视觉语言模型引导的图像复原

崔欣 杨[a]，荣康 董[a] 和Kin-ManLam[a]，*

[a]电机及电子工程学系，香港理工大学，中国香港

文章信息

关键词：图像复原 视觉语言模型 退化 语义理解

摘 要

许多图像复原（IR）任务既要求像素级保真度，也要求高层语义理解，才能复原出细节丰富、逼真的照片。然而，以往的方法往往难以有效利用视觉和语言知识。近期的一些工作尝试将擅长对齐视觉和文本特征的视觉语言模型（VLMs）引入通用图像复原。尽管如此，这些方法未能利用语言先验来确保复原过程中的语义连贯性。为解决这一问题，本文提出视觉语言模型引导的图像复原（VLMIR）框架，利用 CLIP 等视觉语言模型丰富的视觉语言先验，通过增强视觉感知和语义理解来提升图像复原性能。我们的方法分为两个阶段：基于视觉语言模型的特征提取和基于扩散的图像复原。第一阶段中，我们通过视觉语言模型凝练视觉感知和高层语义先验，提取输入图像互补的视觉和语言表征。具体而言，我们使用余弦相似度损失并结合 LoRA微调，对齐低质量图像与高质量图像的图像说明嵌入，同时利用退化预测器将退化嵌入与干净图像内容嵌入分解开来。随后，通过交叉注意力机制将这些互补的视觉和文本嵌入整合到基于扩散的模型中，以增强复原效果。大量实验和消融研究表明，VLMIR 在通用图像复原任务和特定退化图像复原任务中均取得了优异的性能，凸显了整合视觉语言模型中的视觉和语言知识对于提升图像复原能力至关重要。

arXiv:2512.17292v1 [cs.CV] 19 Dec 2025

1. 引言

图像复原（IR）是一项基础且广受欢迎的计算机视觉任务，旨在从低质量（LQ）图像中复原出高质量（HQ）图像 [14, 57]。它在监控、自动驾驶和多媒体娱乐等领域有着广泛的应用。去噪 [48], 超分辨率 [58], 和图像修复[41], 等图像复原任务旨在恢复精细细节，同时保留语义完整性。然而，大多数现有的图像复原方法 [63,24], 依赖多样的网络架构来利用低层图像统计特征 [15, 53] 和训练损失[11], , 难以兼顾像素级保真度与高层语义连贯性。

视觉语言模型（VLM） [61],的兴起为应对上述挑战带来了令人期待的机遇：利用多模态知识引导复原过程。CLIP [37]和 DALL-E [43],等视觉语言模型在大规模图文对数据集上进行预训练，能够在共享嵌入空间中对齐视觉特征与相应的文本描述。凭借在视觉和语言理解方面的先进能力，视觉语言模型已成功应用于多种任务 [1, 38], , 包括图文检索、目标检测和分割。然而，将视觉语言模型应用于图像复原任务仍面临若干挑战。一个主要问题是领域不匹配：视觉语言模型通常在配有描述性图像说明的干净、高质量图像上进行预训练，因此在处理退化的低质量图像时效果较差。近期进展 [32, 34, 55] 表明，将视觉语言模型融入图像复原能够进一步提升性能。这些方法通常会训练一个图像控制器来预测低质量图像的退化类型，从而分离干净图像内容与退化嵌入，再将二者整合到图像复原模型中。然而，这些方法往往忽视了语言先验的运用，例如从图像描述中提取的文本嵌入；这类先验能够在图像复原过程中提供高层语义理解和连贯性。另一种称为智能体图像复原 [6, 67, 68],的方法则将视觉语言模型用作智能体，以描述图像退化并制定复原计划，借助其先进的推理和规划能力。然而，智能体图像复原主要侧重于生成描述或指令，而没有利用视觉语言模型嵌入中编码的细粒度特征级先验。这限制了视觉语言模型的视觉语言知识在图像复原建模中的充分利用。

为充分利用嵌入视觉语言模型中的视觉语言知识，本文提出了视觉语言模型引导的图像复原（Vision-Language Model Guided Image Restoration，VLMIR）框架。受 DA-CLIP [32],启发，并使用 CLIP [42],，我们的方法同时学习图像内容嵌入和图像说明嵌入，这些嵌入构成互补的视觉语言信息，同时还学习退化嵌入。随后，利用这些嵌入引导下游图像复原过程。具体而言，VLMIR 包含两个

*通讯作者地址：香港理工大学电机及电子工程学系，中国香港。cuixin.yang@connect.polyu.hk (C. Yang); rongkang97.dong@connect.polyu.hk (R. Dong)； enkmlam@polyu.edu.hk (K. Lam) 开放研究者与贡献者身份识别码:

主要阶段：基于视觉语言模型的特征提取和基于扩散的图像复原。

在第一阶段，我们利用自举式视觉-语言框架 BLIP [27] 生成低质量图像的图像说明（称为低质量图像描述）及其对应高质量图像的图像说明（真实图像描述）。由干净的高质量图像生成的真实图像描述被假定为准确的描述。由于退化，低质量图像描述可能产生误导且不够准确，如图 1所示。因此，我们使用余弦相似度损失对齐低质量图像描述嵌入与真实图像描述嵌入，并通过 LoRA 微调低质量图像描述的文本编码器 [18]。此外，我们通过对比学习训练退化预测器以识别退化类型，从而使图像编码器能够生成干净图像嵌入 [32]。

在第二阶段，我们通过交叉注意力机制，将提取出的互补视觉和文本特征整合到基于扩散的图像复原模型中。为确保训练稳定，我们采用 IR-SDE [33] 作为骨干模型。

总而言之，本文的主要贡献如下：

- 我们提出了视觉语言模型引导的图像复原（VLMIR）框架，该框架利用视觉语言模型中的互补视觉先验和语言先验来引导图像复原过程。
- 我们提出了一种基于 CLIP 同时学习图像内容嵌入和图像描述嵌入的方法，通过 LoRA微调结合余弦相似度损失对齐低质量图像描述与真实图像描述，并采用对比学习获取干净的图像内容嵌入和退化嵌入。
- 大量实验和消融研究验证了所提方法的有效性，定量和定性评估均表明，该方法在通用图像复原和特定退化图像复原任务中都取得了优异性能。

