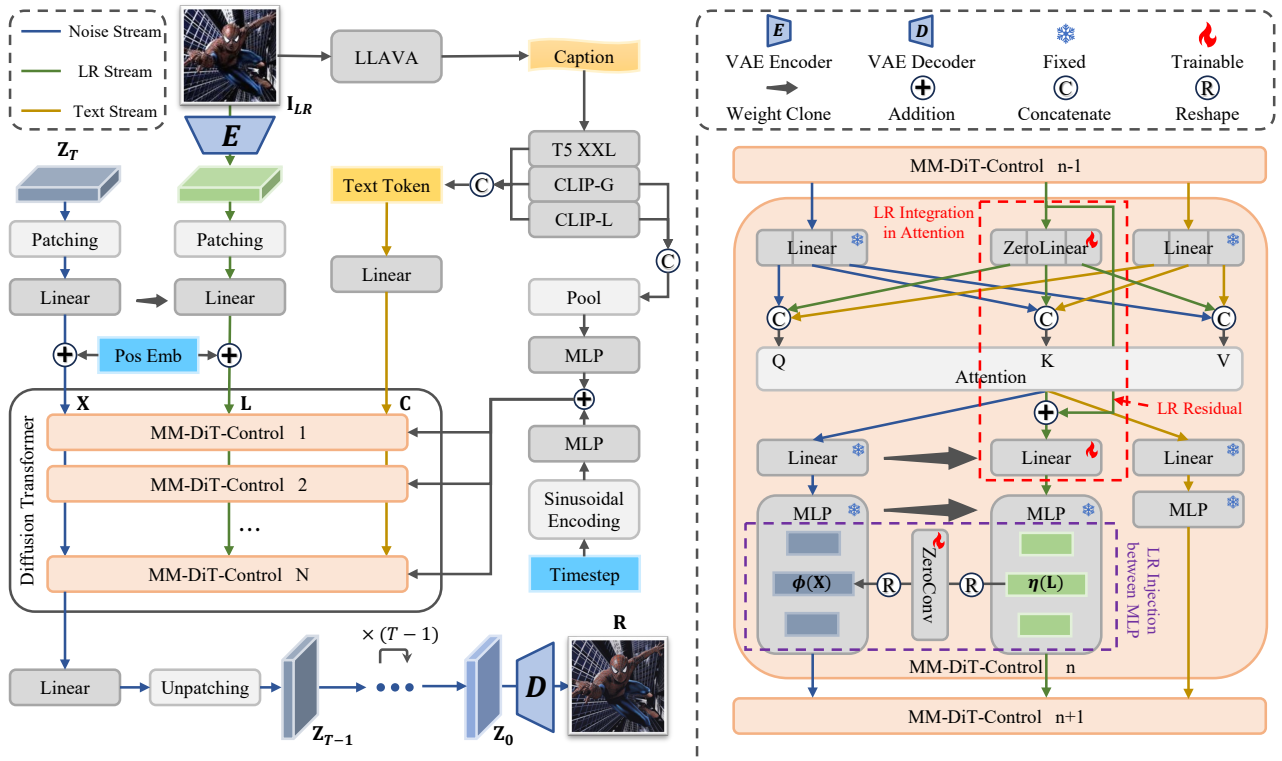
两条流之间的双向交互使得 LR 流能够随着扩散过程共同演化，从而逐步生成更精细、具备上下文感知的引导信息，更好地与生成的潜变量对齐。 为了保留 LR 信息并增强深层模块中的一致性， 我们在 LR 流的注意力模块内引入了额外的 **LR** 残差，直接连接输入与输出。 然而，由于注意力机制是全局操作，仅依赖其本身并不足以胜任超分辨率任务。 捕捉局部信息同样是恢复精细细节的关键。 因此，我们进一步通过卷积层将 LR 引导注入到噪声流中。 这种 **MLP** 之间的 **LR** 注入 使得模型能够聚合局部信息，弥补 DiT 在局部信息捕捉能力上的不足 [61]。

我们的 贡献 总结如下：

- 据我们所知，我们的方法是最早探索如何将大规模 DiT 模型应用于 Real-ISR 的工作之一。
- 我们并未像 ControlNet 那样复制额外模块，而是将 LR 流融入到原始的 DiT 模块中，从而实现 LR 引导与扩散过程之间的双向信息交互。
- 我们引入了一个卷积层，将 LR 引导注入到噪声流中，从而弥补 DiT 在局部信息捕捉能力上的不足。

2. 相关工作

基于深度学习的ISR方法已经取得了显著进展， 其架构从卷积网络 [18, 19, 34, 73–75] 发展到Transformer [10, 13, 14, 31]。 然而， 这些方法在应对真实世界场景下复杂退化以及任务本身的病态特性时依然存在困难。 为了解决前者问题， BSRGAN [69] 和 Real-ESRGAN [55] 探索了更复杂的退化模型。 针对病态特性， 一些基于GAN的方法 [6, 9, 33, 54] 被提出，以学习高质量图像的分布并生成感知上真实的细节。 然而，这些方法仍然受到训练不稳定的影响，并会产生不自然的视觉伪影 [8, 32, 62]。



(a) Overview of DiT4SR (b) One MM-DiT-Control Block

Figure 3. DiT4SR (a) 和 MM-DiT-Control 模块 (b) 的概览。我们的 DiT4SR 将 LR 流融合进 MM-DiT-Control 模块。注意力中的 **LR** 融合 允许 LR 流与 Noise 流之间的双向交互。我们引入了 **LR** 残差 来增强 LR 引导的一致性。此外，**MLP** 之间的 **LR** 注入通过一个可训练的卷积层将 LR 流的引导注入到 Noise 流中。

