

# 基于 JIT 的 diffusion 超分辨模型

---

李孟霖

December 29, 2025

# Introduction

---

# Introduction

---

## 研究背景

# 主流超分辨模型范式

## 回归型方法 (Pixel-wise Regression)

- EDSR (CNN): 以 MSE / L1 为目标, PSNR/SSIM 表现稳定
- SwinIR (Transformer): 更强建模能力, 当前 SR 指标基线

## 生成型方法 (Generative Models)

- GAN-based SR (如 SRGAN): 提升感知质量, 但训练不稳定、指标与视觉存在权衡
- Diffusion-based SR (如 SR3、StableSR): 逐步去噪, **高频细节表达能力强**

## 动机

Diffusion 模型在生成任务中往往具有较强的高频细节建模能力, 本任务中尝试将 Diffusion + Transformer backbone 引入超分辨任务

## 常见去噪结构

- UNet：多尺度卷积结构，局部归纳偏置强，在超分辨等低层视觉任务中被广泛采用
- DiT / JiT (Transformer-based)：使用 ViT 结构作为去噪网络，通过 self-attention 建模全局依赖

## Transformer 的潜在优势

- **全局建模能力强**：self-attention 可直接建模远距离像素/patch 关系
- **结构灵活**：天然适合 token-level 条件注入（时间、类别或图像条件）
- **统一建模范式**：避免多尺度卷积结构的复杂设计

基于以上考虑，本研究采用 Transformer 作为 Diffusion 的去噪模块，并进一步研究其在 **图像条件超分辨率任务**中的表现与局限。

## 三种参数化方式

- $\epsilon$ -prediction: 直接预测噪声  $\epsilon$ , 是早期 Diffusion 模型的常见形式
- $x$ -prediction: 预测原始干净图像  $x_0$ , 更直观, 常用于高分辨率生成
- $v$ -prediction: 预测速度变量  $v$ , 在不同噪声强度下具有更稳定的梯度尺度

## 本任务中的实现方式

- 网络前向输出采用  $x$ -prediction
- 训练时通过固定变换, 将  $x$ -prediction 转换为  $v$ -prediction loss
- 该做法与 JiT / EDM 等工作保持一致

## v-prediction 的定义

Diffusion 前向过程定义为：

$$x_t = \alpha_t x_0 + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

v-prediction 定义为：

$$v_t = \alpha_t \epsilon - \sigma_t x_0$$

在训练中，模型预测  $\hat{x}_0$ ，并通过固定变换得到  $\hat{v}_t$ ，使用 v-prediction loss 进行优化。

# Introduction

---

## 相关工作

何恺明团队 CFG 条件生成模型 (JiT), Back to Basics: Let Denoising Generative Models Denoise

低维流形假设:JiT 研究认为, 真实分布处在整个图像空间中的低维流形中。

以 vit 为核心去噪模块替代 UNet, 模型做  $x$ -prediction,  $x$ -prediction 的预测值变换为  $v$ -prediction 构造损失函数

# JiT: x-prediction 与 v-prediction

前向加噪过程

$$x_t = t x_0 + (1 - t) \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

模型预测  $\hat{x}_0$ , 并通过固定变换构造  $v$ :

$$\hat{v}_t = \frac{\hat{x}_0 - x_t}{1 - t}, \quad v_t = \frac{x_0 - x_t}{1 - t}$$

训练目标:

