研究论文

基于扩散模型的深度估计与人像分割研究

董宗博,王一帆,王立君*,卢湖川

摘要 扩散模型在生成式任务中展现出强大的能力,但其在视觉感知任务(如深度估计与人像分割)中的应用仍有待深入探索。提出一种基于扩散模型的统一框架 Diffusion Perception,实现高质量深度估计与人像分割。通过将传统感知任务重新定义为条件生成问题,该框架利用潜在扩散模型(LDM)的去噪特性,在潜在空间中优化预测结果。创新性设计 3 种核心处理阶段:多模态特征编码阶段、噪声输入预测阶段和文本控制特征提取与重建阶段,使扩散模型从生成范式迁移到视觉感知任务范式上。实验表明,这种方法在自建深度估计数据集上对应的评估指标:相对精度(RR)、平面估计精度(plane)和场景一致性(consistence)分别达到了 93.98%、99.61%、93.61%,均优于现有先进的深度估计方法。此外,在人像分割任务中,对应的交并比(IOU)与平均交并比(mIOU)分别达到了 96.98%、91.98%,均优于现有分割算法。为扩散模型在视觉感知领域的应用提供了新思路,其生成式范式能够自然处理预测不确定性,适用于动态环境下的鲁棒感知任务。

关键词 扩散模型;深度估计;人像分割;全卷积神经网络;深度学习

视觉感知是计算机视觉核心方向,目标为通过算法深度理解视觉场景。从早期基础任务发展到目标检测、语义分割等复杂任务体系,其核心始终是高效准确提取二维图像/视频中有价值信息。深度学习(尤其是卷积神经网络(convolutional neural network, CNN)与 Transformer[1]架构的广泛应用[2])推动感知任务精度与效率显著提升,但现有算法多聚焦单一任务,无法跨任务迁移,需单独设计训练,难以实现任务统一与联合促进;同时在真实场景的鲁棒性、泛化能力及多任务协同上仍存挑战。

扩散模型(diffusion model, DM)作为新兴生成式模型,通过逐步去噪马尔可夫链生成高保真、多样图像,不仅革新图像生成,也为视觉感知提供新思路。其隐式表征学习、鲁棒不确定性建模及灵活条件生成特性,可应对传统感知任务诸多挑战,促使研究者探索其与视觉感知需求结合,开创全新研究范式。

当前视觉感知技术的发展呈现出显著特征。在模型架构方面,从早期的卷积神经网络,如 AlexNet[3]、VGGNet[4]等基础架构,到具有里程碑意义的残差网络(residual network, ResNet[5])通过残差连接解决了深层网络训练难题,再到密集连接卷积网络(densely connected convolutional network, DenseNet[6])通过特征重用进一步提升了网络表征能力,最终发展到如今基于自注意力机制的视觉 Transformer(vision transformer, ViT),特征提取能力得到了显著提升。以 ResNet、EfficientNet[7]等为代表的经典 CNN 架构通过精心设计的卷积块和网络拓扑,在图像分类、目标检测等任务上取得了突破性进展;而以 Swin Transformer[8]为代表的层次化注意力架构,通过局部窗口计算和跨窗口连接,在保持计算效率的同时实现了全局建模能力,

大连理工大学未来技术学院,大连 116024

收稿日期:2025-05-12;修回日期:2025-11-03
基金项目:国家自然科学基金(62276045, 62422610, U23A20386)
作者简介:董宗博,硕士研究生,研究方向为计算机视觉与深度学习,电子信箱:1246363088@mail.dlut.edu.cn;王立君(通信作者),教授,研究方向为计算机视觉与深度学习,电子信箱:ljwang@dlut.edu.cn
引用格式:董宗博, 王一帆, 王立君, 等. 基于扩散模型的深度估计与人像分割研究[J]. 科技导报, 2025, 43(22): 98-107; doi:10.3981/j.issn.1000-7857.2025.05.00059

标志着视觉表征学习进入了新阶段。在多任务学习方面,从早期的 Faster R−CNN[9]、Mask R−CNN[10]等基于 CNN 的检测分割算法,到 Panoptic−DeepLab[11]等统一框架通过共享 ResNet 或 Transformer 骨干网络和任务特定头部的设计,实现了检测、分割等任务的协同优化。在训练范式上,从基于 ImageNet 预训练的微调策略,到自监督预训练技术,如掩码自编码器(masked autoencoder, MAE[12])、自监督视觉 Transformer 的知识蒸馏(self−distillation with no labels, DINO[13])等通过大规模无标注数据学习通用视觉表征,显著降低了对标注数据的依赖,这些技术进步使得现代视觉感知系统在标准测试集上的性能不断提升,但在面对真实场景中的复杂情况时仍显不足。