随着大规模文本到图像（T2I）模型（如SD [43]）的成功， 研究者们 [17, 22, 28, 40, 45, 46, 49, 51, 67] 将注意力转向利用这些强大的预训练模型中蕴含的生成先验来解决Real-ISR问题。 基于第一代UNet架构的SD [43]， StableSR [53] 和 DiffBIR [35] 通过ControlNet [70] 或类似ControlNet的方式注入LR信息作为条件。 PASD [65] 和 SeeSR [60] 进一步结合高层语义信息来引导扩散过程。 在更先进的SDXL中，SUPIR 和 FaithDiff [12] 探讨了在Real-ISR中扩大模型规模的效果。 同时，还有一些其他方法 [7, 29, 36, 37, 47, 56, 59, 63, 68] 致力于提高扩散过程的效率。 随着DiT的发展， DiT-SR [15] 从零开始训练基于DiT的SR模型。 DreamClear [2] 和 DPIR [27] 提出了基于DiT的图像恢复模型， 但依然采用ControlNet注入LR信息， 这限制了其充分发挥DiT优势的能力。 从网络架构的角度来看，我们的DiT4SR是少数开创性工作之一，能够驯服大规模扩散Transformer模型以用于Real-ISR，同时摒弃了ControlNet式的方式。 同时期还有两种方法 [20, 30]， 它们从流轨迹的角度探索一步式的基于DiT的SR模型。

3. 方法

给定一个经历复杂退化的低分辨率图像 $\mathbf{I}_{LR}$， Real-ISR 的目标是恢复对应的高分辨率图像 $\mathbf{R}$。 在去除退化的同时， Real-ISR 还要求模型生成逼真的细节， 从而提升视觉感知效果。 为克服该任务的病态特性， 研究者们将方法构建在大规模预训练的 T2I 模型之上，特别是 SD，以利用其中丰富的生成先验。 随着 DiT 的最新发展逐渐取代传统的基于 UNet 的扩散模型，在细节生成和视觉质量方面已取得了最先进的表现。在本研究中， 我们专注于利用 DiT 的优势，并将其驯化于 Real-ISR 任务。

为此，我们提出了基于 DiT 的 SD3 的 DiT4SR。其整体架构概览在 Section 3.1 中介绍。 不同于 SD3-ControlNet 那样创建额外的 DiT 模块来处理 LR 信息，我们将 LR 流直接融入到 DiT 原有的注意力计算中，并允许其与噪声流和文本流进行双向交互， 细节在 Section 3.2 中描述。 然而，仅依靠注意力机制还不足以让 DiT 胜任 Real-ISR。 为了注入更多的 LR 引导并捕获 LR 特征的局部信息，我们通过深度卷积层将 LR 流中 MLP 的中间特征引入到噪声流的 MLP 中，相关细节在 Section 3.3 中给出。

3.1. 架构概览

由于我们的 DiT4SR 建立在流行的基于 DiT 架构的 SD3 之上， 我们首先简要介绍 SD3。 与以往的 SD 模型 [43] 相同， SD3 也在潜在空间中进行扩散过程。它由一系列 MM-DiT 模块组成，其中不同的权重集合

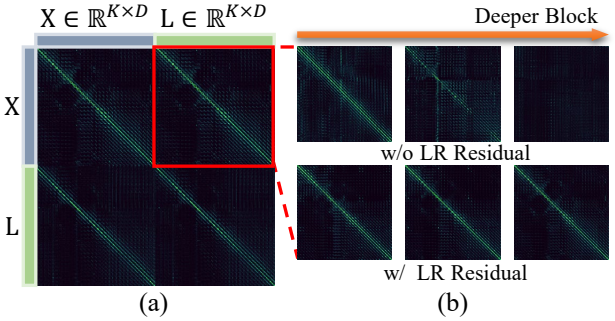


Figure 4. (a) 在第 24 个 MM-DiT-Control 中，噪声图像 token $\mathbf{X}$ 和 LR 图像 token $\mathbf{L}$ 的四个注意力图可视化（$\mathbf{X} \to \mathbf{X}$，$\mathbf{X} \to \mathbf{L}$，$\mathbf{L} \to \mathbf{X}$，$\mathbf{L} \to \mathbf{L}$）。在 $\mathbf{X} \to \mathbf{L}$ 和 $\mathbf{L} \to \mathbf{X}$ 的对角线表示 $\mathbf{X}$ 与 $\mathbf{L}$ 之间的信息交互。(b) 带有和不带有 LR 残差的 $\mathbf{X} \to \mathbf{L}$ 注意力图。在没有 LR 残差的情况下，随着网络块深度的增加（第 1、13 和 24 个 MM-DiT-Control），LR 引导会逐渐减弱。引入 LR 残差能够显著增强 LR 引导的一致性。

分别处理文本和图像嵌入，形成噪声流和文本流，如 Figure 2 (a) 底部所示。该结构为每种模态分别采用两个独立的 Transformer，同时通过注意力机制融合其序列，从而实现跨模态交互。双向交互使得两个流在整个扩散过程中能够共同演化，这是 DiT 的关键优势。如果采用 ControlNet [50, 70] 作为 Real-ISR 的 DiT 控制器，如 Figure 2 (a) 所示，SD3-ControlNet 会在额外的 DiT 模块中处理 LR 流，并通过可训练的线性层将 LR 嵌入直接注入到噪声流中。这种方法只建立了从 LR 流到噪声流的单向信息流，限制了信息交互。相比之下，我们的 DiT4SR 将 LR 流直接融入到原始的 DiT 模块中，如 Figure 2 (b) 所示。这种设计实现了双向信息流，使得 LR 流能够在整个扩散过程中不断自适应，并生成与不断演化的噪声流更有效对齐的引导。