2. 相关工作

2.1. 图像复原

图像复原（IR）是一项经典且重要的低层次计算机视觉任务。早期的 IR 方法主要依赖手工设计先验 [56]和数学模型 [45]来处理图像退化。中值滤波 [22] 和维纳滤波 [52]等去噪技术假设特定的噪声分布，但通常会导致过度平滑和细节丢失。全变分正则化[45] 和稀疏编码 [9] 等方法施加平滑性和稀疏性等约束来重建图像，但难以应对复杂的退化。

深度学习的兴起彻底改变了图像复原，通过引入数据驱动模型，从退化的低质量图像到干净的高质量图像学习复杂的映射关系。DnCNN [62]等模型用于去噪，SRResNet [25] 等模型用于超分辨率，采用深度卷积架构学习层次特征，从而能够高保真地复原精细细节，并显著提升整体图像质量。近年来，基于 Transformer 的架构（如 SwinIR [29], ）通过长程依赖建模捕获全局表征，展现出更强的性能。

新兴生成模型，包括生成对抗网络（GANs） [11] 和扩散模型 [44],，通过生成逼真的纹理和结构，进一步提升了感知质量，尤其是在图像修复和超分辨率等任务中。尽管这些方法凭借优化的训练、量身定制的架构和损失函数，在特定退化场景中表现出色，但由于存在分布外退化，它们往往难以在单个模型中泛化到各种图像复原任务 [26]。为应对复杂多样的现实应用而分别设计训练流程，会导致效率低下。

2.2. 通用图像复原

近年来，为解决特定任务方法的局限性，人们提出了通用图像复原模型，也称为一体化图像复原模型 [26, 36, 37, 21]。这些模-型 [51, 65, 60]旨在通过统一框架 [20], 处理多种彼此独立或混合的退化类型，并且通常会显式或隐式地对退化类型进行分类 [32,19]。早期的通用图像复原方法采用解耦学习[10], 、定制编码器 [28], 或专用解码头[7]等技术，应对多种退化挑战。近期的进展则转向更为集成且稳健的解决方案。例如，AirNet [26]提出通过正负样本对退化表征进行对比学习，并利用这些表征引导复原过程。PromptIR [37]引入了一种即插即用提示模块，可隐式预测退化类型，并将编码后的退化特征注入复原网络。InstructIPT [47]提出了一种简单而有效的权重调制方法，认为对于彼此高度无关的特定任务，权重调制比特征调制更为合适。近期，外部大规模模型 [42] 和生成先验 [17] 也被引入，以进一步提升性能 [32, 66, 34,55]。视觉语言模型（VLM），例如在网络规模数据集上训练的 CLIP [42],，擅长嵌入对齐的视觉与语言特征。InstructIR [8]利用 GPT-4 生成用户提示，并将这些指令用作控制机制，以复原高质量图像。DA-CLIP [32] 使用 CLIP [42]学习退化与图像内容的表征，并利用这些表征促进低质量图像的复原。为拓展 DA-CLIP 的应用范围，作者利用合成退化流程模拟野外退化，使其能够应用于野外图像复原场景 [34]。在 DA-CLIP 的基础上， [55] 的作者提出了一种通过伪真值生成来去除真实场景恶劣天气的方法，并采用 RainMix [13, 4] 数据增强方法，将合成雨纹掩码融合到退化图像中。

然而，这些方法对强大的视觉语言模型所提供的统一视觉和语言表征的利用仍然有限。它们通常只将视觉嵌入和退化嵌入整合到图像复原模型中，却忽略了由图像说明生成的互补语言嵌入。相比之下，本文提出一种视觉语言模型引导的图像复原框架，在整个复原过程中协同处理视觉感知与语言连贯性。

2.3. 视觉语言模型

视觉语言模型（VLMs）结合计算机视觉与自然语言处理，以处理和理解多模态数据，包括图像、视频和文本。视觉语言模型的核心在于，在共享的嵌入空间中对齐视觉感知表征与语言表征。通过同时从视觉和文本上下文中学习，视觉语言模型具备出色的零样本能力，并支持广泛的应用，例如图像描述生成、视觉问答和多模态推理。CLIP通过对比学习对齐图像与文本嵌入。BLIP [27]通过引导合成描述生成，利用噪声网络数据。[3] 在包含任意交错文本和图像的大规模图文数据集上进行训练。基于LLaMA[49], LLaVA [30] 将预训练语言模型与视觉编码器相结合，以相对较低的计算需求处理、理解并生成多模态输出。DALL-E [43] 和 Stable Diffusion [44] 基于生成模型，例如自回归框架 [12]和扩散模型 [17]，以根据文本提示生成图像。

近年来，视觉语言模型被引入通用图像复原领域。诸如 [32, 34, 55]利用 CLIP [42] 获取干净图像内容和退化嵌入，并将其作为提示来辅助图像复原过程。另一方面，智能体图像复原方法利用大语言模型/视觉语言模型来调度复原计划，而不是利用特征级先验。本文将有效利用视觉语言模型的特征级视觉与语言先验，以指导图像复原过程。

3. 方法

3.1. 概述

所提出方法的概览如图1所示。训练过程分为两个阶段：基于视觉语言模型的特征提取和图像复原。第一阶段的目标是通过基于视觉语言模型的特征提取，获得干净图像嵌入，并准确预测低质量图像的退化类型。如图1所示，具体做法是将低质量图像的嵌入 f_{LQ}^{I} 与真实图像描述的嵌入 f_{GT}^{T}对齐，并将预测退化f_{D}^{I} 与文本编码器生成的退化嵌入 f_{D}^{T}相匹配。为进一步增强模型的语义理解能力，我们将低质量图像描述的文本嵌入 f_{LQ}^{T}与真实图像描述的嵌入 F_{GT}^{T}对齐。需要注意的是，合成的真实图像描述和低质量图像描述均使用引导式视觉语言框架 BLIP [27]生成。随后，这些特征将用于训练下游图像复原模型，以提升其性能。

3.2. 基于视觉语言模型的特征提取

语言信息已被证明在许多计算机视觉任务中十分有效 [31, 59],，尽管其在图像复原中的应用仍未得到充分探索。为增强上下文理解并生成更高质量的图像，我们提出在退化图像的复原过程中同时引入视觉和语言表征。我们采用预训练的 CLIP [37]模型作为特征提取的骨干网络。CLIP 是一种多模态视觉语言模型，通过对比学习同时学习图像和文本，并在二者之间建立联系。从 CLIP 的图像编码器和文本编码器中提取的潜在特征，旨在为图文对提供良好对齐的表征。如图 1所示，退化的文本描述和真实图像描述被输入文本编码器。通过预测退化与文本编码器生成的退化嵌入之间的对比学习，训练退化预测器对低质量图像的退化类型进行分类 f_{D}^{I} 。这使图像编码器能够生成由真实图像描述嵌入 f_{D}^{T}引导的干净图像嵌入 f_{LQ}^{I} 。对比损失的公式如下：

$$L_c(m,n) = -\frac{1}{K}\sum_i \log\left(\frac{e^{m_i^T n_i/s}}{\sum_j e^{m_i^T n_j/s}}\right), \tag{1}$$

