

DL-Diff: 基于暗转亮扩散模型的低光照视频增强方法

王雨轩, 邢桢, 吴祖煊*, 姜育刚

(复旦大学计算机科学技术学院 上海 200438)
(zxwu@fudan.edu.cn)

摘要: 针对低光照视频增强 (low-light video enhancement, LLVE) 任务中细节恢复不足与时空一致性缺失的问题, 提出一种基于暗转亮扩散模型的低光照视频增强方法 DL-Diff。首先, 基于预训练的潜在扩散模型构建基础模型, 将低光照视频增强任务转换为有条件的视频到视频的生成任务; 其次, 设计协同工作的恢复组件与时序组件, 其中恢复组件实现低光照视频到正常光照视频的映射, 时序组件则确保连续帧之间的时序一致性; 此外, 设计了一个多阶段训练流程, 通过分步优化网络参数从而逐步提高视频的恢复质量。在成对的 DID 和 SDS D 数据集上进行了充分的实验。定量实验表明, DL-Diff 在各个指标上达到了最优和次优的效果; 尤其在 DID 数据集上, DL-Diff 的空间指标 FID 达到 41.29, LPIPS 达到 0.17, 时序指标 AB (Var) 达到 25.40, MABD 达到 0.08, 均超越对比 LLVE 方法。同时, 定性结果也表明, DL-Diff 能够生成兼具空间对齐和时间连续性的正常光照视频, 且视觉效果优于对比方法。

关键词: 潜在扩散模型; 低光照视频增强; 恢复组件; 时序组件
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2024-00277

DL-Diff: Advancing Low-Light Video Enhancement through Dark-to-Light Diffusion Models

Wang Yuxuan, Xing Zhen, Wu Zuxuan*, and Jiang Yugang

(School of Computer Science, Fudan University, Shanghai 200438)

Abstract: To address the challenges of insufficient detail recovery and spatiotemporal inconsistency in low-light video enhancement (LLVE), this thesis proposes DL-Diff, a novel dark-to-bright diffusion model. The framework first establishes a conditional video-to-video generation task based on a pre-trained latent diffusion model. It then incorporates two synergistic components: a restoration component that learns the mapping from low-light to normal-light domains, and a temporal component that maintains inter-frame consistency. Furthermore, this thesis develops a multi-stage training strategy to optimize network parameters for enhanced video quality progressively. Extensive experiments on paired DID and SDS D datasets demonstrate DL-Diff's superior performance. On the DID benchmark, this method achieves state-of-the-art results with spatial metrics of 41.29 (FID) and 0.17 (LPIPS), along with temporal metrics of 25.40 (AB(Var)) and 0.08 (MABD), outperforming existing LLVE approaches. Qualitative evaluations confirm that DL-Diff generates visually pleasing videos with improved spatial alignment and temporal coherence compared to competing methods.

Key words: latent diffusion model; low-light video enhancement; restoration component; temporal component

收稿日期: 2024-06-04; 修回日期: 2025-04-24。国家自然科学基金 (62032006)。王雨轩 (2000—), 女, 硕士研究生, 主要研究方向为计算机视觉、视频内容理解; 邢桢 (1998—), 男, 博士研究生, 主要研究方向为计算机视觉、视频内容理解; 吴祖煊 (1991—), 男, 博士, 副教授, 博士生导师, CCF 会员, 论文通信作者, 主要研究方向为计算机视觉、视频内容理解; 姜育刚 (1981—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为计算机视觉。

低光照视频增强 (low-light video enhancement, LLVE) 是一项极具挑战性的任务, 低光照视频的噪声、颜色失真和细节丢失等各种因素使得恢复正常光照视频变得极其困难。然而, 该技术具有广泛的潜在应用, 包括摄影、自动驾驶系统和智能家居设备等。传统的深度学习方法主要针对低光照图像的恢复, 通过特定设计的模型结构^[1-5]提高低光照图像的可见性, 这些方法往往依赖于探索黑暗图像和正常光照图像之间的内在物理关系, 如 Retinex 理论和伽马强度校正。当前, 扩散模型的卓越生成能力使其在低光照图像增强 (low-light image enhancement, LLIE) 任务中展现出了巨大的潜力^[6-7]。这些基于扩散模型的 LLIE 方法^[8-10]通过学习低光照图像逆转成光照图像的扩散过程, 可以生成与正常光照图像近似的增强图像。与 LLIE 方法不同, 恢复低光照视频更加困难。一方面, 收集空间像素对齐的动态低光照-正常光照视频对很困难; 另一方面, 由于图像和视频之间存在领域差异, LLIE 方法在应用于视频时的效果往往不如在图像上的表现。LLVE 方法近年来取得了显著的进展, 现有方法^[11-14]设计专门的恢复模型, 以更好地处理空间和时间信息, 从而提高低光照视频的可见性, 同时保持相邻帧间的时间连续性。

针对 LLVE 任务, 本文提出了一种基于潜在扩散模型的低光照视频增强方法 (dark-to-light diffusion, DL-Diff), 不仅利用扩散模型的生成能力恢复低光照视频, 还能同时保持相邻帧间的时间连续性。DL-Diff 基于潜在扩散模型 (latent diffusion model, LDM) 构建了基础模型, 将预训练多模态生成式模型强大的生成能力迁移到视频增强任务中; 该方法设计了恢复组件和时序组件, 恢复组件用来提升低光照视频帧的能见度, 时序组件则用于在扩散模型的推理过程中确保生成视频帧之间的时序一致性, 并通过采取不同的装配方式使得组件间协同工作解决 LLVE 任务的难点; DL-Diff 提出了一个多阶段训练流程, 通过将预训练的文生图 (text-to-image, T2I) 扩散模型转化成视频到视频 (video-to-video, V2V) 扩散模型, 从而在 LLVE 任务场景下实现可控的视频生成, 这种设计使 DL-Diff 既能够有效地增强低光照视频的能见度, 又能保持视觉内容的连续性和时序一致性。

本文通过大量的定量和定性实验, 证明了 DL-Diff 的有效性。相较文中对比的 LLVE 方法, DL-Diff 在定量结果上取得了优异的恢复性能, 尤其在感知质量上实现了显著提升; 此外, 其在相邻

帧之间能保持更好的时序一致性。定性结果进一步表明, DL-Diff 在提高低光照视频的视觉质量方面具有出色的恢复能力。增强的视频帧显示出了良好的可见性和自然的色彩再现, 并且相邻帧之间保持了较好的时序连续性。实验结果验证了 DL-Diff 在 LLVE 任务中的突出表现, 并展示了其在极低光照视频增强中的优越性能。