如 Figure 3 所示，我们的 DiT4SR 与 SD3 具有相似的架构。具体来说，在输入扩散 Transformer 之前，带噪声的潜变量 $\mathbf{Z} \in \mathbb{R}^{H \times W \times C}$ 首先被展平成长度为 K 的块序列，其中 $K = \frac{H}{2} \cdot \frac{W}{2}$，然后通过一个线性层投影到 D 维空间。随后，将位置嵌入加入到得到的带噪图像 token $\mathbf{X} \in \mathbb{R}^{K \times D}$ 中。为了引入 LR 信息，我们首先通过预训练的 VAE 编码器将 LR 图像 $\mathbf{I}_{LR}$ 编码到潜在空间中。考虑到 LR 潜变量和带噪潜变量都是视觉表征的形式，我们采用相同的处理流程，并加入相同的位置嵌入，从而得到 LR 图像 token $\mathbf{L} \in \mathbb{R}^{K \times D}$。参考 SD3 [21]，用于描述 $\mathbf{I}_{LR}$ 的输入的文本由三个预训练的文本模型进行编码，包括 CLIP-L [41]、CLIP-G [16] 和 T5 XXL [42]。两个 CLIP 模型输出的文本表征会进行池化，并与时间步 t 结合，用于调制 DiT 的内部特征。除了池化后的文本表征，文本 token $\mathbf{C} \in \mathbb{R}^{M \times D}$ 也会被构建，长度为 M，由三个文本表征共同组成。

以带噪图像 token $\mathbf{X}$ 和文本 token $\mathbf{C}$ 作为输入，原始的 MM-DiT 模块为这两种模态分别采用独立的权重集合，分别称为噪声流和文本流。在我们的 DiT4SR 中，我们引入了一个额外的流，称为 LR 流，用于处理 LR 图像 token $\mathbf{L}$。MM-DiT 模块被修改为 MM-DiT-Control 模块，从而使得 LR 信息能够指导 HR 潜变量的生成。经过 N 个 MM-DiT-Control 模块以及 unpatch 操作后，噪声流输出该时间步 t 的去噪潜变量。重复扩散过程 T 步，并解码得到干净潜变量 $\mathbf{Z}_0$，最终我们可以得到所需的 HR 结果 $\mathbf{R}$。在接下来的两个部分中，我们将介绍 MM-DiT-Control 模块中的具体改动。

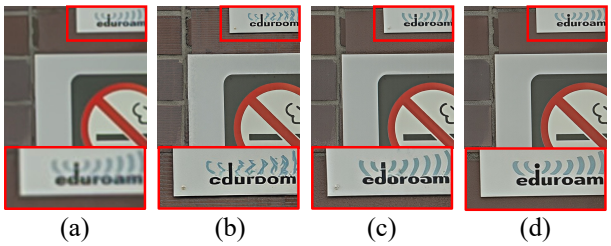


Figure 5. (a) 为 LR 输入。(b) 为在 MLP 之间去除 LR 注入后的结果。(c) 通过线性层注入 LR 信息。(d) 为我们的 DiT4SR 的结果，其将线性层替换为卷积层。卷积层有助于捕捉更精确的局部信息，在恢复精细结构时表现出显著优势。所有结果均是在相同设置下重新训练各自模型后获得的。

3.2. 注意力中的 LR 融合

遵循原始 MM-DiT 的设计，在 MM-DiT-Control 模块中，三个流依次经过联合注意力机制和 MLP 操作。

对于联合注意力，其输入可以表示为

$$\begin{aligned} \mathbf{Q} &= P_{\mathbf{Q}}^{\mathbf{X}}(\mathbf{X}) \copyright P_{\mathbf{Q}}^{\mathbf{L}}(\mathbf{L}) \copyright P_{\mathbf{Q}}^{\mathbf{C}}(\mathbf{C}), \\ \mathbf{K} &= P_{\mathbf{K}}^{\mathbf{X}}(\mathbf{X}) \copyright P_{\mathbf{K}}^{\mathbf{L}}(\mathbf{L}) \copyright P_{\mathbf{K}}^{\mathbf{C}}(\mathbf{C}), \\ \mathbf{V} &= P_{\mathbf{V}}^{\mathbf{X}}(\mathbf{X}) \copyright P_{\mathbf{V}}^{\mathbf{L}}(\mathbf{L}) \copyright P_{\mathbf{V}}^{\mathbf{C}}(\mathbf{C}), \end{aligned} \tag{1}$$

其中 $P_{\mathbf{Q}}^{\mathbf{X}}$、$P_{\mathbf{K}}^{\mathbf{X}}$、$P_{\mathbf{V}}^{\mathbf{X}}$、$P_{\mathbf{Q}}^{\mathbf{C}}$、$P_{\mathbf{K}}^{\mathbf{C}}$ 和 $P_{\mathbf{V}}^{\mathbf{C}}$ 是针对 $\mathbf{X}$ 和 $\mathbf{C}$ 的预训练固定线性投影。$\copyright$ 表示在 token 长度维度上的拼接操作。$P_{\mathbf{Q}}^{\mathbf{L}}$、$P_{\mathbf{K}}^{\mathbf{L}}$ 和 $P_{\mathbf{V}}^{\mathbf{L}}$ 是为 $\mathbf{L}$ 新建的可训练线性投影，其权重初始化为零。这样一来，在训练初期可以忽略 $\mathbf{L}$ 的影响，而其作用会随着训练逐渐增强。MM-DiT-Control 中的联合注意力计算为