其中， m和 n表示图像-文本嵌入对的归一化向量批次，即 f_{LQ}^{I}& f_{GT}^{T}和 f_{D}^{I}& f_{D}^{T}，其中 m_i和 n_i是正样本对的嵌入，而n_j 表示所有其他样本（包括负样本）的嵌入； K 是训练批次中的配对特征数量； s是一个用于调整对比强度的可学习因子。

语义描述已在各种计算机视觉任务中展现出有效性 [31, 59]。为了丰富下游图像复原模型可用的文本信息，我们引入了一个额外的文本编码器，用于编码低质量图像的图像说明（低质量图像说明）。由于退化图像的视觉质量较低且存在失真，生成的低质量图像说明可能无法准确描述图像内容，从而引入噪声（见图 1）。为修正低质量图像说明的潜在嵌入，即 f_{LQ}^{T}，我们将其与对应真实图像说明的潜在嵌入，即 f_{GT}^{T}进行对齐。具体而言，低质量图像说明文本编码器的初始参数从预训练的 CLIP 文本编码器复制而来。随后，我们使用 LoRA 对其进行微调，采用 f_{LQ}^{T}与 f_{GT}^{T}之间的余弦相似度损失，定义如下：

$$L_{cos}(m,n) = 1 - \cos(m,n), \tag{2}$$

其中，$\cos(\cdot,\cdot)$ 表示两个归一化向量 m和 n之间的余弦相似度。综上，第一阶段的总训练

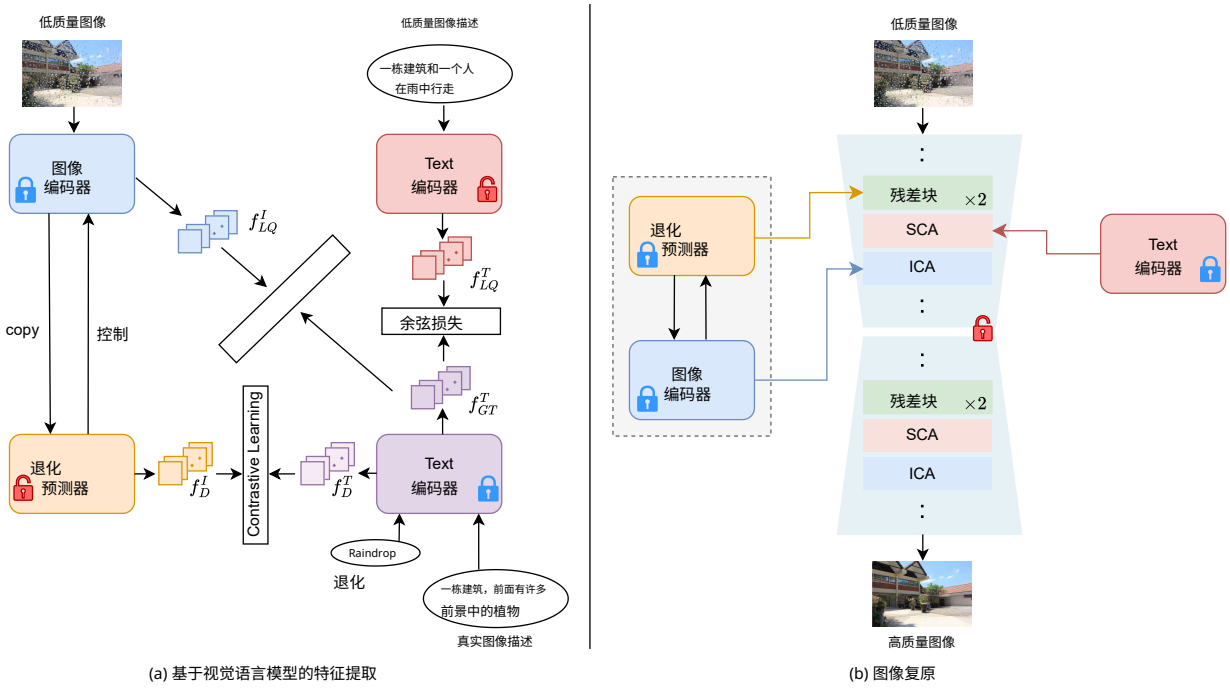

图 1: 所提出的视觉语言模型引导的图像复原（VLMIR）框架概览 .

损失表示如下:

$$L_1 = L_c(f_{LQ}^I, f_{GT}^T) + L_c(f_D^I, f_D^T) + \alpha \times L_{cos}(f_{LQ}^T, f_{GT}^T), \quad (3)$$

其中， α 是余弦相似度损失的加权因子。该框架充分利用了视觉和文本特征，可将其集成到下游图像复原模型中，以提升模型性能。

3.3. 图像复原

在图像复原阶段，我们采用 IR-SDE [33],作为骨干网络。IR-SDE 是一种基于 U-Net 的扩散模型。骨干模型的所有参数均从头开始训练，而图像编码器、文本编码器和退化预测器则保持固定。为了同时利用低质量图像描述的文本嵌入和干净图像嵌入，我们在扩散模型中引入了两种机制：语义交叉注意力（SCA）和图像交叉注意力（ICA）。ICA 模块位于 SCA 模块之后。通过以文本嵌入和图像嵌入为条件，扩散模型能够增强对图像的语义理解，从而提升复原性能。此外，退化预测器生成的退化嵌入用于判别退化后的低质量图像的具体退化类型。这使模型能够通过单一模型处理各种类型的退化。具体而言，退化嵌入被注入残差块，并通过可学习提示进行融合，以生成用于变换输入特征的缩放因子和平移因子 [32]。给定时刻 x_t 的噪声状态 t以及作为条件的低质量图像 μ，扩散模型表示为 $\epsilon_\theta(x_t, \mu, t, f_{LQ}^T, f_{LQ}^I, f_D^T)$。