1 相关工作

1.1 低光照视频增强方法

LLIE 方法通常以恢复低光照图像的空间信息为主要目标, 与该方法不同, LLVE 任务则更加注重处理相邻帧之间的时间连续性。为此, 现有 LLVE 方法通常通过设计恢复网络的结构, 以更充分地挖掘视频的时序信息。MBLLVEN^[11]采用 3D 卷积层从低光照视频中提取时间域的特征; StableLLVE^[13]提取各视频帧的光流, 并将其与低光照输入拼接以学习视频的运动场, 构建正常光照视频的序列; SMID^[15]中提出一种 ResUNet 结构, 并在原始格式的视频上完成 LLVE 任务, 其随机选择低光照视频中的 2 帧, 计算其与 1 帧真实的正常光照帧之间的差异; Wang 等^[16]提出动态低光照场景视频数据集 SDS, 并逐步将低光照视频与正常光照视频在空间上对齐; LAN^[17]是一个迭代的动态增强低光照视频框架, 同时构建了动态的低光照-正常光照视频数据集 DID。

尽管上述方法在 LLVE 任务中达到了较为优秀的表现, 但它们仍存在一些局限性和不足。MBLLVEN 只应用了简单的 3D 卷积, 导致对嘈杂的低光照视频内容的理解不充分; StableLLVE 是一个复杂的多阶段过程, 首先需要对单帧进行实例分割, 然后从分割结果中提取光流, 这些操作对于大规模数据可能耗时较长; SMID 仅关注静态暗视频, 未能有效地处理真实动态场景中的时间信息; 虽然 LAN 在 DID 数据集上表现良好, 获得了最高的重建分数, 但其缺乏时序评估且未公开源代码, 限制了其在实际中的广泛使用。与这些 LLVE 方法不同, DL-Diff 中引入预训练的 LDM 并利用扩散模拟过程恢复低光照视频, 生成效果更接近真实数据分布。与已有的仅评估重建质量的方法不同, 本文同时从模型理解和感知的角度评估被恢复视频的感知质量; 同时, 在时间序列的建模更加全面, 并在时间评估指标上取得了显著的性能提升。

1.2 用于低光照图像增强的扩散模型

与其他生成类模型相比,扩散模型能生成更真实的细节,并具有更强大的降噪能力,因此它们也被应用于包括 LLIE 任务在内的图像恢复任务。WCDM^[8]中提出一种基于小波的条件扩散模型,通过在小波域中整合高频细节增强低光照图像的可见性;CLE Diffusion^[9]通过将 SAM^[18]集成到条件扩散模型中,调节指定图像区域的亮度;Diff-Retinex^[10]将低光照图像分解为反射域图和光照度图,在这 2 个域中应用扩散过程,并将它们合并以获得增强的正常光照图像。与之前应用像素级扩散过程的方法不同,Diff-Plugin^[19]利用预训练的 Stable Diffusion^[7]的生成能力,在潜在空间中恢复低光照图像;通过设计自然语言指令和插件模块,指导 Stable Diffusion 的生成过程。

尽管上述方法在静态图像的低光照增强任务中表现出色,但它们在处理视频时面临帧间一致性问题,尤其是在增强过程中保持时序信息方面。为了解决这一问题,本文提出一种基于扩散模型的低光照视频增强方法,通过在视频帧间建立时序相关性,在获得更高质量的增强效果的同时保持帧间一致性;该方法不仅继承扩散模型的细节恢复和降噪优势,还专注于视频特有的时间一致性需求,进一步拓展了扩散模型在低光照条件下的应用。

1.3 基于预训练扩散模型的视频增强方法

鉴于扩散模型在图像增强任务中表现出色,近期研究开始探索将其应用于视频领域,尤其是基于预训练好的 T2I 扩散模型^[7]的方法。Upscale-A-Video^[20]将时间层集成到扩散模型的架构中,并设计了一个流引导的循环潜在传播模块解决视频超分辨率问题;VIDiff^[21]利用多模态指令控制扩散模型的编辑过程,并探索其在低光照增强任务中的应用;AVID^[22]在扩散模型中结合运动模块和结构引导,用于视频修复任务;LDMVFI^[23]是第 1 个使用预训练扩散模型解决视频帧插值任务的工作。

尽管上述方法展示了扩散模型在视频增强中的潜力,但它们大多关注的是视频分辨率、修复和插值等任务,相比之下,低光照视频增强不仅需要恢复清晰细节和去噪,还必须保持亮度稳定和时序一致。本文方法基于预训练扩散模型,并进一步引入低光照增强的条件约束,专注于在增强视频亮度的同时保证帧间一致性。

2 本文方法

首先介绍扩散模型的基础知识;然后描述

DL-Diff 的整体架构,包括基础模型、恢复组件和时序组件的结构和装配方式;最后介绍本文提出的多阶段训练方案。

2.1 扩散模型的基础知识

DDPM (denoising diffusion probabilistic model)^[24]由前向扩散和后向去噪 2 个过程组成。这 2 个过程通过模拟马尔可夫链,使模型能够生成接近真实数据分布的样本。在前向过程中,噪声 $\varepsilon \sim N(0, I)$ 被逐步引入干净图像 x_0 ,直到在时间步 t 得到噪声图像 x_t ,该过程可表示为

$$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \varepsilon \quad (1)$$

其中, α_t 表示预先定义的随时间变化的方差时间表。在后向过程中,模型在学习到的权重参数和噪声时间表的指导下,迭代地去除 x_t 的噪声,恢复原始的干净图像 x_0 。模型的训练目标是使得被添加的噪声 ε 与预测噪声函数 ε_θ 之间的均方误差最小,表示为

$$L_{\text{diff}} = E \left[\left\| \varepsilon - \varepsilon_\theta \left(\sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \varepsilon, t \right) \right\|^2 \right] \quad (2)$$

与原始的像素域 DDPM 不同, LDM^[7]的扩散过程是在低维的潜在空间中进行的。输入数据通过变分自编码器 (variational auto-encoder, VAE) 压缩到其潜在空间,并在这个低维潜在空间执行扩散过程,实现训练目标,表示为

$$L_{\text{LDM}} = E \left[\left\| z - z_\theta \left(\sqrt{\alpha_t} z_0 + \sqrt{1 - \alpha_t} \varepsilon, c, t \right) \right\|^2 \right] \quad (3)$$

其中, z 表示通过 VAE 获得的潜在特征变量, z_0 表示模型预测的潜在特征变量, c 表示文本条件。输出潜变量 z_θ 将输入到改进后的 U-Net^[25]结构中,进而重新映射到像素维度。该结构包括 4 个下采样块、1 个中间块和 4 个上采样块。LDM 还包含 1 个交叉注意层,用于在扩散过程中添加文本控制条件 c 。

2.2 DL-Diff 框架

首先介绍 DL-Diff 中 3 个组件——基础模型、恢复组件和时序组件的架构与功能;然后阐述如何将恢复组件和时序组件集成到基础模型中。

2.2.1 基础模型

DL-Diff 以预训练的 LDM^[7]作为基础模型。样本经过 VAE 的编码器后得到其潜在表征,将它们输入到改进后的 U-Net^[25]模型中,经扩散过程后输出处理过的潜在表征;然后将这些表征输入 VAE 的解码器,生成干净图像。