$$\mathcal{L} = \mathbb{E} \left[ \|\hat{v}_t - v_t\|_2^2 \right]$$

## Methods

---

主体与 JIT 相同, 去噪模块做  $x$ -prediction, 换算为  $v$ -prediction 构造 loss

去噪模块为多个 vit 模块的串联, 串联长度由超参数控制

删去 CFGtoken 拼接, 尝试按下述两种方式将低清图信息引入

## 低清图插入方式

- 将低清图和带噪声图像按通道拼接
- 将低清图编码为 token, 和带噪声图像拼接成一个序列

# 实验

---

# 实验

---

## 性能指标

# 通道拼接 (Concat) 方案：性能指标

PSNR: 6.98      SSIM: 0.146

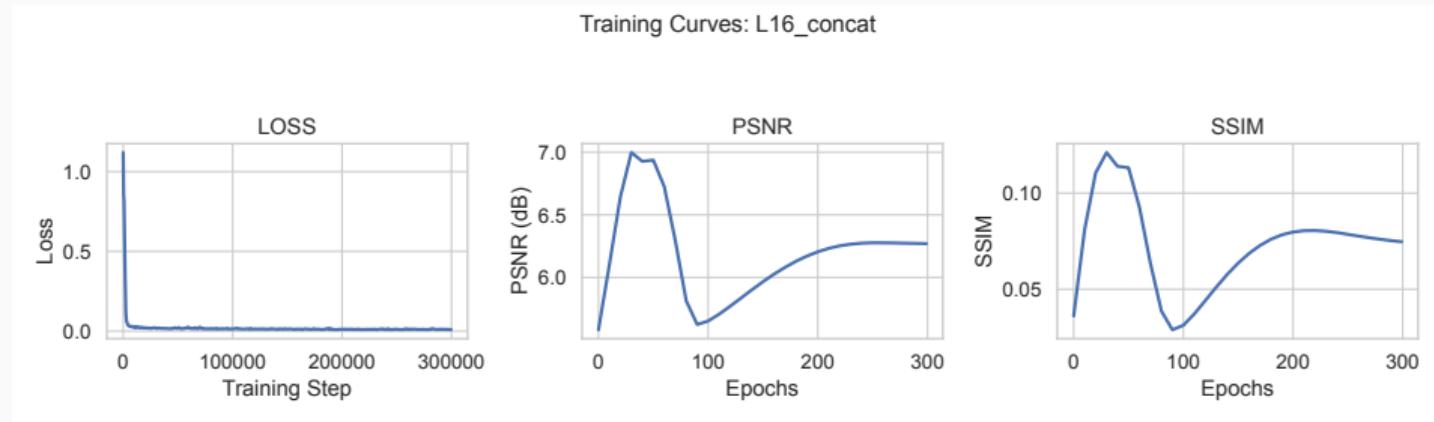


Figure 1. L16 通道拼接方案在验证集上的性能曲线

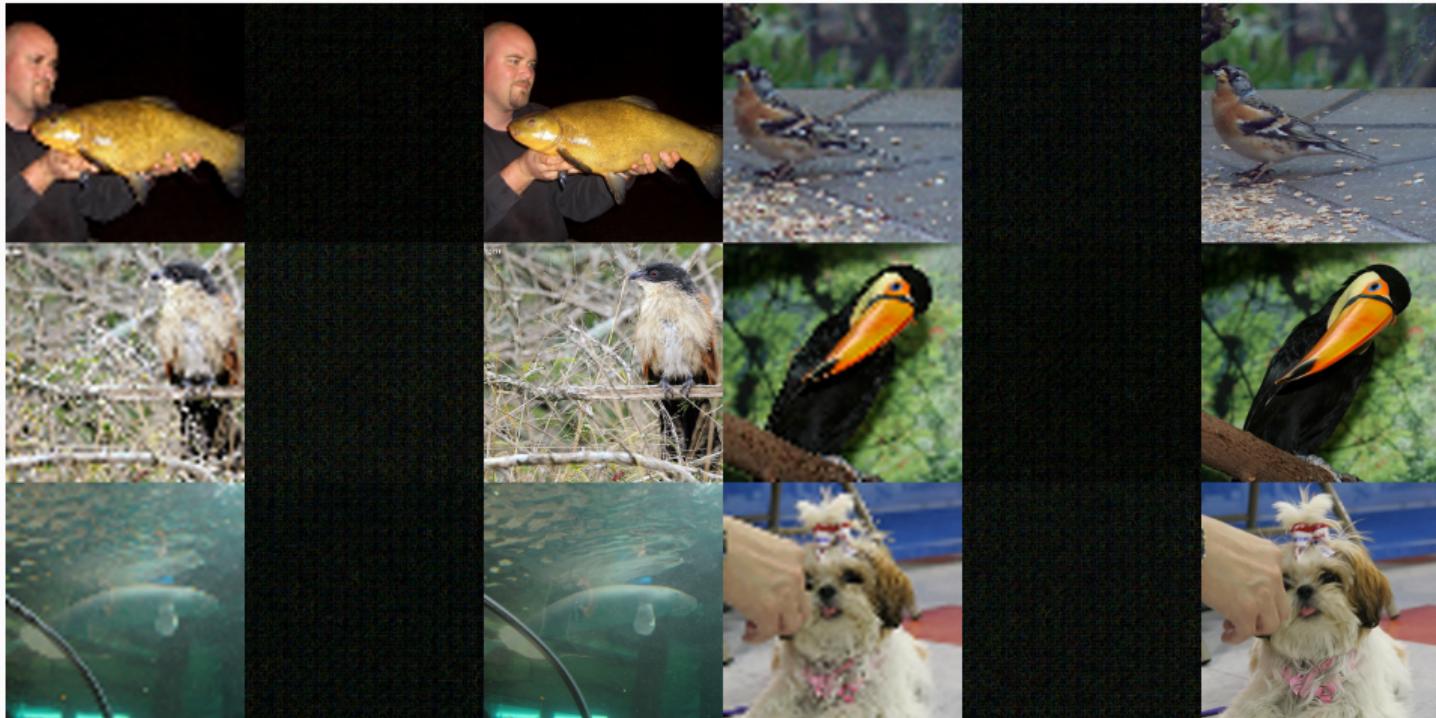


Figure 2.16 通道拼接方案效果图

# Token 拼接方案：性能指标

PSNR: 20.47

SSIM: 0.617

参数量: 131M

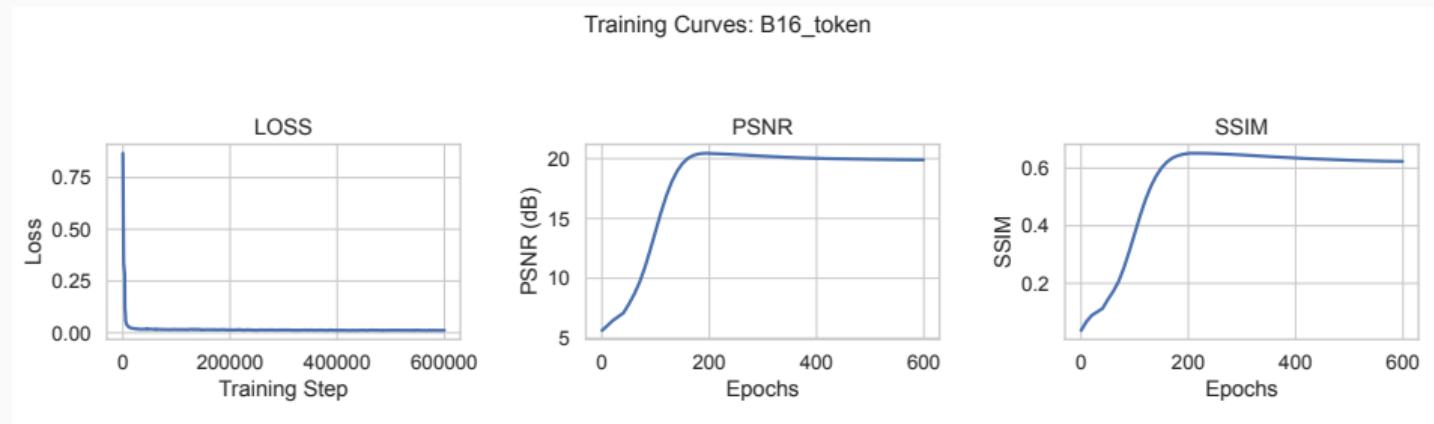


Figure 3.B16 Token 拼接方案在验证集上的性能曲线

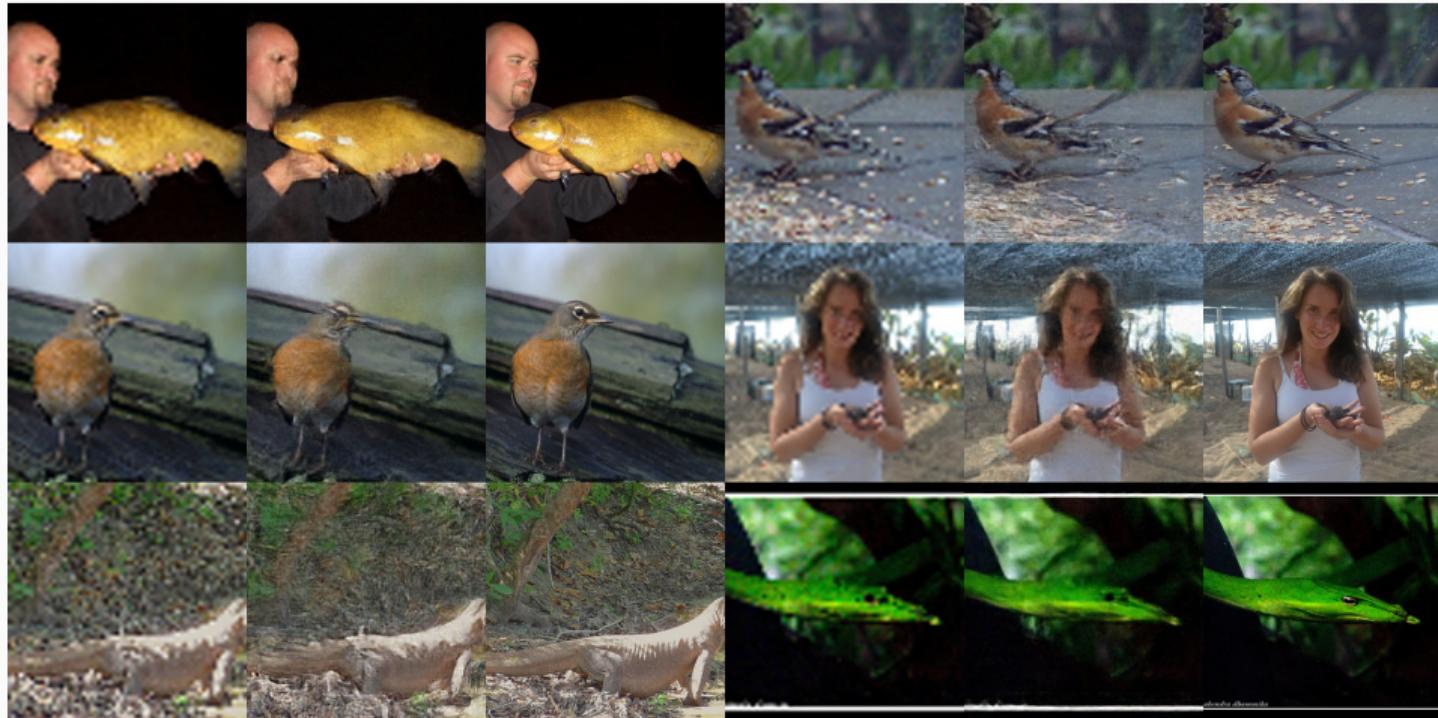


Figure 4.B16 token 拼接方案效果图

# Token 拼接方案：性能指标

PSNR: 21.10