现有视觉感知面临 3 大挑战:复杂场景适应性差(光照、遮挡等易致漏检/误分割[14])、判别与生成模型割裂[15]、动态场景(非线性运动)建模不足。扩散模型[16]为解决挑战提供新可能,其通过逐步去噪学习数据分布,可捕捉细粒度特征提升感知精度,概率化输出能量化不确定性适配困难样本,灵活条件生成机制支持多任务统一建模。据此提出基于扩散模型的统一视觉感知框架 Diffusion Perception,核心创新是将感知任务重定义为条件生成问题,通过任务自适应协同训练、专用条件注入及预训练大模型赋能,实现多任务协同优化,提升精度与收敛速度。

1 已有方案综述

1.1 视觉感知模型的发展与局限

视觉感知模型的演进历程反映了计算机视觉领域方法论的根本性转变。从传统手工特征时代过渡到深度学习驱动的特征学习,这一转变始于 2012 年 AlexNet[3]在 ImageNet 竞赛中的突破性表现。以 VGGNet[4]、ResNet[5]为代表的基础架构通过系统性的网络深度研究,确立了 CNN 在视觉感知任务中的主导地位。ResNet 的创新性残差连接设计不仅解决了深层网络训练中的梯度消失问题,更催生了一系列衍生架构,如聚合残差变换网络(aggregated residual transformations network, ResNeXt[17])通过分组卷积扩展了网络宽度,DenseNet 则通过特征重用机制实现了更高效的梯度传播。这些基础架构构成了第 1 代深度学习视觉感知系统的核心,支撑了包括 Faster R−CNN[9]、Mask R−CNN[10]等经典检测分割框架的发展。

随着注意力机制的兴起,视觉感知模型进入了架构创新的新阶段。Vision Transformer 首次证明了纯注意力架构在图像分类任务中的竞争力,而层次化视觉 Transformer(swin transformer, SwinT[8])通过引入层次化设计和局部窗口计算,在保持计算效率的同时实现了全局建模能力。这种架构创新带来了多任务学习范式的转变,金字塔视觉 Transformer(pyramid vision transformer, PVT[18])、SwinT 等统一骨干网络能够同时支持检测、分割等多种下游任务。端到端目标检测 Transformer(end−to−end object detection with transformers, DETR[19])系列工作则彻底改变了目标检测的范式,将检测任务重新定义为集合预测问题,从而实现统一的视觉感知任务,基于 Transformer 的编码器与解码器架构如图 1 所示。

1.2 扩散模型的进展与视觉应用

扩散模型作为生成式模型的新范式,其发展轨迹与视觉感知模型形成了有趣的对比。稳定扩散模型[20](stable diffusion, SD)模型作为当前最先进的文本到图像生成模型,其核心算法采用了潜在扩散模型(latent diffusion models, LDM)架构。该模型首先通过预训练的变分自编码器将高分辨率图像压缩到潜在空间(通常为 64×64×4 的维度),然后在潜在空间中进行扩散过程。SD 模型的去噪网络采用改进的 U−Net[21]结构,其中包含多尺度特征提取模块、时间嵌入层和交叉注意力机制,这些设计为后续的视觉感知任务应用提供了重要基础。

在深度估计领域,Marigold[22]和 Lotus[23]等算法基于 SD 模型进行了创新性改进。Marigold 通过将深度估计任务重新定义为条件生成过程,利用 SD 模型的潜在空间表示能力,在保持细节精度的同时实现了对复杂场景的鲁棒处理。Lotus 算法则进一步引入了几何一致性约束,通过在扩散过程中加入表面法线估计的辅助任务。基于预训练扩散模型的视觉感知(visual perception with a pre−trained diffusion model[24])模型创新性地设计了任务特定的解码器架构,针对深度估计任务优化了特征聚合机制。

在图像分割领域,DatasetDM[25]展现了扩散模型在结构化预测任务中的潜力。面向扩散模型感知的

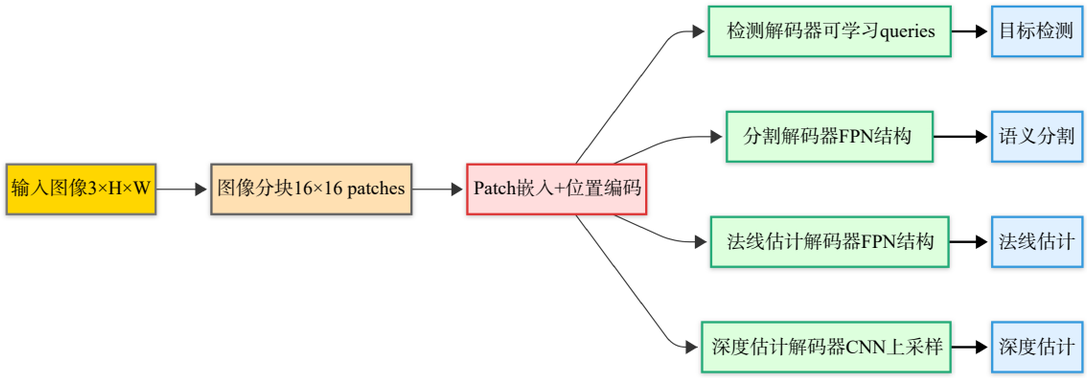

图 1 基于 Transformer 的编码器与解码器架构

文本图像对齐(text-image alignment for diffusion-based perception[26])进一步扩展了这一思路,针对不同感知任务(如语义分割、实例分割)设计了专用的解码器网络,通过动态路由机制实现任务自适应特征处理。然而,这些方法同样受限于对预训练特征的依赖,未能完全释放扩散模型在跨任务知识迁移方面的优势。