为了实现 V2V 的转换任务,并利用预先训练好的 T2I 扩散模型的先验知识,本文参考文献

[21,26-27], 将原始的 2D-UNet 膨胀为 3D-UNet 来处理视频输入。给定输入视频特征 $x \in \mathbb{R}^{b \times c \times f \times h \times w}$, 其中, b 表示批次, f 表示输入视频帧数在空间特征处理层, 特征图的维度被重塑为 $x' \in \mathbb{R}^{(b \times f) \times c \times h \times w}$ 。该操作允许模型的空间处理层在保留空间信息的同时, 独立地处理每个输入帧。在时序处理层中, 本文将特征图的维度重塑为 $x' \in \mathbb{R}^{(b \times h \times w) \times c \times f}$, 则模型可以处理输入的整个视频片段的视频帧在时间维度上的信息, 而忽略空间维度。通过时间层处理后, 特征图维度将重新恢复为原始形式 $x \in \mathbb{R}^{b \times c \times f \times h \times w}$, 确保最终输出的特征保持原始视频特征的结构。

2.2.2 恢复组件

恢复组件控制预训练扩散模型生成的内容。训练恢复组件时, 使用随机抽取的低光照和正常光照配对视频单帧引导控制: 利用低光照帧作为恢复组件的控制条件, 将其对应的正常光照帧作为生成目标训练恢复组件。将恢复组件作为特定层的旁路分支装备到预训练的 T2I 扩散模型上, 对低光照-正

常光照图像对进行训练, 如图 1 所示。恢复组件接收来自预训练扩散模型的潜在表征, 指导该模型的生成过程。恢复组件的输入包括: 来自主网络 U-Net 编码器的特征, 以及来自分支网络 (恢复组件) 的给定低光照图像的控制特征。恢复组件的输出是一组额外的残差特征, 这些残差将被添加到主网络 2D-UNet 的下采样解码器和中间模块中, 与模型对应层的原始输出以一定比例权重相加, 实现对主网络生成恢复图像的控制。修改后的特征图将通过模型的余下层影响最终输出。恢复组件的网络设计参考文献[28]的设置, 将主网络中 U-Net 编码器部分复制为 1 个独立且可训练的分支网络 (恢复组件), 如图 1 所示。恢复组件的结构由多个带有交叉注意力的下采样层和 1 个无交叉注意力的下采样层组成, 这种配置不仅可保留预训练模型的原始生成能力, 还能根据恢复组件的影响指导生成过程。该过程可以让恢复组件有效地控制生成过程, 使模型生成的图像与真实的正常光照图像非常接近。

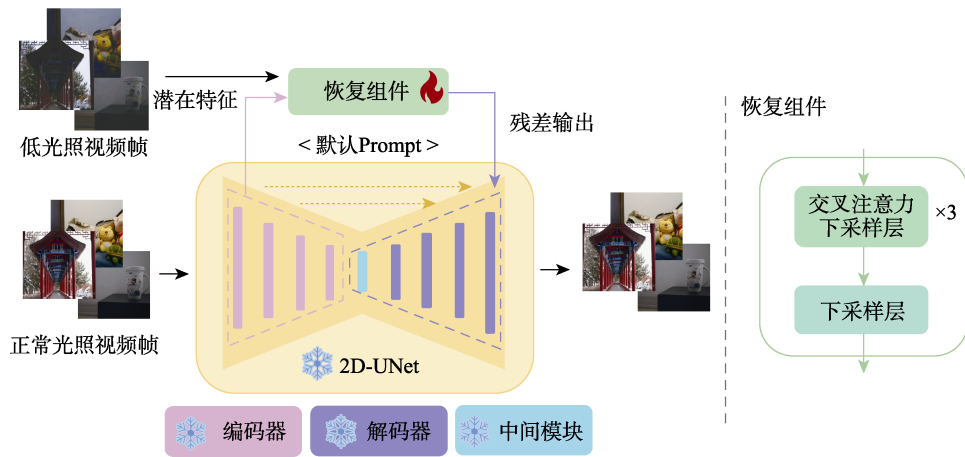


图 1 恢复组件的装配方式

2.2.3 时序组件

为了将图像生成 T2I 模型转换为视频生成 V2V 模型, 在 DL-Diff 的基础模型中加入时序组件, 该组件被插入经过膨胀处理的 3D-UNet 模型的空间处理层之后进行训练。时序组件的目的是沿时间轴学习连续帧间的动态变化和时间依赖关系。如第 2.2.1 节所述, 在输入时序组件之前, 特征将被重塑为 $x' \in \mathbb{R}^{(b \times h \times w) \times c \times f}$ 的形式, 使得时序组件能够在时间维度上对特征进行处理, 并实现跨帧的特征信息交换。

时序组件接收来自基础模型的潜在表征, 并输出一组中间时态特征, 如图 2 所示。与恢复组件通

过残差特征影响模型输出不同, 时序组件直接集成到基础模型的各层中, 通过操作基础模型的特征表征实现对时序信息的建模。本文中, 时序组件的网络设计参考文献[27]的设置, 通过在时间维度采用 Transformer 自注意力机制对动态先验进行建模; 使用多层自注意力模块捕捉帧间依赖关系, 并通过注意力机制加权融合每帧的信息, 确保帧间的连续性和一致性。

2.2.4 可控的视频转视频模型

已有的 LLVE 方法通常将时间信息纳入模型, 但模型的输入仍然是视频的各个帧, 即输入特征维度为 $[b, c, h, w]$ 。与之不同, 本文将这项任务视为

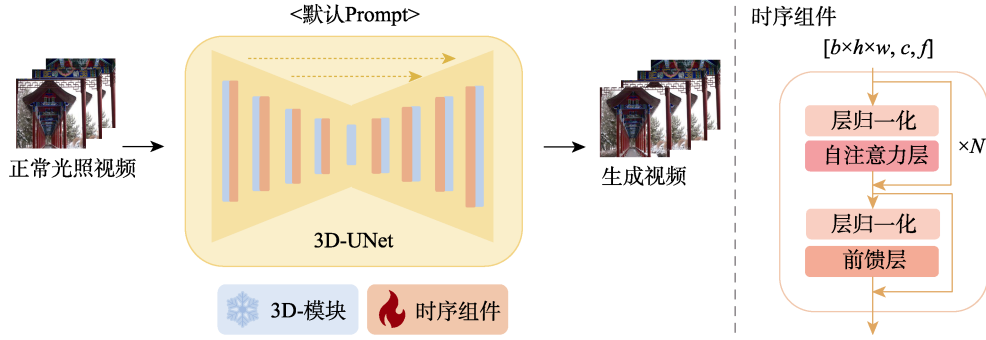


图2 时序组件的装配方式

V2V 的转换, 因此模型的输入是整个视频序列, 而不是分别对单帧进行操作, 即本文模型的输入特征维度为 $[b, f, c, h, w]$ 。本文将基础模型与恢复组件和时序组件相结合, 构建了一个可控的 V2V 模型, 其特征维度编号如图 3 所示。