扩散模型采用似然学习从头开始训练，并遵循 IR-SDE[33] 的稳定训练策略。

4. 实验

4.1. 实验设置

4.1.1. 数据集

我们在两种图像复原训练设置下开展实验：通用图像复原和特定退化图像复原。特定退化图像复原模型在仅包含一种特定退化类型图像的数据集上进行训练。通用图像复原模型也称为一体化图像复原模型，在包含多种退化类型的混合数据集上进行训练，旨在通过统一框架处理各种图像复原任务。在本研究中，我们在三项图像复原任务上评估模型性能：雨滴去除、去雾和去噪。对于雨滴去除，我们采用 RainDrop 数据集 [39],，其中包含 861 张训练图像和 58 张测试图像。对于去雾，我们采用 RESIDE-6k 数据集[40]。由于其规模较大，可能导致数据不平衡，因此我们选取前 1,000 张训练图像和前 50 张测试图像，分别构成训练数据集和测试数据集。对于去噪，我们采用广泛使用的 DIV2K 数据集 [2], 作为训练数据集，其中包含 900 张图像；并采用 CBSD68 [35], 作为测试数据集，其中包含 68 张图像。低质量噪声图像通过添加噪声水平为 50 的高斯噪声合成。真实图像描述和低质量图像描述使用视觉语言模型

表1不同通用图像复原方法在各类图像复原任务上的定量比较。最佳结果以红色突出显示，第二佳结果以蓝色标示。

方法	退化	PSNR	SSIM	Y通道峰值信噪比	Y通道结构相似性指数	学习感知图像块相似度	FID
AirNet [26]	Raindrop	22.09	0.8218	23.66	0.8732	0.0757	119.50
PromptIR [37]		28.01	0.8771	29.83	0.9134	0.0429	72.33
IR-SDE [33]		27.89	0.8707	29.57	0.9039	0.0633	32.37
DA-CLIP [32]		29.09	0.8856	30.89	0.9161	0.0568	21.72
VLMIR		29.36	0.8866	31.14	0.9175	0.0553	25.99
AirNet [26]	去雾	27.45	0.9475	28.96	0.9697	0.0150	29.59
PromptIR [37]		27.58	0.9479	29.13	0.9699	0.0162	26.25
IR-SDE [33]		18.67	0.8411	20.11	0.8933	0.0928	19.80
DA-CLIP [32]		30.01	0.9488	31.62	0.9742	0.0199	7.45
VLMIR		30.22	0.9523	31.85	0.9745	0.0199	8.99
AirNet [26]	去噪	28.00	0.7975	29.67	0.8214	0.0747	86.85
PromptIR [37]		27.06	0.7517	28.88	0.7884	0.0965	89.19
IR-SDE [33]		25.54	0.7057	27.18	0.7427	0.1833	77.03
DA-CLIP [32]		25.71	0.7157	27.28	0.7493	0.1766	74.09
VLMIR		25.74	0.7171	27.32	0.7507	0.1793	71.82

BLIP [27]。鉴于真实图像具有较高质量，我们认为其真实图像描述是准确的，并将其用作第一阶段模型训练期间的监督信号。

4.1.2. 评估指标

在评估中，我们采用失真度指标和感知指标。失真度指标包括峰值信噪比、结构相似性、Y通道峰值信噪比和Y通道结构相似性指数 [54], ，感知指标则包括学习感知图像块相似度 [64] 和弗雷歇起始距离 [16]。峰值信噪比和结构相似性用于评估图像中RGB通道的质量。Y通道峰值信噪比和Y通道结构相似性指数用于评估YCbCr色彩空间中亮度（Y）通道的质量。峰值信噪比、结构相似性、Y通道峰值信噪比和Y通道结构相似性指数的值越高，性能越好；而学习感知图像块相似度和弗雷歇起始距离的值越低，感知质量越好。

4.1.3. 实验配置

为训练模型的第一阶段，我们将初始学习率设为 0.00003，并在训练过程中采用余弦退火调度进行调整。模型训练 300 个训练轮次，批量大小为 256。训练时采用权重为 0.5 的余弦相似度损失进行引导。模型使用四块 NVIDIA GeForce RTX 4090 GPU 进行训练。我们选择在 LAION-2B 上训练的 CLIP ViT-B/32 模型，即 LAION-5B 的一个包含 20 亿个样本的英语子集 [46], ，作为预训练CLIP模型。训练第二阶段图像复原模型时，初始学习率设为 0.0001，并使用余弦退火调度器动态调整，$\beta_1 = 0.9$和 $\beta_2 = 0.99$。模型总共训练 150,000 次迭代，批量大小为 48。图块大小为 256×256。为增强泛化能力，我们采用了随机垂直和水平翻转等数据增强技术。训练使用六块 NVIDIA GeForce RTX 4090 GPU 进行。两个训练阶段均使用 Adam优化器 [23] 。

4.2. 与最先进方法的比较*4.2.1.* 通用图像复原

我们将提出的方法与多种先进的通用图像复原方法进行了比较，包括 AirNet [26],PromptIR [37], IR-SDE [33], 和 DA-CLIP[32]。请注意，AirNet [26]是一种基于卷积神经网络 (CNN) 的模型，而 PromptIR [37] 基于 Transformer 架构构建。AirNet 和 PromptIR 均使用 $l1$ 损失进行重建，以确保像素级保真度。IR-SDE [33], 、DA-CLIP [32],和我们的方法均基于生成式扩散模型。

表1 展示了这些方法在三项图像复原任务（即去除雨滴、去雾和去噪）上的定量性能。对于雨滴去除和去雾，我们的方法在所有失真度指标（即峰值信噪比、结构相似性、Y通道峰值信噪比和Y通道结构相似性指数）上均取得最佳性能，在学习感知图像块相似度和弗雷歇起始距离上也表现出竞争力。对于去噪，尽管基于CNN的AirNet [26] 和基于Transformer的PromptIR [37] 在失真度指标和学习感知图像块相似度上表现良好，但如图2所示，它们的输出往往过于平滑或视觉效果不佳，因此其弗雷歇起始距离得分不如基于生成模型的方法。

图 2 展示了通用图像复原方法在雨滴去除、去雾和去噪任务上的定性比较。AirNet [26] 和 PromptIR [37]难以去除雨滴。IR-SDE [33] 和 DA-CLIP[32]能够有效去除大多数雨滴，但仍会留下痕迹，也难以处理相机镜头上的雨滴造成的橙色伪影。相比之下，本文提出的方法能够去除所有雨滴和橙色伪影，复原出最接近目标高质量图像的干净图像。在去雾方面，本文提出的方法能够更有效地去除雾霾，生成视觉效果更佳的结果。在去噪方面，如上所述，由于 $l1$ 损失约束，AirNet 和 PromptIR 会生成过度平滑的图像，因此峰值信噪比和结构相似性更高。我们的模型在.....的引导下