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \underbrace{\text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})}_{\text{attention map}} \mathbf{V}, \tag{2}$$

这使得三个流之间能够进行更加全面的交互。我们进一步可视化了关于 LR 图像 token $\mathbf{X}$ 和噪声图像 token $\mathbf{L}$ 的注意力图，如 Figure 4 (a) 所示。我们观察到在自注意力区域和交叉注意力区域中，对角线均被明显激活，这表明在 $\mathbf{X}$ 和 $\mathbf{L}$ 对应位置之间存在信息交互。这种双向交互不仅使噪声流能够受到 LR 引导的影响，同时也使 LR 流能够根据噪声流的状态进行自适应，从而提供更准确且具有上下文感知的引导。

另一个值得注意的现象是，$\mathbf{L}$ 与 $\mathbf{X}$ 之间的信息交互在连续的注意力块中逐渐衰减，这在 Figure 4 (b) 中展示。可能的原因是，在联合注意力的作用下，LR 信

Datasets	Metrics	Real-ESRGAN	SwinIR	ResShift	StableSR	SeeSR	DiffBIR	OSEDiff	SUPIR	DreamClear	SD3-ControlNet	DiT4SR
DrealSR	LPIPS ↓	0.282	0.274	0.353	0.273	0.317	0.452	0.297	0.419	0.354	0.323	0.365
	MUSIQ ↑	54.267	52.737	52.392	58.512	65.077	65.665	64.692	59.744	44.047	55.956	64.950
	MANIQA ↑	0.490	0.475	0.476	0.559	0.605	0.629	0.590	0.552	0.455	0.545	0.627
	ClipIQA ↑	0.409	0.396	0.379	0.438	0.543	0.572	0.519	0.518	0.379	0.449	0.548
	LIQE ↑	2.927	2.745	2.798	3.243	4.126	3.894	3.942	3.728	2.401	3.059	3.964
RealSR	LPIPS ↓	0.271	0.254	0.316	0.306	0.299	0.347	0.292	0.357	0.325	0.305	0.319
	MUSIQ ↑	60.370	58.694	56.892	65.653	69.675	68.340	69.087	61.929	59.396	62.604	68.073
	MANIQA ↑	0.551	0.524	0.511	0.622	0.643	0.653	0.634	0.574	0.546	0.599	0.661
	ClipIQA ↑	0.432	0.422	0.407	0.472	0.577	0.586	0.552	0.543	0.474	0.484	0.550
	LIQE ↑	3.358	2.956	2.853	3.750	4.123	4.026	4.065	3.780	3.221	3.338	3.977
RealLR200	MUSIQ ↑	62.961	63.548	59.695	63.433	69.428	68.027	69.547	64.837	65.926	65.623	70.469
	MANIQA ↑	0.553	0.560	0.525	0.579	0.612	0.629	0.606	0.600	0.597	0.587	0.645
	ClipIQA ↑	0.451	0.463	0.452	0.458	0.566	0.582	0.551	0.524	0.546	0.526	0.588
	LIQE ↑	3.484	3.465	3.054	3.379	4.006	4.003	4.069	3.626	3.775	3.733	4.331
RealLQ250	MUSIQ ↑	62.514	63.371	59.337	56.858	70.556	69.876	69.580	66.016	66.693	66.385	71.832
	MANIQA ↑	0.524	0.534	0.500	0.504	0.594	0.624	0.578	0.584	0.585	0.568	0.632
	ClipIQA ↑	0.435	0.440	0.417	0.382	0.562	0.578	0.528	0.483	0.502	0.509	0.578
	LIQE ↑	3.341	3.280	2.753	2.719	4.005	4.003	3.904	3.605	3.688	3.639	4.356

Table 1. 在四个真实场景基准上的与最新 Real-ISR 方法的定量比较。最佳和次佳结果分别用 红色 和 蓝色 标记。我们的 DiT4SR 在四个基准上都取得了最佳或具有竞争力的性能。

Ours vs.	SeeSR	DiffBIR	SUPIR	DreamClear
Realism	82.1%	83.6%	81.7%	72.7%
Fidelity	68.9%	79.5%	75.4%	64.5%

Table 2. 真实场景数据上的用户研究结果。数字表示在图像真实感和保真度方面，我们的 DiT4SR 相比于对比方法的获胜率。

息会与噪声流一同演化，并可能受到不期望的扰动，从而逐渐削弱其对噪声流的引导效果。为了增强 LR 引导的一致性，我们额外引入了一条捷径，将输入的 LR 信息直接传递到联合注意力机制的输出。通过引入 **LR** 残差，可以在更深的 transformer 块中有效保留 LR 引导，确保其在整个扩散过程中对噪声流的持续影响。带有和不带有 LR 残差的模型之间的视觉对比展示在 Figure 7 (c) 和 (f) 中。

3.3. MLP之间的LR注入

由于联合注意力在全局层面上运行，并且仅依赖位置嵌入来提供空间信息，因此不足以有效地将 DiT 适配到 Real-ISR。这一局限性产生的原因是：局部信息在恢复精细细节时同样重要，而单纯的全局注意力可能会忽视这些信息。

如 Figure 5 (b) 所示，仅依赖联合注意力在恢复文本和保持图像保真度方面依然面临挑战。

为了加强 LR 信息的引导，我们进一步将 LR 流中的 MLP 中间特征与噪声流通过一个 3×3 深度卷积层连接，该卷积层的权重初始化为零。具体而言，在 LR 流和噪声流的 MLP 中，隐藏状态的维度首先被扩展为原始维度的 4 倍，然后通过两个线性投影