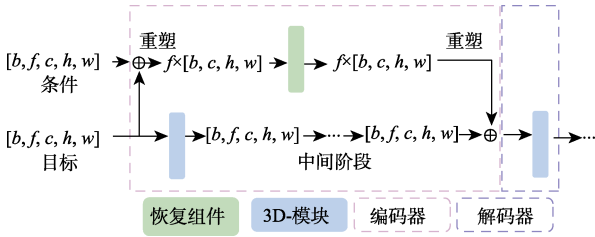


图3 V2V 模型内部特征维度变化

正常光照视频经过预先训练的 VAE 处理后生成潜在表征; 然后通过嵌入层将这些表征与低光照视频的表征合并, 合并后的表征经过重塑后被输入到恢复组件中, 按照视频帧的顺序输出 f 帧的残差特征。这些残差将按照相同的顺序添加到模型原始输出特性的相应帧位置中, 确保每个残差特征都能与相应的帧准确匹配; 此外, 模型的原始输出特性已通过时序组件进行处理, 提高了还原视频的时态一致性: 这 2 个过程的整合, 可确保视频修复在空间上的连贯性和时间上的一致性。

2.3 多阶段训练流程

在 LLVE 任务中, 本文执行的是 V2V 的转换。因此, 训练视频数据的构造为 $\langle V_d, V_l \rangle$, 其中, V_d 表示输入的低光照视频, V_l 表示相应的目标正常光照视频。预训练的 LDM 需要一个文本提示指导生成过程, 为防止生成无关内容, 本文采用不包含任何上下文信息的默认提示 c 。因此, 训练数据的形式为 $\langle V_d, V_l, c \rangle$ 。多阶段训练过程如图 4 所示。第 1 阶段中, 使用从训练视频对 $\langle V_d, V_l \rangle$ 中采样的成对帧训练恢复组件。由于恢复组件的训练基于 T2I 模

型, 因此训练时针对单帧低光照图像, 即输入特征维度为 $[b, c, h, w]$ 。该阶段中, 冻结预先训练好的 2D-LDM 的参数, 只对恢复组件中的参数进行训练。本文将随机采样的低光照视频帧作为恢复组件的输入条件, 将对应的正常光照帧作为目标来训练恢复组件。恢复组件将学习如何恢复低光照帧中丢失的细节的先验知识, 并将其增强到与正常光照帧相匹配的空间质量。第 2 阶段中, 时序组件被插入已经膨胀后的 3D-UNet 基础模型中, 并在不改变基础模型中预先训练的权重的情况下训练时序组件。该阶段中, 利用正常光照视频 $\langle V_l \rangle$ 训练时序模块, 目的是捕捉目标正常光照视频的时态依赖性。在最后阶段, 恢复组件和时序组件将同时被装配到基础模型上, DL-Diff 通过最终微调执行可控的 V2V 还原任务。本文使用配对视频 $\langle V_d, V_l \rangle$ 作为输入和目标, 通过微调可控 V2V 模型中的特定层, 生成的视频将保持更好的结构重建质量和细粒度。

3 实验及结果分析

3.1 实验设置

本文使用 Stable Diffusion v1.5^[7]作为基础的预训练 T2I 模型, 其构建基于 Stable Diffusion v1.5 架构。固定 Stable Diffusion 的预训练权重, 并将原始 2D-UNet 膨胀为 3D-UNet 构建基础模型; 此外, 分别按照 ControlNet^[28]的结构设计和 AnimateDiff^[27]的运动模块结构更好地初始化 DL-Diff 的恢复组件和时序组件。

在多阶段训练过程中, 单帧的训练数据被设置为 256×256 像素分辨率。为了避免引入不相关的内容信息, 将训练提示默认为“A detailed professional high-quality image/video shot under normal illumination”; 将扩散模型训练过程的时间步数设置为 1 000。在第 1 阶段训练恢复组件, 使用从训练视

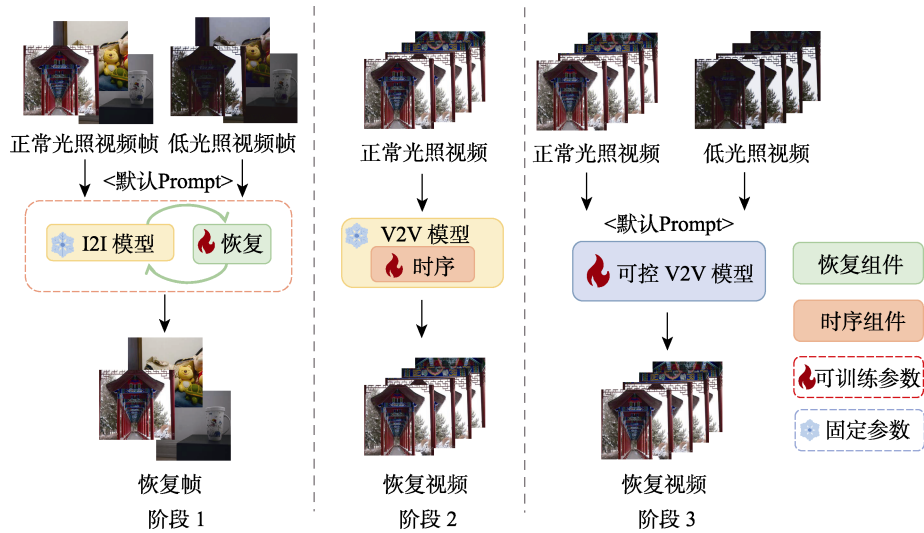


图4 DL-Diff的多阶段训练流程

频中随机采样的暗光帧对（静态），设置学习率为 1×10^{-5} ，并使用 AdamW 优化器；在第 2 阶段训练时序组件，使用连续采样的 16 帧暗光视频对（动态），将学习率设定为 1×10^{-4} ，并使用 Adam 优化器；在最后的微调阶段对可控 V2V 模型进行微调，参数设置与第 2 阶段的训练一致；在 DL-Diff 的推理阶段，模型的推断步数设置为 100，采用 DDIM 采样方法^[29]加速光照视频的生成。

3.2 数据集

本文使用真实世界中的低光照-正常光照配对视频数据进行训练和评估 DL-Diff，包括 DID 数据集^[17]和 SDS D 数据集^[16]。DID 数据集是 1 个动态视频数据集，具有明显的摄影运动特征，即视频中包含明显的物体运动；此外，包含 413 对从真实世界场景中收集的低光照-正常光照视频数据对，增强了数据集的多样性和真实性，分辨率为 2560×1440 像素，这些高质量视频对在像素维度上进行额外的对齐操作，使得这些视频在空间上精密对齐。SDSD 数据集包含 150 对视频序列，其中，包括 70 对室内视频和 80 对室外视频；每个视频包含 100~300 帧，分辨率均为 1920×1080 像素；尽管 SDS D 也是一个动态数据集，但它并没有在像素空间上达到精确的对齐度。