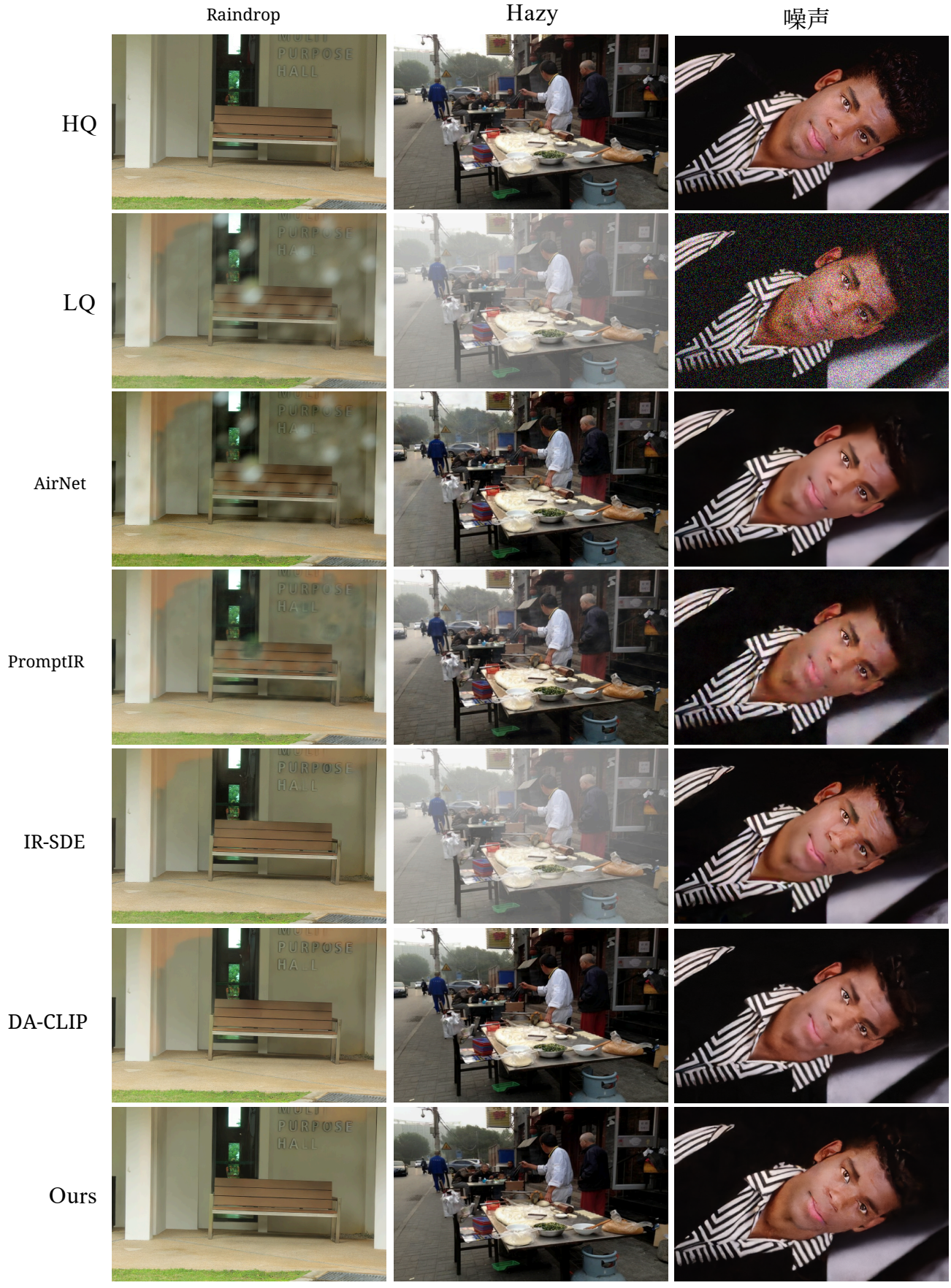

图 2： 不同通用图像复原方法在各种图像复原任务中的视觉效果。

表 2不同去雾方法在图像去雾任务上的定量比较。最佳结果以红色突出显示，次佳结果以蓝色标记。

方法	PSNR	SSIM	Y通道峰值信噪比	Y通道结构相似性指数	学习感知图像块相似度	FID
MAXIM [50]	31.05	0.9653	32.69	0.9814	0.0215	9.34
IR-SDE [33]	20.56	0.8671	22.03	0.9152	0.0718	15.93
DA-CLIP [32]	30.18	0.9513	31.78	0.9738	0.0189	8.40
Ours	30.34	0.9579	31.92	0.9757	0.0182	8.11

表 3图像去雾任务中各种变体的比较。最佳结果以粗体标出。

变体	PSNR	SSIM	Y通道峰值信噪比	Y通道结构相似性指数	学习感知图像块相似度	FID
SCA	29.51	0.9497	31.09	0.9741	0.0207	9.11
ICA	29.67	0.9499	31.24	0.9738	0.0206	**8.58**
SCA和ICA	**30.22**	**0.9523**	**31.85**	**0.9745**	**0.0199**	8.99

来自CLIP的视觉和文本嵌入，并由生成模型驱动，生成了细节更丰富、更逼真的图像。

4.2.2. 特定退化场景下的图像复原 *ion*

为了进一步评估所提出方法的有效性，我们不仅在通用图像复原设置下评估模型，还在特定退化场景下进行评估。具体而言，我们在去雾任务上开展实验。我们将所提出的方法与当前最先进的去雾方法进行比较，包括 MAXIM [50], IR-SDE[33],和DA-CLIP [32]。与通用图像复原不同，特定退化方法仅使用去雾数据集进行训练。

表2展示了各种去雾方法的定量比较结果。MAXIM [50],是一种基于Transformer的方法，使用Charbonnier损失 [5], ，该损失类似于$l1$ 损失。尽管MAXIM在基于失真的评估指标上取得了较高的结果，但由于 $l1$ 或 $l2$ 损失函数的局限性，其难以复原逼真且自然的图像。所提出的方法在峰值信噪比和结构相似性指标上优于其他基于生成模型的方法，同时在学习感知图像块相似度和弗雷歇起始距离方面也实现了更优的感知质量。这些结果证明了所提出方法的有效性和优越性。