Model	LR Integation	LR Residual	LR Injection	MUSIQ ↑	MANIQA ↑
FULL	✔	✔	**Conv**	71.832	0.632
A	✗	✔	**Conv**	66.963	0.574
B	✔	✗	**Conv**	70.887	0.614
C	✔	✔	✗	71.202	0.610
D	✔	✔	**Linear**	71.607	0.621

Table 3. 在 RealLQ250 数据集上的 DiT4SR 消融实验结果。所有变体均在与完整模型相同的设置下进行训练。

映射回原始大小。我们将这两个中间特征分别记为 $\phi(\mathbf{X}) \in \mathbb{R}^{K \times 4D}$ 和 $\eta(\mathbf{L}) \in \mathbb{R}^{K \times 4D}$。我们首先将 $\eta(\mathbf{L})$ 重新整形为 $\eta(\mathbf{L})' \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 4D}$，然后将其输入到 3×3 的深度卷积层中。在重新整形成图像 token 的形式后，LR 信息便被有效地注入到噪声流中。

与此同时，3×3 的深度卷积层也弥补了 DiT 在捕获局部信息方面的不足。如前所述，联合注意力是在全局层面上计算的，因此缺乏来自 LR 信息的局部引导。即便使用一个线性层来注入 LR 引导，修复诸如文字等精细结构依然具有挑战性，如 Figure 5 (c) 所示。3×3 的深度卷积有助于从 LR 流中捕获更精确的局部信息，并将这些 LR 引导注入到噪声中，从而实现更好的性能，尤其在精细结构的恢复上，如 Figure 5 (d) 所示。

4. 实验

4.1. 实验设置

数据集. 遵循 SeeSR [60] 的设置，我们在训练过程中采用了来自 DIV2K [1]、DIV8K [23]、Flickr2K [48] 以及 FFHQ [25] 中前 10K 张人脸图像的组合。为了充

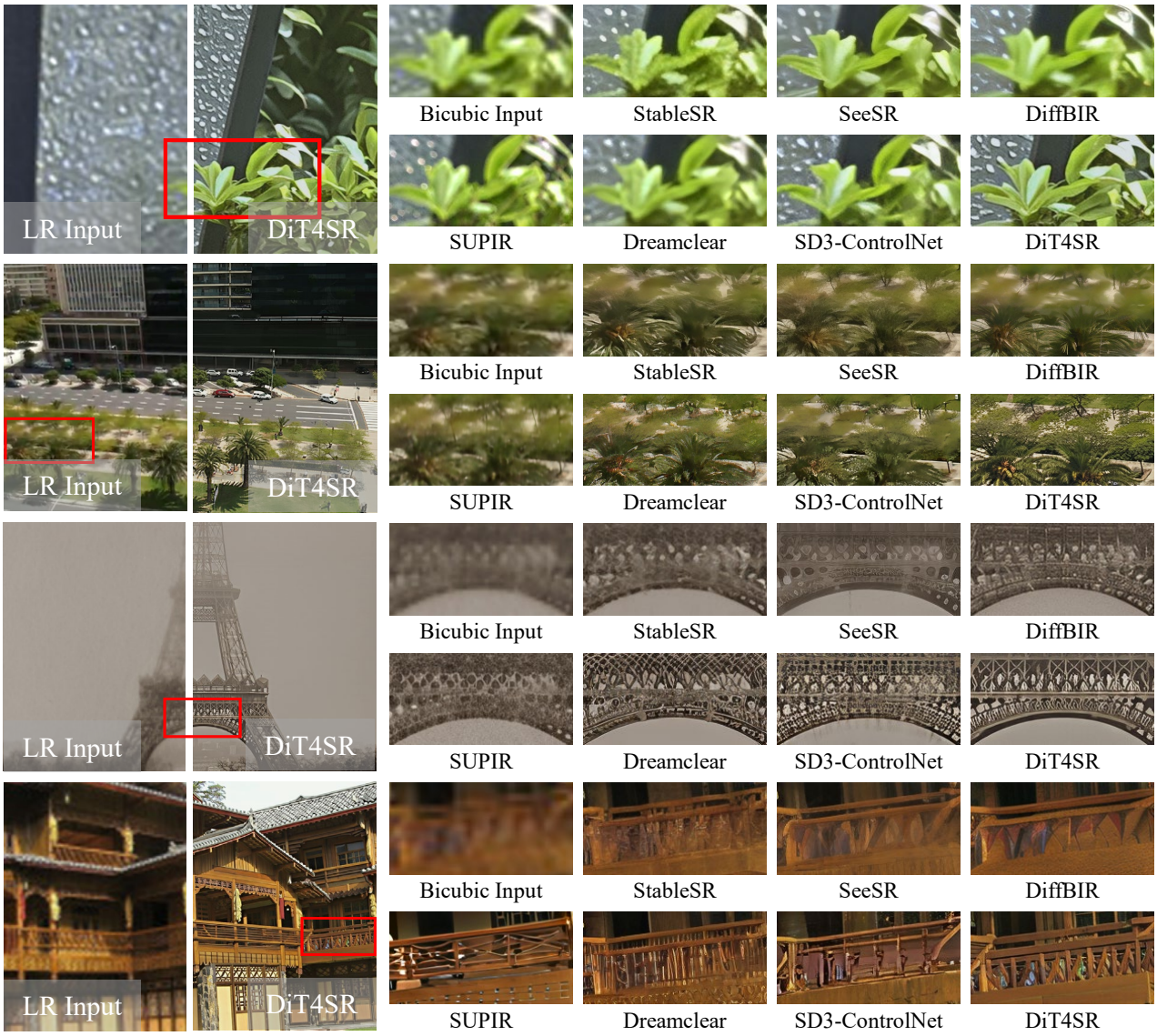


Figure 6. 在 RealSR（第一行）、RealLR200（第二行和第三行）以及 RealLQ250（第四行）上与最新的 Real-ISR 方法进行定性比较。我们的 DiT4SR 在图像真实感和细节生成方面取得了最佳表现，同时保持了对输入 LR 图像的保真度，特别是在细结构的保留上表现突出。更多的可视化结果可以在补充材料中找到。