3.3 定量评估

由于 DL-Diff 专注于 V2V 的转换任务，因此本文采用全面的评估方法。将 DL-Diff 与已有的 LLVE 方法进行了对比，包括 MBLLVEN^[11]、StableLLVE^[13]、SMID^[15]、SDSD^[16]和 LAN^[17]方法。本文对恢复的视频帧的空间质量和时序一致性都进行多指标的评估。对于空间质量，评估了生成视

频帧与真实正常光照帧之间的感知质量和重构质量；因为时空一致性对于保持视频序列的连贯性至关重要，所以也对生成帧在时空一致性上的表现进行评估。通过综合评估，验证了 DL-Diff 既能保持视频序列的结构完整性，也能在时序一致性上提高视频质量的能力。

3.3.1 空间质量评估

本文从感知质量和重建质量 2 个角度评估 DL-Diff 的空间质量。（1）感知质量从人类观察者的角度判断恢复帧与原始帧的相似程度。本文通过学习感知图像块相似度（learned perceptual image patch similarity, LPIPS）和（Fréchet inception distance, FID）指标报告感知质量，它们的值越低表示感知质量越高。（2）重建质量衡量恢复后的帧与原始帧的像素一致性。本文使用峰值信噪比（peak signal-to-noise ratio, PSNR）和结构相似性（structural similarity index measure, SSIM）指标来呈现 DL-Diff 的恢复质量，它们的值越高表示重建质量越高。

在 DID 和 SDS D 数据集上，将 DL-Diff 与已有的 LLVE 方法进行对比实验，恢复视频的感知质量和重建质量对比如表 1 所示。

从表 1 可以看出：

（1）DL-Diff 在 FID 和 LPIPS 指标上均优于其他方法，表明其具有更优越的感知质量。在 DID 数据集上，DL-Diff 的 FID 和 LPIPS 分数分别为 41.29 和 0.17，明显低于对比方法，这表明 DL-Diff 能够更准确地恢复视频的真实感知质量；与 LLVE 方法相比，DL-Diff 借助扩散模型出色的建模能力，FID 和 LPIPS 分数有较高的提升，其增益效果得益

表1 在 DID 和 SDS D 数据集上 6 种方法的感知和重建质量对比

方法	DID				SDSD			
	PSNR/dB↑	SSIM↑	FID↓	LPIPS↓	PSNR/dB↑	SSIM↑	FID↓	LPIPS↓
MBLLVEN ^[11]	<u>24.82</u>	<u>0.91</u>	49.55	0.45	21.79	0.65	55.87	0.19
StableLLVE ^[13]	22.26	0.66	46.62	0.36	20.57	<u>0.74</u>	<u>39.20</u>	0.27
SMID ^[15]	22.97	0.87	<u>42.74</u>	<u>0.21</u>	24.09	0.69	204.23	0.21
SDSD ^[16]	21.88	0.83			<u>24.92</u>	0.73		0.13
LAN ^[17]	29.01	0.94			27.25	0.85		
DL-Diff	22.92	0.72	41.29	0.17	22.01	0.70	34.88	<u>0.14</u>

注: 粗体表示最优值, 下划线表示次优值。

于多阶段训练流程, 在最后阶段, DL-Diff 通过结合 Stable Diffusion 框架进行微调, 加入 LPIPS 和 (mean squared error, MSE) 损失强化解码器的恢复映射能力, 其中 LPIPS 指标表现尤为突出, 生成的视频更符合人类视觉感知的需求。

(2) 在重建质量方面, DL-Diff 也表现出很强的竞争力, 与对比方法接近。在 DID 数据集上, DL-Diff 的 PSNR 达到 23.12 dB, SSIM 达到 0.72; 在 SDS D 数据集上, DL-Diff 的 PSNR 达到 24.01 dB, SSIM 达到 0.74。这些结果表明, DL-Diff 能够在重建视频的细节和结构方面取得令人满意的效果。此外, 与像素级 LLVE 方法相比, 虽然 DL-Diff 在被压缩后的潜空间中进行扩散对重构产生一定的影响, 但也能得到与像素域方法相近的重建质量。

3.3.2 时序一致性评估

在 LLVE 任务中, 恢复视频的时序一致性是评估其生成质量的一个重要标准, 该标准衡量视频是否出现明显的亮度变化或不连续性。本文通过平均亮度方差 (average brightness variance, AB (Var))^[30] 指标和平均绝对亮度差值 (mean absolute brightness difference, MABD)^[12] 指标评估恢复视频的时序一致性, 它们的值越低, 表示恢复视频的平滑度和连续性越好。AB (Var)^[30] 度量的光照视频和正常光照视频之间的平均亮度和方差, 可以反映恢复视频的稳定性。MABD^[12] 指标则表示视频序列中每个像素亮度值随时间的平均变化率, 用于评估视频的平滑度和连续性。

在 DID 和 SDS D 数据集上, 采用不同方法进行时序一致性对比实验, 结果如表 2 所示。可以看出, DL-Diff 在 DID 数据集上取得了最好的时间结果, 在 SDS D 数据集上取得了次优的结果。(1) 在 DID 数据集上, DL-Diff 的 AB (Var) 结果更优, 分别比 MBLLVEN 低 9.06, 比 StableLLVE 低 2.53, 比 SMID 低 2.47; DL-Diff 的 MABD 得分为 0.08, 比

MBLLVEN、StableLLVE 和 SMID 分别降低 0.04、0.03 和 0.01。(2) 在 SDS D 数据集上, DL-Diff 也达到了次优的性能, 仅略逊色于 SDS D, 但相比其他方法有较高的提升。DL-Diff 在时序一致性上的优势可以归功于其时序组件的设计。其他方法只将其他时间信息 (如光流) 作为模型的额外输入, 模型的输入仍然是维度为 $[b, c, h, w]$ 的逐帧视频帧, 而 DL-Diff 直接将维度为 $[b, f, c, h, w]$ 的视频作为输入, 有效地捕捉了视频的时间动态。

表2 在 DID 和 SDS D 数据集上 5 种方法的 AB (Var) 和 MABD 指标对比

方法	DID		SDSD	
	AB (Var) ↓	MABD ↓	AB (Var) ↓	MABD ↓
MBLLVEN ^[11]	34.46	0.12	40.38	1.53
StableLLVE ^[13]	27.93	0.11	8.03	1.63
SMID ^[15]	<u>27.87</u>	<u>0.09</u>	10.94	1.47
SDSD ^[16]			2.37	0.94
DL-Diff	25.40	0.08	<u>7.45</u>	<u>1.02</u>