SSIM: 0.634

参数量: 462M

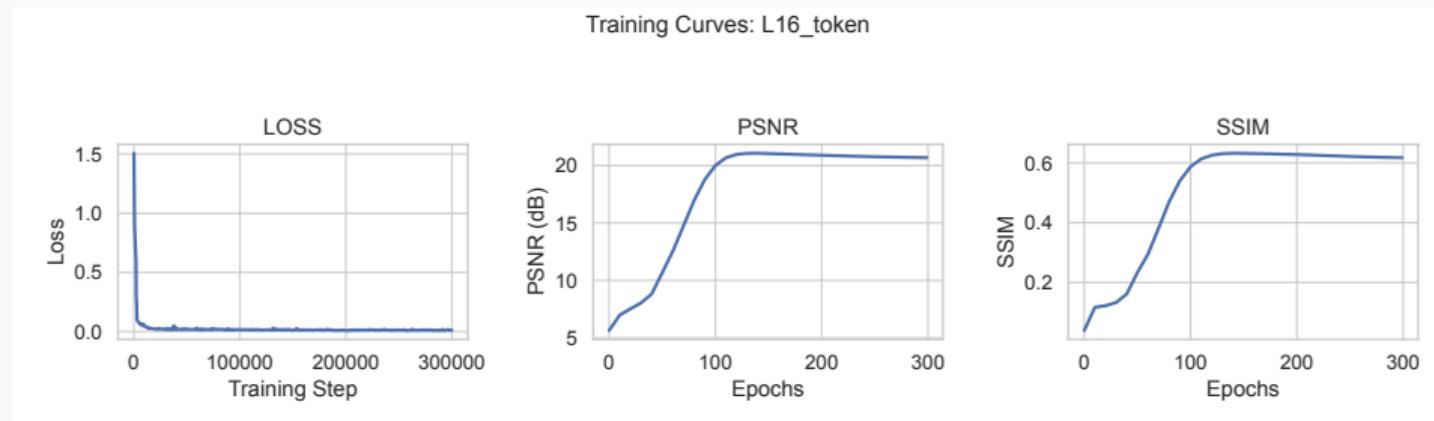


Figure 5. L16 Token 拼接方案在验证集上的性能曲线

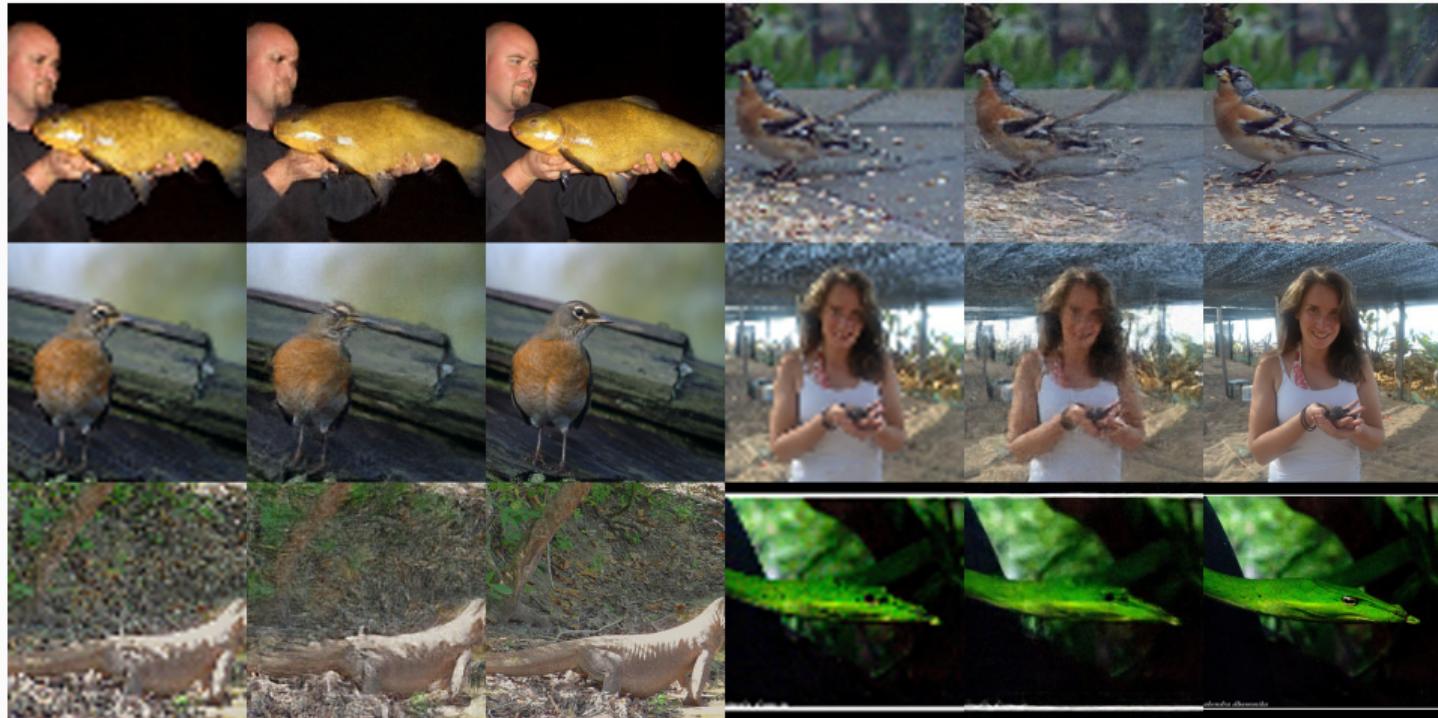


Figure 6.L16 token 拼接方案效果图

## 结果分析

---

# 性能对比

算法	psnr	ssim
bicubic	19.67	0.579
EDSR	22.92	0.702
SRGAN	21.87	0.692
SwinIR	23.77	0.733
本模型	20.47	0.617

Table 1.性能对比

## 整体性能总结

- 在 ImageNet  $\times 4$  超分辨率任务上，本方法在 PSNR 与 SSIM 指标上 **明显优于 bicubic 插值**。