图 3 展示了不同去雾方法的视觉效果。与其他方法相比，MAXIM [50] 去除雾霾的效果最差。IR-SDE [33] 也难以去除雾霾，尤其是树木周围。DA-CLIP [32]生成的视觉效果比 IR-SDE 更清晰，但建筑物上仍残留一些雾霾。在所有方法中，本文提出的方法生成的输出最清晰，细节也最丰富。

总之，通过结合视觉感知与语言理解，所提出的方法能够有效利用 VLM 提供的指导，复原受各种单一退化影响的图像。

表 4将不同文本描述作为输入提供给文本编码器的比较。“空值”表示未向文本编码器提供文本。最佳结果以粗体突出显示。

Text	PSNR	SSIM	学习感知图像块相似度	FID
“这是一张照片”	29.90	0.9541	**0.0190**	**8.572**
Null	29.55	**0.9554**	0.0198	8.86
低质量图像描述	**30.22**	0.9513	0.0199	8.99

4.3. 消融研究

4.3.1. 模块的影响

SCA 和 ICA 模块对于处理文本嵌入和视觉嵌入至关重要。为评估它们的影响，我们在图像去雾任务上进行了消融实验。表3 展示了不同模型变体的定量比较。“SCA”指仅包含 SCA 模块的变体，“ICA”仅包含 ICA 模块，而“SCA&ICA”则同时包含 SCA 和 ICA 模块，即我们提出的模型。“ICA”变体在所有基于失真的指标和感知指标上均优于“SCA”，这表明在图像复原中，图像内容比文本描述更为重要。然而，同时包含 SCA 和 ICA 的完整模型在大多数评估指标上优于其他所有变体，并取得了相当的弗雷歇起始距离（FID）分数。这凸显了视觉嵌入和文本嵌入对于生成逼真且细节丰富的复原图像的重要贡献。

4.3.2. 不同低质量图像描述的影响

我们还研究了不同文本输入对文本编码器的影响。低质量图像描述是指对低质量图像内容的文本描述。作为替代方案，我们使用“This is aphoto”和空输入（即不输入文本）。表4 显示，不向文本编码器输入文本描述会导致像素级保真度显著下降。与空输入相比，通用描述“This is a photo”能够提升图像保真度，但效果仍不及使用实际的低质量图像描述。使用低质量图像描述时，学习感知图像块相似度和弗雷歇起始距离等感知指标略有下降，这可能是由于对低质量图像描述不准确所引入的噪声。尽管如此，低质量图像描述使模型能够生成保真度更高、感知质量相当的图像，凸显了其在图像复原过程中的重要性。

5. 结论

本文介绍了视觉语言模型引导的图像复原（VLMIR）框架。该框架利用视觉语言模型（VLM）的视觉和语言先验互补优势，推进通用图像复原。VLMIR采用两阶段方法：首先进行基于视觉语言模型的特征提取，随后进行基于扩散的图像复原。在第一阶段，我们通过LoRA微调，使用余弦相似度损失对齐低质量（LQ）和高质量（HQ）图像描述嵌入，并通过对比学习提取干净图像内容和

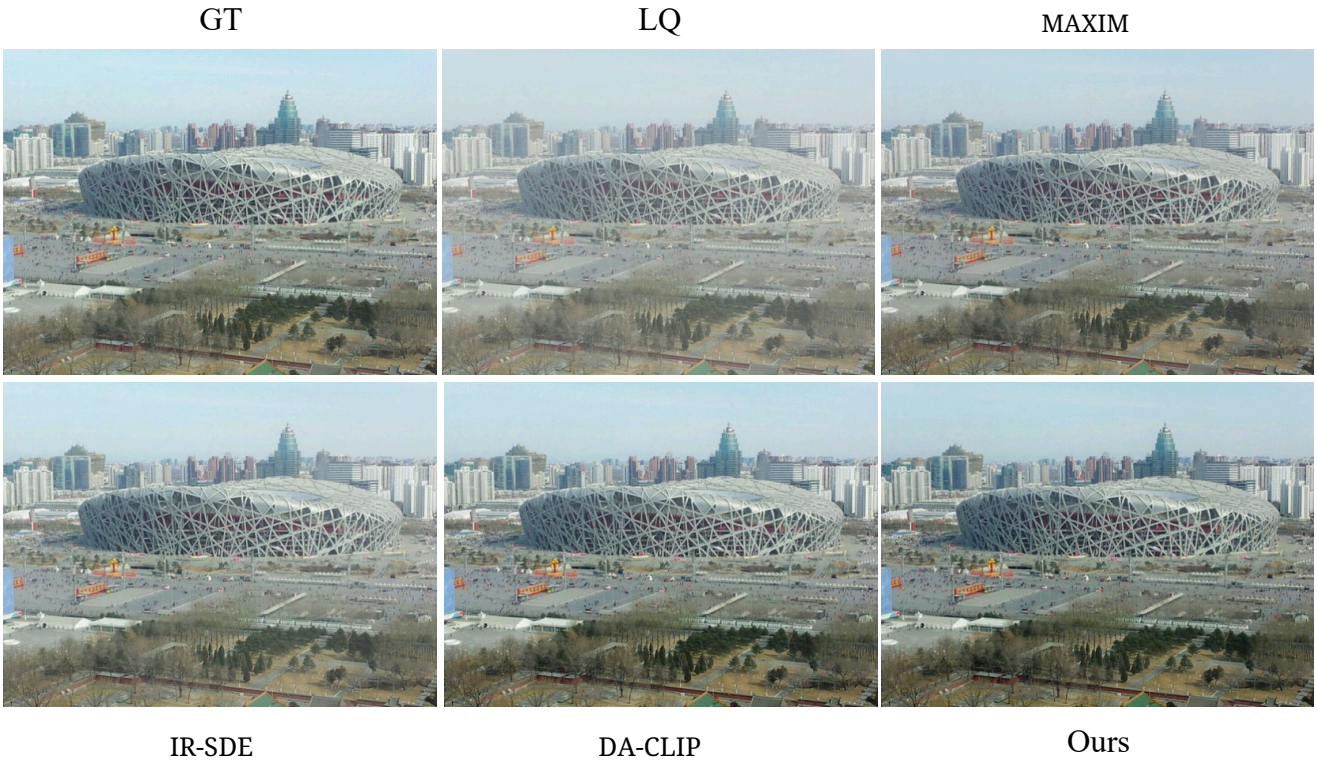

图 3： 不同去雾方法在图像去雾任务上的视觉效果。

退化特定嵌入。随后，这些特征通过交叉注意力机制无缝集成到基于扩散的模型（IR-SDE）中。大量实验和消融研究表明，VLMIR在通用复原任务和特定退化复原任务中均优于现有方法。我们的研究结果凸显了整合视觉语言模型中的视觉和语言知识的重要作用，为实现更具泛化能力和更有效的图像复原方案铺平了道路。