注: 粗体表示最优值, 下划线表示次优值。

3.4 定性评估

在 DID 和 SDS D 数据集上, 对 DL-Diff 与已有的 LLVE 方法进行定性评估, 结果分别如图 5 和图 6 所示。其中, 图 5a 和图 6a 所示为输入的原始低光照帧, 图 5f 和图 6f 所示为真实正常光照帧。可以看出, MBLLVEN^[11] 和 StableLLVE^[13] 方法都能生成保留原始视频结构的帧, 但这 2 种方法的生成结果在色彩饱和度方面仍与真实正常光照帧存在一定差异; SMID 生成的帧具有更好的色彩饱和度和亮度还原度, 然而它往往会过度锐化, 带来明显的噪点; 与 LLVE 方法相比, DL-Diff 能够生成更加优秀的视觉效果, 且没有出现色彩失真。值得注意的是, 在极低光照的环境中, DL-Diff 生成的效果明显优于对比方法, 如图 7 所示。可以看出, 与其他方法相比, DL-Diff 不仅在能见度方面表现

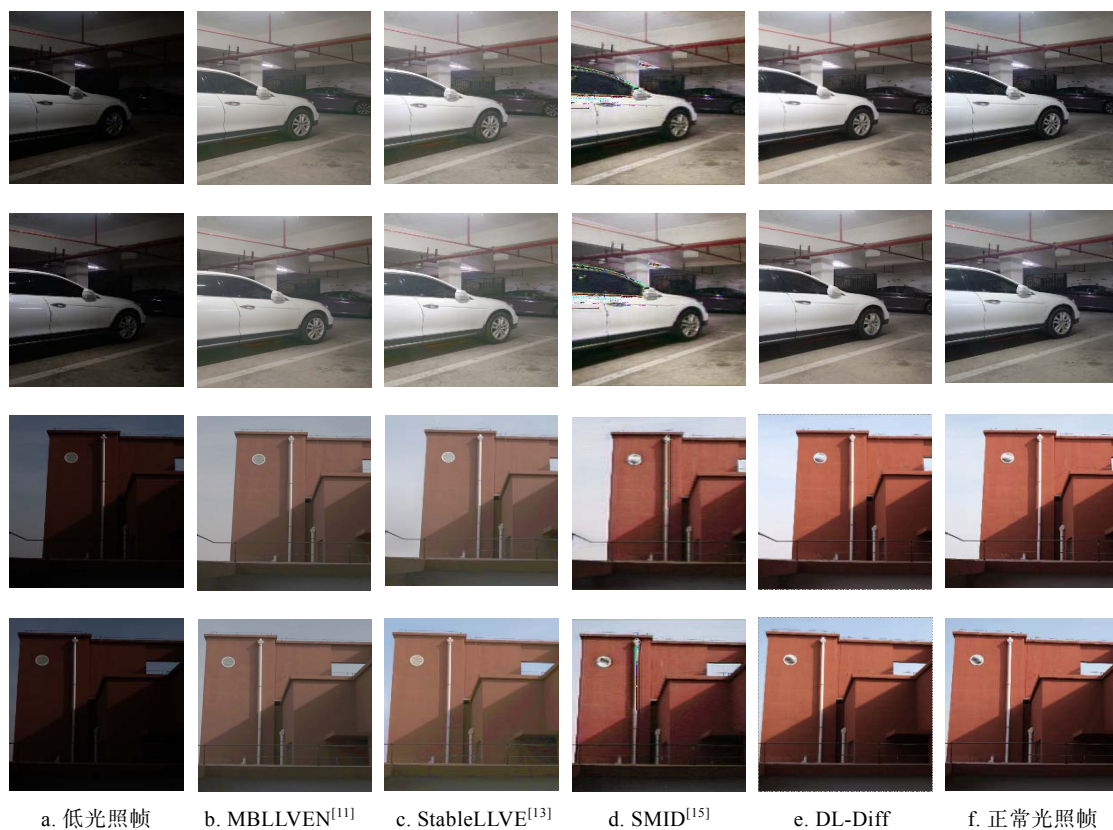


图 5 在 DID 数据集上 4 种方法的效果对比

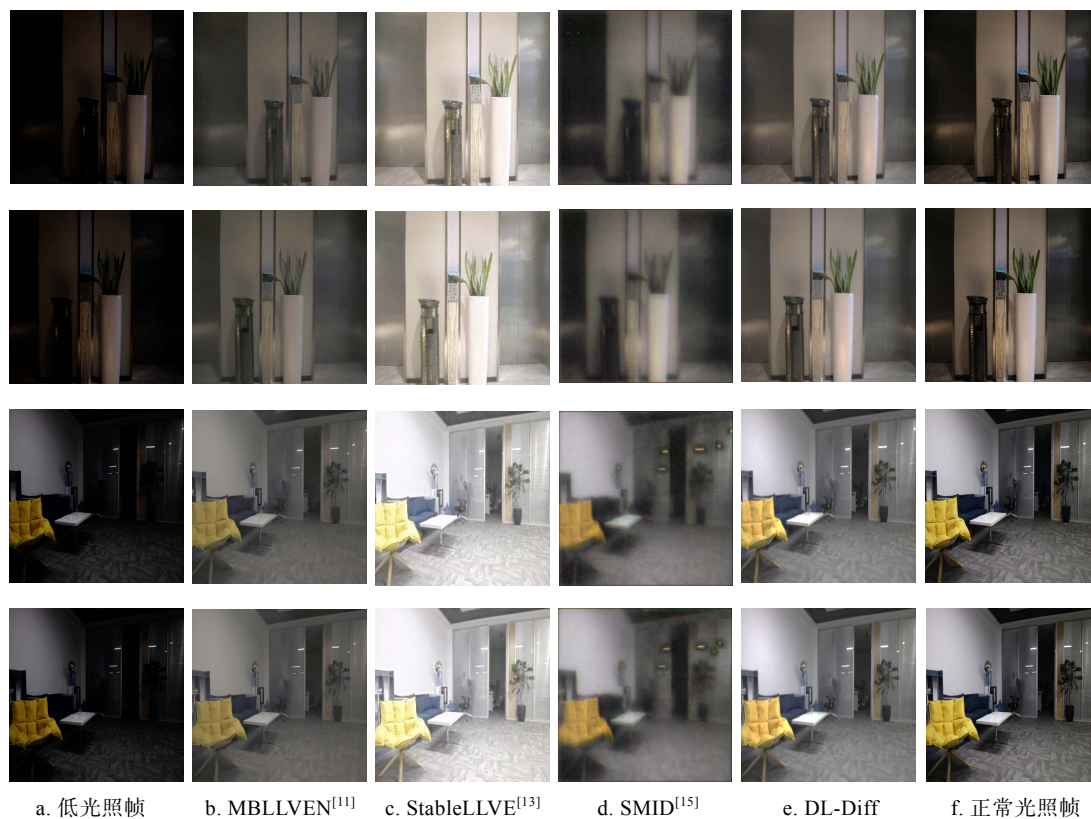


图 6 在 SDSD 数据集上 4 种方法的效果对比



图7 在极低光照条件下4种方法的效果对比

突出, 而且在色彩饱和度方面也呈现出非常优秀的视觉复原结果, 这种对真实世界照明条件的模拟源于 Stable Diffusion 模型强大的预训练数据, 广泛而多样的数据集使该模型能够学习和理解各种照明场景, 增强模拟和适应各种照明条件的能力。