分发挥我们方法的潜力并扩充训练数据集，我们另外加入了 1K 张自采集的高分辨率图像。训练过程中采用 Real-ESRGAN [55] 的退化流程来合成 LR-HR 训练对，其参数配置与 SeeSR 保持一致。需要注意的是，LR 与 HR 图像的分辨率分别设置为 128×128 和 512×512。

由于我们的方法专注于 Real-ISR 任务，我们在四个常用的真实场景数据集上对模型进行了评估，包括 DrealSR [58]、RealSR [5]、RealLR200 [60] 和 RealLQ250 [2]。所有实验均在放大因子 $\times 4$ 的条件下进行。DrealSR 和 RealSR 分别包含 93 张和 100 张图像。与 SeeSR 相同，这两个数据集均采用中心裁剪，LR 图像的分辨率设置为 128×128。RealLR200 是 SeeSR 提出的数据集，包含 200 张分辨率差异显著的图像。RealLQ250 是由 DreamClear [2] 构建的，包含 200 张固定分辨率为 256×256 的图像。需要指出的是，RealLR200 与 RealLQ250 都没有对应的 GT 图像。

评价指标. 正如先前研究所指出的 [4, 24, 66]，全参考指标如 PSNR 与 SSIM [57] 难以准确反映恢复结果的视觉效果。因此，大多数研究仅使用感知度量 LPIPS [71] 来衡量图像保真度。我们在用户研究中额外加入了对图像保真度的主观评估。我们还使用 MUSIQ [26]、MANIQA [64]、ClipIQA [52] 和 LIQE [72] 作为无参考指标来衡量图像质量。

实现细节. 已在补充材料中给出。

4.2. 和其他方法的对比

我们将我们的方法与最先进的Real-ISR方法进行了比较，这些方法包括基于GAN的方法（*i.e.* Real-ESRGAN [55] 和 SwinIR [31]）、基于扩散的UNet架构方法（*i.e.* ResShift [67], StableSR [53], SeeSR [60], DiffBIR [35], OSEDiff [59], 和 SUPIR [66]）、以及基于DiT架构的扩散方法（*i.e.* DreamClear [2] 和 SD3-ControlNet）。SD3-ControlNet基于SD3.5-medium参数初始化，采用Diffusers [50]提供的默认配置，并在与我们方法相同的设置下进行训练。

定量比较. 在四个基准数据集上的最先进Real-ISR方法的定量比较结果展示在Table 1中。在DrealSR和RealSR数据集上，虽然SeeSR和DiffBIR表现出强大的性能，但我们的方法取得了有竞争力的结果，接近或超过了它们的表现。在RealLR200和RealLQ250数据集上，我们的方法表现出压倒性的优势，在所有无参考指标上都取得了最佳表现。这些结果凸显了我们的方法在生成高质量修复结果方面的能力。

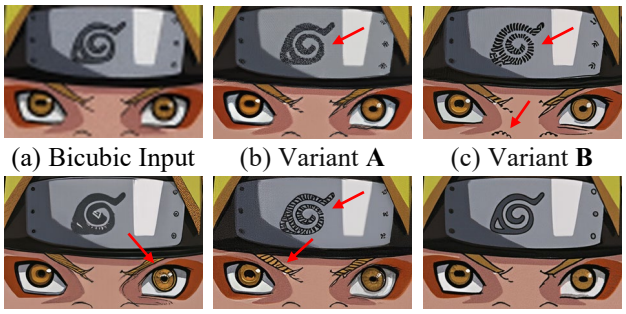
定性比较. 与其他方法的定性比较结果如Figure 6所示。从前两行可以看出，即使在遇到严重模糊退化的情况下，我们的方法也能够生成清晰度更高、细节更丰富的结果，相较于其他方法更具优势。这归功于我们的方法充分利用了SD3出色的生成能力。此外，如最后两行所示，我们的方法在处理精细结构（如建筑结构）方面展现出明显优势。值得注意的是，即使是同样基于SD3构建的SD3-ControlNet，在这些方面的表现也不如我们的方法。这进一步凸显了我们控制机制相较于ControlNet的优越性，更全面的信息交互使得模型能够更好地利用LR信息，从而生成高质量、保真度更高的修复结果。

用户研究. 为了进一步验证修复质量，我们邀请了80位志愿者进行了用户研究。我们从四个数据集（DrealSR、RealSR、RealLR200 和 RealLQ250）中随机选择了60张LR图像，并采用四种最新的方法（SeeSR、DiffBIR、SUPIR 和 DreamClear）进行比较。在用户研究中，每次评估会向参与者展示三张图像：原始LR输入图像、由我们的方法生成的修复结果，以及由随机选择的另一种方法生成的修复结果。参与者需要回答两个问题：(1) 哪个修复结果具有更高的图像真实感？(2) 哪个修复结果对原始图像内容的保真度更高？ Table 2中报告的结果表明，我们的方法优于其他方法。