CRediT 作者贡献声明

Cuixin Yang： 撰写——初稿、概念化、方法、软件、可视化、验证、数据整理。 **Rongkang Dong：** 撰写——方法、验证。 **Kin-Man Lam：** 撰写——审阅与编辑、监督。

参考文献

[1] Abdelfattah, R., Guo, Q., Li, X., Wang, X., Wang, S., 2023年。Cdul：用于多标签图像分类的 CLIP驱动的无监督学习，载于：IEEE/CVF国际计算机视觉会议论文集，第1348–1357页。[2] Agustsson, E., Timofte, R., 2017。NTIRE 2017挑战赛：单幅图像超分辨率数据集与研究，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。[3] Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., 等，2022。Flamingo：一种用于少样本学习的视觉语言模型。《神经信息处理系统进展》35，23716–23736。[4] Ba, Y., Zhang, H., Yang, E., Suzuki, A., Pfahnl, A., Chandrappa, C.C., De Melo, C.M., You, S., Soatto, S., Wong, A., 等，2022年。Not Just Streaks：迈向单幅图像去雨的真实值，载于：欧洲计算机视觉会议，施普林格出版社。第723–740页。[5] Barron, J.T., 2019年。一种通用且自适应的鲁棒损失函数，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第4331–4339页。[6] Chen, H., Li, W., Gu, J., Ren, J., Chen, S., Ye, T., Pei, R., Zhou, K., Song, F., Zhu, L., 2024年。RestoreAgent：基于多模态大语言模型的自主图像复原智能体。《神经信息处理系统进展》37，110643–110666。[7] Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W., 2021年。预训练图像处理变换器，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第12299–12310页。[8] Conde, M.V., Geigle, G., Timofte, R., 2024年。InstructIR：遵循人类指令的高质量图像复原，载于：欧洲计算机视觉会议，施普林格出版社。第1–21页。[9] Elad, M., Aharon, M., 2006年。通过学习字典上的稀疏和冗余表示进行图像去噪。《IEEE图像处理汇刊》15，3736–3745。[10] Fan, Q., Chen, D., Yuan, L., Hua, G., Yu, N., Chen, B., 2019年。一种用于参数化图像算子的通用解耦学习框架。《IEEE模式分析与机器智能汇刊》43，33–47。[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020年。生成对抗网络。《美国计算机协会通讯》63，139–144。[12] Gregor, K., Danihelka, I., Mnih, A., Blundell, C., Wierstra, D., 2014年。深度自回归网络，载于：国际机器学习会议，PMLR。第1242–1250页。

[13] Guo, Q., Sun, J., Juefei-Xu, F., Ma, L., Xie, X., Feng, W., Liu, Y., Zhao, J., 2021年。Efficientderain：通过学习逐像素膨胀滤波实现高效单幅图像去雨，载于：AAAI人工智能会议论文集，第1487–1495页。[14] Guo, Y., Dai, L., Gan, X., Huang, Y., Ruan, M., Huang, D., 2025年。通过退化去除和文本提示实现真实场景图像超分辨率的一步扩散。《图像与视觉计算》，105699。[15] He, K., Zhang, X., Ren, S., Sun, J., 2016年。深度残差网络中的恒等映射，载于：欧洲计算机视觉会议，施普林格出版社。第630–645页。[16] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017年。采用双时间尺度更新规则训练的生成对抗网络收敛于局部纳什均衡。《神经信息处理系统进展》30。[17] Ho, J., Jain, A., Abbeel, P., 2020年。去噪扩散概率模型。《神经信息处理系统进展》33，6840 6851–。[18] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 等，2022年。LoRA：大语言模型的低秩适配。国际学习表征会议 1，3。[19] Hu, J., Jin, L., Yao, Z., Lu, Y., 2025年。通过退化分类进行通用图像复原预训练。arXiv预印本 arXiv:2501.15510。[20] Jiang, J., Zuo, Z., Wu, G., Jiang, K., Liu, X., 2024a。一体化图像复原综述：分类体系、评估与未来趋势。arXiv预印本 arXiv:2410.15067。[21] Jiang, Y., Zhang, Z., Xue, T., Gu, J., 2024b。AutoDIR：基于潜在扩散的自动一体化图像复原，载于：欧洲计算机视觉会议，施普林格出版社。第340–359页。[22] Justusson, B., 2006年。中值滤波：统计特性，载于：《二维数字信号处理II：变换与中值滤波器》。施普林格出版社，第161–196页。[23] Kingma, D.P., Ba, J., 2014年。Adam：一种随机优化方法。arXiv预印本 arXiv:1412.6980。[24] Kwon, M.J., Nam, S.H., Yu, I.J., Lee, H.K., Kim, C., 2022年。学习JPEG压缩伪影以进行图像篡改检测与定位。《国际计算机视觉期刊》130，1875–1895。[25] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., 等，2017年。利用生成对抗网络实现照片级真实感单幅图像超分辨率，载于：IEEE计算机视觉与模式识别会议论文集，第4681–4690页。[26] Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X., 2022a。针对未知退化的一体化图像复原，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第17452–17462页。[27] Li, J., Li, D., Xiong, C., Hoi, S., 2022b。BLIP：通过自举实现语言-图像预训练，以统一视觉语言理解与生成，载于：国际机器学习会议，PMLR。第12888–12900页。[28] Li, R., Tan, R.T., Cheong, L.F., 2020年。使用架构搜索实现一体化恶劣天气去除，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第3175–3185页。[29] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021年。SwinIR：使用Swin Transformer进行图像复原，载于：IEEE/CVF国际计算机视觉会议论文集，第1833–1844页。[30] Liu, H., Li, C., Wu, Q., Lee, Y.J., 2023a。视觉指令微调。《神经信息处理系统进展》36，34892–34916。[31] Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R.K., Mahadevan, V., Manmatha, R., 2023b。Polyformer：将指代表达图像分割视为序列多边形生成，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第18653–18663页。[32] Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B., 2023a。控制视觉语言模型以实现多任务图像复原。

arXiv 预印本 arXiv:2310 .01018 .