3.5 消融实验

通过消融实验, 探究 DL-Diff 各组件以及多阶段训练策略对整体性能的影响。鉴于恢复组件接受低光照视频作为输入, 是 DL-Diff 的主要控制框架, 因此在消融实验中始终将其装配在基础模型上。

3.5.1 多阶段训练流程的有效性

为了验证多阶段训练过程的有效性, 尤其是最后一步微调的有效性, 对是否使用多阶段训练过程的模型的性能进行消融实验, 结果如表3所示。可以看出, 多阶段训练过程大大提高了重建质量, 使 PSNR 从 18.96 dB 提高到 22.92 dB, SSIM 从 0.44

表3 DID数据集上多阶段训练过程消融实验结果对比

方法	PSNR/dB↑	SSIM↑
无多阶段训练	18.96	0.44
DL-Diff	22.92	0.72

注: 粗体表示最优值。

提高到 0.72。

本文还对是否经过最后一步微调进行消融实验, 可视化结果如图8所示。可以看出, 采用多阶段流程 (尤其是微调步骤) 后, 训练的模型在结构重建能力方面优于视觉结果, 微调步骤能使模型更好地适应低光照视频的独特特性, 实现更精确的色彩还原和更出色的细节保留; 未经过微调步骤时, 模型生成的结果往往色彩表现不够准确, 而且会产生更明显的伪影, 尤其是在光线条件复杂多变的场景中。



图8 多阶段训练流程对 DL-Diff 效果的影响对比

进一步, 通过低光照增强后的视频帧间差分图展示相邻帧之间的像素差异, 以评估 DL-Diff 多阶段训练流程中微调阶段对帧间一致性、噪声抑制和细节保留的影响, 如图9所示。帧间差分图的平滑程度反映了相邻帧间的一致性: 过多不规则亮点表明增强过程存在不连续性或伪影; 而清晰且均匀的

边缘差分则说明模型能够较好地保留结构特征; 相反, 随机噪声表明模型在动态变化的建模方面存在不足。通过分析放大的细节区域可以看出, 图9b中存在较为明显的不规则亮点, 表现出增强过程中帧间一致性不足以及噪声较高的现象; 图9c中, 引入微调流程后, DL-Diff 显著地抑制差分图中的

噪声,同时增强了帧间的一致性,使差分图更加平滑;此外,清晰且均匀的边缘区域进一步表明,DL-Diff 微调流程对细节特征的保留能力更强,充分验证了DL-Diff的多阶段训练流程中微调阶段的必要性和有效性。

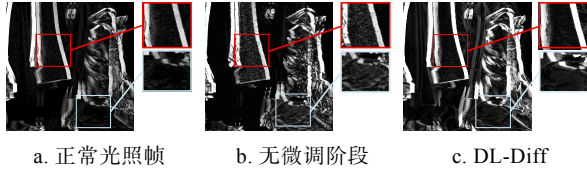


图9 微调阶段对DL-Diff效果的差分图结果对比

3.5.2 时序组件的有效性

为验证时序组件的有效性,本文进行了消融实验,并从客观指标和视觉结果2个维度进行对比分析。其中,定量对比结果见表4所示,可视化对比

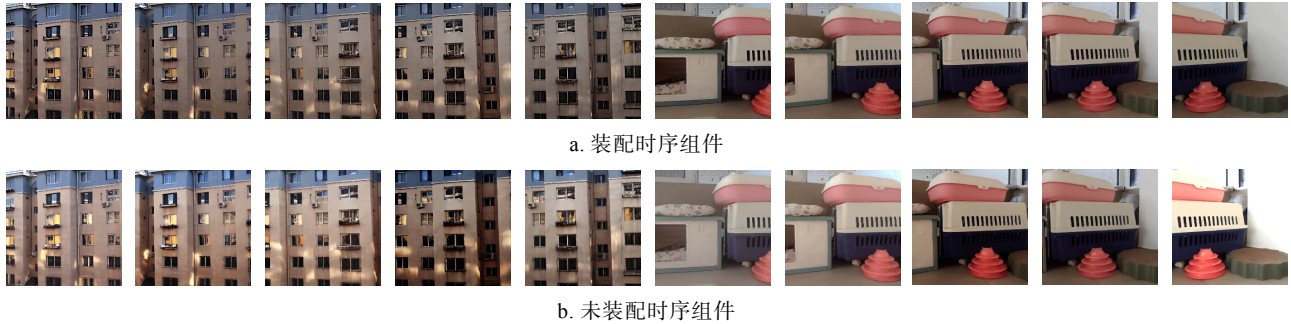


图10 时序组件对DL-Diff效果的影响对比

3.5.3 默认文本提示的有效性

在DL-Diff模型的训练和生成过程中,本文使用默认文本 Prompt 作为扩散模型的文本提示,以避免生成内容的不可控性。默认文本提示能够简化模型的控制条件,并通过恢复组件、时序组件以及多阶段训练流程,充分地挖掘预训练扩散模型在LLVE任务中的潜力。与复杂的文本提示设计相比,默认提示减少了人为干预和实验变量,更专注于评估DL-Diff模型的核心恢复能力。为了验证默认文本提示的有效性,引入自动生成的详细视频内容描述作为文本提示,对DL-Diff进行训练与测试。由于低光照视频通常缺乏明确的内容描述,实验中首先对低光照视频进行初步增强,然后利用多模态大语言模型LLaVA-NEXT^[31]生成与视频内容相关的详细文本描述,包括场景、物体、动作及光照条件等。

引入额外文本控制条件实验结果如图11所示。可以看出,将生成的文本描述作为提示引入模

型后,生成的视频在整体结构和物体位置上展现出一定相似性,但细节上表现出显著偏差,尤其在物体形状和关键特征上。文本提示的不准确性,使生成内容在某些情况下偏离原始光照视频目标。而默认文本提示在低光照恢复任务中具有以下优势:

表4 在DID数据集上时序组件消融实验结果对比

方法	AB (Var) ↓	MABD ↓
无时序组件	27.77	0.10
DL-Diff	25.40	0.08

注:粗体表示最优值。

从图10可以看出,移除时序组件后,恢复的连续帧的亮度和色彩饱和度受到影响,导致生成视频帧之间的不一致性;装配了时序组件的恢复帧在时间上保持较好的一致性,包括各帧的饱和度和亮度,视觉效果更好。