- 与经典回归型超分模型（如 EDSR、SwinIR）相比，当前模型在 PSNR / SSIM 指标上 **仍存在差距**。
- 该结果表明：基于 Diffusion 的 JiT-SR 方法在像素级重建指标上 **尚未达到当前主流 SR 方法的性能水平**。

## 反思：连续型生成模型与重建任务的适配性

- 本工作采用的 Diffusion / JiT 属于 **连续型生成模型**，其核心目标是学习条件分布  $p(x_{\text{HR}} | x_{\text{LR}})$ ，而非最小化像素级误差。
- 对于超分辨这类 **重建型任务**，PSNR 与 SSIM 本质上衡量的是 **条件均值解的像素一致性**，与生成模型的优化目标并不完全一致。
- 在 ImageNet 等高复杂度数据集上，同一低清图像可能对应多种合理的高分辨结果，连续型模型更倾向于生成 **感知上合理但非像素最优**的样本。
- 因此，Diffusion 模型在感知质量与细节多样性上具有优势，但在 PSNR / SSIM 等像素级重建指标上 **不一定占优**，这一现象具有一定的范式必然性。

## 反思：低清图条件信息引入方式

简单的 token 拼接的条件信息引入形式, 会导致条件信息 token 和带噪声图 patch 的 token, 在 self attention 的 Q K V 中没有区别, 它们完全对等地 “互相参考”

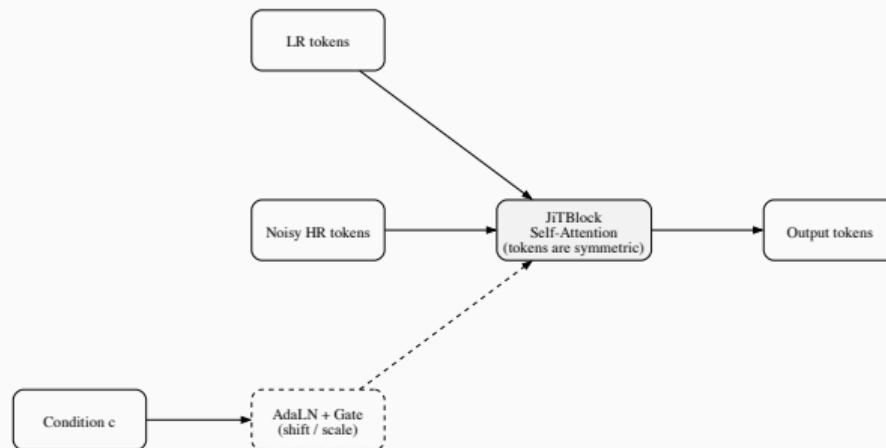


Figure 7.JitBlock self attention 的结构

谢谢大家！