[33] Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B., 2023b. 基于均值回复随机微分方程的图像复原。arXiv预印本，arXiv:2301.11699。[34] Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B., 2024. 使用受控视觉语言模型在自然场景中实现照片级真实感图像复原，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第6641–6651页。[35] Martin, D., Fowlkes, C., Tal, D., Malik, J., 2001. 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，载于：第八届IEEE国际计算机视觉会议。ICCV 2001，IEEE，第416–423页。[36] Park, D., Lee, B.H., Chun, S.Y., 2023. 使用针对特定退化的自适应判别滤波器实现未知退化的一体化图像复原，载于：2023年IEEE/CVF计算机视觉与模式识别会议（CVPR），IEEE，第5815–5824页。[37] Potlapalli, V., Zamir, S.W., Khan, S.H., Shahbaz Khan, F., 2023. PromptIR：面向一体化图像复原的提示。神经信息处理系统进展36，71275–71293。[38] Qian, Q., Hu, J., 2024. 使用CLIP进行在线零样本分类，载于：欧洲计算机视觉会议，施普林格，第462–477页。[39] Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J., 2018. 用于从单幅图像中去除雨滴的注意力生成对抗网络，载于：IEEE计算机视觉与模式识别会议论文集，第2482–2491页。[40] Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H., 2020. FFA-Net：用于单幅图像去雾的特征融合注意力网络，载于：AAAI人工智能会议论文集，第11908–11915页。[41] Quan, W., Chen, J., Liu, Y., Yan, D.M., Wonka, P., 2024. 基于深度学习的图像与视频修复：综述。《国际计算机视觉期刊》132，2367–2400。[42] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等，2021. 从自然语言监督中学习可迁移视觉模型，载于：国际机器学习会议，PMLR，第8748–8763页。[43] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I., 2021. 零样本文本到图像生成，载于：国际机器学习会议，PMLR，第8821–8831页。[44] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. 基于潜扩散模型的高分辨率图像合成，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第10684–10695页。[45] Rudin, L.I., Osher, S., Fatemi, E., 1992年. 基于非线性全变分的去噪算法。Physica D：非线性现象60，259–268。[46] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., 等，2022. LAION-5B：用于训练下一代图文模型的开放大规模数据集。神经信息处理系统进展35，25278–25294。[47] Tian, Y., Han, J., Chen, H., Xi, Y., Ding, N., Hu, J., Xu, C., Wang, Y., 2024. Instruct-IPT：通过权重调制实现的一体化图像处理变换器。arXiv预印本，arXiv:2407.00676。[48] Tirer, T., Giryes, R., 2018年. 通过迭代去噪和反向投影进行图像复原。《 IEEE 图像处理汇刊》28，1220–1234。[49] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., 等，2023年. LLaMA：开放且高效的基础语言模型。arXiv预印本，arXiv:2302.13971。[50] Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y., 2022. MAXIM：用于图像处理的多轴MLP，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第5769–5780页。[51] Valanarasu, J. M.J., Yasarla, R., Patel, V.M., 2022. TransWeather：基于变换器的恶劣天气条件退化图像复原，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第

视觉与模式识别，第2353–2363页。[52] Vaseghi, S.V., 1996. 维纳滤波器，载于：《高级信号处理与数字降噪》。施普林格，第140–163页。[53] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. 注意力就是你所需要的一切。神经信息处理系统进展30。[54] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. 图像质量评估：从误差可见性到结构相似性。《IEEE图像处理汇刊》13，600–612。[55] Wei, Y., Zhang, Y., Li, K., Wang, F., Tang, S., Zhang, Z., 2025. 利用视觉语言提示实现真实场景图像复原与增强。《计算机视觉与图像理解》250，104222。[56] Weiss, Y., Freeman, W.T., 2007. 什么样的模型才是自然图像的良好模型？，载于：2007年IEEE计算机视觉与模式识别会议，IEEE，第1–8页。[57] Yang, C., Dong, R., Xiao, J., Zhang, C., Lam, K.M., Zhou, F., Qiu, G., 2025. 几何失真引导的全向图像超分辨率变换器。《IEEE视频技术电路与系统汇刊》。[58] Yang, C., Xiao, J., Ju, Y., Qiu, G., Lam, K.M., 2023a. 使用蒙特卡罗方法提升单幅图像超分辨率模型的鲁棒性，载于：2023年IEEE国际图像处理会议（ICIP），IEEE，第2135–2139页。[59] Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., 等，2023b. Reco：区域控制的文本到图像生成，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第14246–14255页。[60] Zeng, H., Li, J., Zheng, Z., Xiong, Z., 2025. 一体化图像压缩与复原，载于：2025年IEEE/CVF冬季计算机视觉应用会议（WACV），IEEE，第609–619页。[61] Zhang, J., Huang, J., Jin, S., Lu, S., 2024. 面向视觉任务的视觉语言模型：综述。《IEEE模式分析与机器智能汇刊》46，5625–5644。[62] Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L., 2017. 超越高斯去噪器：用于图像去噪的深度CNN残差学习。《IEEE图像处理汇刊》26，3142–3155。[63] Zhang, K., Zuo, W., Zhang, L., 2018a. FFDNet：迈向快速灵活的基于CNN的图像去噪方案。《IEEE图像处理汇刊》27，4608–4622。[64] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018b. 深度特征作为感知度量的非凡有效性，载于：IEEE计算机视觉与模式识别会议论文集，第586–595页。[65] Zhang, X., Ma, J., Wang, G., Zhang, Q., Zhang, H., Zhang, L., 2025. Perceive-IR：通过更好地感知退化实现一体化图像复原。《IEEE图像处理汇刊》。[66] Zheng, D., Wu, X.M., Yang, S., Zhang, J., Hu, J.F., Zheng, W.S., 2024. 基于扩散模型的通用图像复原选择性沙漏映射，载于：IEEE/CVF计算机视觉与模式识别会议论文集，第25445–25455页。[67] Zhu, K., Gu, J., You, Z., Qiao, Y., Dong, C., 2024. 面向复杂图像复原问题的智能代理系统。arXiv预印本，arXiv:2410.17809。[68] Zuo, Y., Zheng, Q., Wu, M., Jiang, X., Li, R., Wang, J., Zhang, Y., Mai, G., Wang, L.V., Zou, J., 等，2025. 4KAgent：智能代理式任意图像到4K超分辨率。arXiv预印本，arXiv:2507.07105。