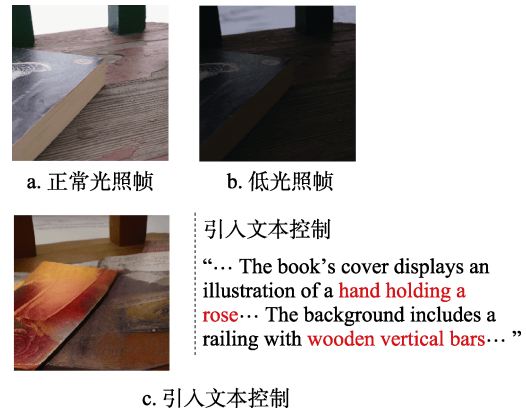


图11 引入额外文本控制条件对比结果

型后,生成的视频在整体结构和物体位置上展现出一定相似性,但细节上表现出显著偏差,尤其在物体形状和关键特征上。文本提示的不准确性,使生成内容在某些情况下偏离原始光照视频目标。而默认文本提示在低光照恢复任务中具有以下优势:

(1) 通过减少外部变量干扰,确保所有测试样本

的一致性; (2) 避免复杂文本控制条件引发的内容偏差, 提升模型在生成细节和结构恢复上的鲁棒性。虽然自动生成的文本提示展示了一定的可控性潜力, 但在低光照恢复任务中, 其误导性提示可能导致生成内容偏离原始目标。

4 结 语

针对 LLVE 任务, 本文提出 DL-Diff, 利用预先训练好的 LDM, 通过建立基于 3D-UNet 的基础模型处理 V2V 转换任务; 设计了特定的恢复组件和时序组件, 前者用于增强低光照帧的恢复质量, 后者可提高生成帧之间的时序一致性; 将这些组件集成到一个统一的框架中, 并采用多阶段训练策略逐步提高恢复质量。在 DID 和 SDS 数据集上进行广泛的实验, 验证了 DL-Diff 的有效性, 该方法能提供更好的感知质量和时间连贯性。

参考文献 (References):

- [1] Guo C L, Li C Y, Guo J C, *et al.* Zero-reference deep curve estimation for low-light image enhancement[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2020: 1780-1789
- [2] Jiang Y F, Gong X Y, Liu D, *et al.* EnlightenGAN: deep light enhancement without paired supervision[J]. IEEE Transactions on Image Processing, 2021, 30: 2340-2349
- [3] Wang Y F, Wan R J, Yang W H, *et al.* Low-light image enhancement with normalizing flow[C] //Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 2604-2612
- [4] Ma L, Ma T Y, Liu R S, *et al.* Toward fast, flexible, and robust low-light image enhancement[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 5637-5646
- [5] Hou J H, Zhu Z Y, Hou J H, *et al.* Global structure-aware diffusion process for low-light image enhancement[C] //Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2023: Article No.3490
- [6] Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis[C] //Proceedings of the 35th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2021: Article No.672
- [7] Rombach R, Blattmann A, Lorenz D, *et al.* High-resolution image synthesis with latent diffusion models[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2022: 10684-10695
- [8] Jiang H, Luo A, Fan H Q, *et al.* Low-light image enhancement with wavelet-based diffusion models[J]. ACM Transactions on Graphics, 2023, 42(6): Article No.238
- [9] Yin Y Y, Xu D J, Tan C C, *et al.* CLE diffusion: controllable light enhancement diffusion model[C] //Proceedings of the 31st ACM International Conference on Multimedia. New York: ACM Press, 2023: 8145-8156
- [10] Yi X P, Xu H, Zhang H, *et al.* Diff-retinex: rethinking low-light image enhancement with a generative diffusion model[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 12302-12311
- [11] Lv F F, Lu F, Wu J H, *et al.* MBLLEN: low-light image/video enhancement using CNNs[C] //Proceedings of the British Machine Vision Conference. Guildford: British Machine Vision Association Press, 2018: Article No.4
- [12] Jiang H Y, Zheng Y Q. Learning to see moving objects in the dark[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 7324-7333
- [13] Zhang F, Li Y, You S D, *et al.* Learning temporal consistency for low light video enhancement from single images[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 4967-4976
- [14] Liu L, An J F, Liu J Z, *et al.* Low-light video enhancement with synthetic event guidance[C] //Proceedings of the 37th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2023: 1692-1700
- [15] Chen C, Chen Q F, Do M, *et al.* Seeing motion in the dark[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 3185-3194
- [16] Wang R X, Xu X G, Fu C W, *et al.* Seeing dynamic scene in the dark: a high-quality video dataset with mechatronic alignment[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 9700-9709
- [17] Fu H Y, Zheng W K, Wang X C, *et al.* Dancing in the dark: a benchmark towards general low-light video enhancement[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 12877-12886
- [18] Kirillov A, Mintun E, Ravi N, *et al.* Segment anything[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 4015-4026
- [19] Liu Y H, Ke Z H, Liu F, *et al.* Diff-plugin: revitalizing details for diffusion-based low-level tasks[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 4197-4208
- [20] Zhou S C, Yang P Q, Wang J Y, *et al.* Upscale-A-Video: temporal-consistent diffusion model for real-world video super-resolution[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2024: 2535-2545

- [21] Xing Z, Dai Q, Zhang Z H, *et al.* VIDiff: translating videos via multi-modal instructions with diffusion models[OL]. [2024-06-04]. <https://arxiv.org/abs/2311.18837>
- [22] Zhang Z X, Wu B C, Wang X Y, *et al.* AVID: any-length video inpainting with diffusion model[OL]. [2024-06-04]. <https://arxiv.org/abs/2312.03816>
- [23] Danier D, Zhang F, Bull D. LDMVFI: video frame interpolation with latent diffusion models[C] //Proceedings of the 38th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2024: 1472-1480
- [24] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[C] //Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2020: Article No.574
- [25] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[C] //Proceedings of 18th International Conference on Medical Image Computing and Computer-Assisted Intervention-MICCAI. Heidelberg: Springer, 2015: 234-241
- [26] Guo Y W, Yang C Y, Rao A Y, *et al.* AnimateDiff: animate your personalized text-to-image diffusion models without specific tuning[OL]. [2024-06-04]. <https://arxiv.org/abs/2307.04725>
- [27] Blattmann A, Rombach R, Ling H, *et al.* Align your latents: high-resolution video synthesis with latent diffusion models[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2023: 22563-22575
- [28] Zhang L M, Rao A Y, Agrawala M. Adding conditional control to text-to-image diffusion models[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 3836-3847
- [29] Song J M, Meng C L, Ermon S. Denoising diffusion implicit models[OL]. [2024-06-04]. <https://arxiv.org/abs/2010.02502>
- [30] Chen Z Y, Abidi B R, Page D L, *et al.* Gray-level grouping (GLG): an automatic method for optimized image contrast enhancement-part II: the variations[J]. IEEE Transactions on Image Processing, 2006, 15(8): 2303-2314
- [31] Liu H T, Li C Y, Li Y H, *et al.* LLaVA-NeXT: improved reasoning, OCR, and world knowledge[EB/OL]. [2024-06-04]. <https://llava-vl.github.io/blog/2024-01-30-llava-next>