5. 消融实验

为了进一步验证各个组件的有效性，我们在RealLQ250 数据集上进行了消融实验，并采用 MUSIQ 和 MANIQA 作为评估指标。所有变体均在与完整模型相同的设置下训练，以保证公平的比较。每个变体的详细框架可在补充材料中找到。

LR 集成的有效性. 为了检验注意力机制中 LR 集成的有效性，我们在保持其他组件不变的情况下，从注意力计算中移除了 LR 流。从 Table 3 报告的结果可以看出，MUSIQ 和 MANIQA 均有显著下降，这表明仅依赖于在 MLP 层之间注入 LR 是不足以生成高质量结果的。Figure 7 (b) 也表明，如果没有 LR 流与生成潜变量之间的双向信息交互，退化问题无法被有效去除。这是因为在缺乏这种信息交互时，LR 引导无法根据不断演化的噪声潜变量进行自适应调整，从而限制了模型的自适应修复能力。持续的交互使得 LR 信息能够逐步调整并提供更准确的引导，这对于有效应对复杂退化至关重要。



(a) Bicubic Input (b) Variant **A** (c) Variant **B**

(d) Variant **C** (e) Variant **D** (f) Full Model

Figure 7. 用于消融实验的可视化比较。变体 **A**、**B** 和 **C** 分别移除了 LR 集成、LR 残差和 LR 注入。变体 **D** 在 LR 注入中将卷积层替换为线性层。

LR 残差的有效性. 我们引入 LR 残差的目的在于保持深层 DiT 模块中 LR 引导的一致性。如 Table 3 所示，在没有 LR 残差的情况下，MUSIQ 和 MANIQA 的性能均出现下降。从 Figure 7 (c) 可以看到，结果中包含一些明显的伪影，从而降低了图像内容的保真度。这一问题主要源于 LR 流的不稳定演化，可能会受到不期望的干扰。通过引入 LR 残差，我们的 DiT4SR 有效地稳定了 LR 流，生成了具有更高保真度的结果。

LR 注入的有效性. 首先，我们移除了 MLP 层之间的 LR 注入。在 Table 3 中可以观察到评估指标有轻微下降。这表明仅依赖于注意力机制中的 LR 集成可以生成勉强可接受的结果。然而，从 Figure 7 (d) 中，特别是在眼睛区域，可以观察到明显的内容失真。这主要是因为注意力机制是全局操作，而 SR 任务还依赖于局部信息来准确恢复细节。我们还将 3×3 深度可分卷积替换为线性层。尽管 Table 3 显示线性层与卷积在性能上相近，但 Figure 7 (e) 表明伪影和失真问题并未得到缓解。3×3 深度可分卷积有效补偿了 DiT 在局部信息捕捉能力上的不足，并显著提升了我们 DiT4SR 的保真度，而这一点在无参考指标数据中无法完全体现。

6. 结论

在本文，我们提出了 DiT4SR，这是首批将大规模 DiT 模型应用于真实图像超分辨率（Real-ISR）的探索性工作之一。与 ControlNet 直接注入从低分辨率（LR）图像中提取的嵌入不同，我们的方法将 LR 嵌入整合进 DiT 的原始注意力机制中。这使得 LR 潜变量与生成潜变量之间能够进行双向信息流动。此外，我们

引入了一个跨流卷积层，将 LR 引导信息注入到生成潜变量中。这一设计不仅增强了 LR 引导，还弥补了 DiT 在捕获局部特征方面的不足。通过这些改进，DiT4SR 在 Real-ISR 上取得了优越的性能，并通过大量实验证明了其有效性。我们的工作突出了利用 DiT 进行高质量图像恢复的潜力，为未来研究开辟了新的方向。

7. 致谢

本工作部分受中国国家自然科学基金（62306153，62225604）、天津市自然科学基金（24JCJQJC00020）、中国科协青年人才托举工程（YESS20240686）、中央高校基本科研业务费（南开大学，070-63243143）、以及深圳市科技计划（JCYJ20240813114237048）的资助。计算设备部分由南开大学超级计算中心（NKSC）支持。

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, 2017. 5

[2] Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. In *NeurIPS*, 2025. 1, 3, 6, 7

[3] blackforestlabs.ai. Flux, offering state-of-the-art performance image generation, 2024. Accessed: 2024-10-07. 2

[4] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *CVPR*, 2018. 6

[5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *ICCV*, 2019. 6

[6] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *CVPR*, 2021. 2

[7] Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression for real-world image super-resolution. *arXiv preprint arXiv:2411.13383*, 2024. 3

[8] Chaofeng Chen, Xinyu Shi, Yipeng Qin, Xiaoming Li, Xiaoguang Han, Tao Yang, and Shihui Guo. Real-world blind super-resolution via feature matching with implicit high-resolution priors. In *ACM MM*, 2022. 2

[9] Du Chen, Jie Liang, Xindong Zhang, Ming Liu, Hui Zeng, and Lei Zhang. Human guided ground-truth generation for realistic image super-resolution. In *CVPR*, 2023. 2

[10] 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

[11] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 2

[12] Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *CVPR*, 2025. 3

[13] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, 2023. 2

[14] Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, and Fisher Yu. Dual aggregation transformer for image super-resolution. In *ICCV*, 2023. 2

[15] Kun Cheng, Lei Yu, Zhijun Tu, Xiao He, Liyu Chen, Yong Guo, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Effective diffusion transformer architecture for image super-resolution. In *AAAI*, 2025. 3

[16] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 4

[17] Qinpeng Cui, Yixuan Liu, Xinyi Zhang, Qiqi Bao, Zhongdao Wang, Qingmin Liao, Li Wang, Tian Lu, and Emad Barsoum. Taming diffusion prior for image super-resolution with domain shift sdes. *arXiv preprint arXiv:2409.17778*, 2024. 3

[18] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, 2019. 2

[19] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, 2014. 2

[20] Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *CVPR*, 2025. 3

[21] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 2, 4

[22] Junhao Gu, Peng-Tao Jiang, Hao Zhang, Mi Zhou, Jinwei Chen, Wenming Yang, and Bo Li. Consissr: Delving deep into consistency in diffusion-based image super-resolution. *arXiv preprint arXiv:2410.13807*, 2024. 3

[23] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. In *ICCVW*, 2019. 5

[24] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *ECCV*, 2020. 6

[25] Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019. 5

[26] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021. 6

[27] Dehong Kong, Fan Li, Zhixin Wang, Jiaqi Xu, Renjing Pei, Wenbo Li, and WenQi Ren. Dual prompting image restoration with diffusion transformers. In *CVPR*, 2025. 3

[28] Hao Li, Xiang Chen, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Foundir: Unleashing million-scale training data to advance foundation models for image restoration. *arXiv preprint arXiv:2412.01427*, 2024. 3

[29] Jianze Li, Jiezhang Cao, Zichen Zou, Xiongfei Su, Xin Yuan, Yulun Zhang, Yong Guo, and Xiaokang Yang. Distillation-free one-step diffusion for real-world image super-resolution. *arXiv preprint arXiv:2410.04224*, 2024. 3

[30] Jianze Li, Jiezhang Cao, Yong Guo, Wenbo Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation. *arXiv preprint arXiv:2502.01993*, 2025. 3

[31] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, 2021. 2, 7

[32] Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *CVPR*, 2022. 2

[33] Jie Liang, Hui Zeng, and Lei Zhang. Efficient and degradation-adaptive network for real-world image super-resolution. In *ECCV*, 2022. 2

[34] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, 2017. 2

[35] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, 2024. 1, 3, 7

[36] Yong Liu, Hang Dong, Jinshan Pan, Qingji Dong, Kai Chen, Rongxiang Zhang, Lean Fu, and Fei Wang. Patchscaler: An efficient patch-independent diffusion model for super-resolution. *arXiv preprint arXiv:2405.17158*, 2024. 3

[37] Mehdi Noroozi, Isma Hadji, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. You only need one step: Fast super-resolution with stable diffusion via scale distillation. In *ECCV*, 2024. 3

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[39] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[40] Yunpeng Qu, Kun Yuan, Kai Zhao, Qizhi Xie, Jinhua Hao, Ming Sun, and Chao Zhou. Xpsr: Cross-modal priors for diffusion-based image super-resolution. In *ECCV*, 2024. 3

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4

[42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 4

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3

[44] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 2

[45] Haoze Sun, Wenbo Li, Jianzhuang Liu, Haoyu Chen, Renjing Pei, Xueyi Zou, Youliang Yan, and Yujiu Yang. Coser: Bridging image and language for cognitive super-resolution. In *CVPR*, 2024. 3

[46] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. *arXiv preprint arXiv:2412.03017*, 2024. 3

[47] Lingchen Sun, Rongyuan Wu, Zhengqiang Zhang, Hongwei Yong, and Lei Zhang. Improving the stability of diffusion models for content consistent super-resolution. *CoRR*, 2024. 3

[48] 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 *CVPRW*, 2017. 5

[49] Li-Yuan Tsao, Hao-Wei Chen, Hao-Wei Chung, Deqing Sun, Chun-Yi Lee, Kelvin CK Chan, and Ming-Hsuan Yang. Holisdip: Image super-resolution via holistic semantics and diffusion prior. *arXiv preprint arXiv:2411.18662*, 2024. 3

[50] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 2, 4, 7

[51] Yuhao Wan, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, Ming-Ming Cheng, and Bo Li. Clearsr: Latent low-resolution image embeddings help diffusion-based real-world super resolution models see clearer. *arXiv preprint arXiv:2410.14279*, 2024. 3

[52] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 6

[53] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 2024. 3, 7

[54] 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 *ECCVW*, 2018. 2

[55] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, 2021. 2, 6, 7

[56] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C

Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *CVPR*, 2024. 3

[57] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 6

[58] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *ECCV*, 2020. 6

[59] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *arXiv preprint arXiv:2406.08177*, 2024. 3, 7

[60] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024. 1, 3, 5, 6, 7

[61] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 2

[62] Liangbin Xie, Xintao Wang, Xiangyu Chen, Gen Li, Ying Shan, Jiantao Zhou, and Chao Dong. Desra: detect and delete the artifacts of gan-based real-world super-resolution models. *arXiv preprint arXiv:2307.02457*, 2023. 2

[63] Rui Xie, Chen Zhao, Kai Zhang, Zhenyu Zhang, Jun Zhou, Jian Yang, and Ying Tai. Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *arXiv preprint arXiv:2404.01717*, 2024. 3

[64] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, 2022. 6

[65] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *ECCV*, 2024. 3

[66] Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *CVPR*, 2024. 1, 6, 7

[67] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *NeurIPS*, 2024. 3, 7

[68] Aiping Zhang, Zongsheng Yue, Renjing Pei, Wenqi Ren, and Xiaochun Cao. Degradation-guided one-step image super-resolution with diffusion priors. *arXiv preprint arXiv:2409.17058*, 2024. 3

[69] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021. 2

[70] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *CVPR*, 2023. 1, 2, 3, 4

[71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[72] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *CVPR*, 2023. 6

[73] Xindong Zhang, Hui Zeng, Shi Guo, and Lei Zhang. Efficient long-range attention network for image super-resolution. In *ECCV*, 2022. 2

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

[75] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, 2018. 2