在计算方法上,扩散模型展现出独特的优势。其迭代式的去噪过程本质上是对数据分布的渐进式优化,每一步都包含了对全局信息的整合。这种特性使扩散模型能够自然地处理多模态输出,并量化预测的不确定性。研究表明,扩散模型在视觉感知任务中展现出巨大潜力。一些开创性工作尝试将检测、分割等传统感知任务重新定义为生成问题,利用扩散模型的迭代优化特性提升模型性能。

2 模型构建

2.1 SD 模型架构

稳定扩散模型(stable diffusion, SD)模型是一种基于潜在扩散模型(latent diffusion model, LDM)的生成架构,其核心结构由 3 个关键组件构成:变分自编码器(variational autoencoder, VAE[27])、U-Net 去噪网络和文本编码器(CLIP text encoder[28])。该模型通过将扩散过程置于低维潜在空间,显著提高了生成效率,同时保持了高质量的图像合成能力。SD 模型的创新之处在于巧妙地结合了深度学习领域多项前沿技术,构建了一个高效且强大的多模态生成框架,具体结构如图 2 所示。

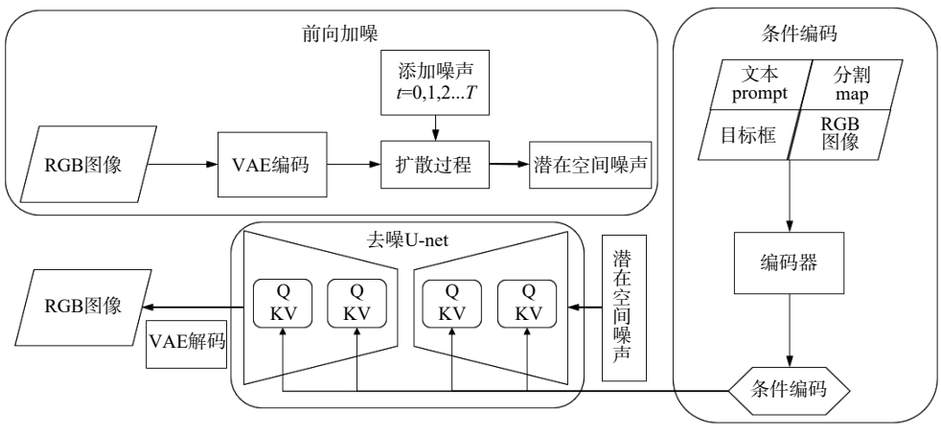

图 2 Stable Diffusion 模型架构

架构设计上,SD 模型采用分阶段处理策略:U-Net 去噪网络是 SD 模型的核心创新,它采用了多层次的特征提取和融合策略。该网络包含 4 级下采样和上采样路径,每层都配备了残差连接和注意力机制。文本编码器部分采用了 CLIP ViT-L/14 的预训练模型,CLIP 模型在大规模图文对数据上预训练获得的语义理解能力,为 SD 提供了强大的文本条件编码能力。VAE 组件将 512×512 等高分辨率图像压缩至 64×64×

4 潜在空间(压缩比约 48 倍),显著降低计算开销。VAE 模型的损失函数如下

$$L_{\mathrm{VAE}}=\mathbb{E}\left[\|x-D(E(x))\|^2\right]+\beta D_{KL}(q(z|x)\|P(z)) \quad (1)$$

其中,L_{VAE}是模型的总损失;$\mathbb{E}[\cdot]$代表对输入数据分布的期望;x 是模型的原始输入数据(如图像、文本等);$E(x)$是编码器的输出,作用是将输入x编码为隐变量z的“近似后验分布”$q(z|x)$的参数;$D(\cdot)$则是解码器的输出,负责将隐变量z解码为重建数据,以还原原始输入x;而$q(z|x)$是编码器学习到的、给定输入x时隐变量z的概率分布,$p(z)$则是预先设定的隐变量z的先验分布。

在训练目标方面,SD 采用了改进的噪声预测损失,其中噪声调度采用了余弦计划,这个选择相比线性调度能更好地平衡不同扩散阶段的学习重点,对应的噪声预测损失如下

$$Z_{t-1}=\sqrt{\alpha_{t-1}}\theta f(Z_t,t,c)+\sqrt{1-\alpha_{t-1}}\theta\varepsilon \quad (2)$$

式中,Z_{t-1}是扩散过程中第 t−1 步的状态变量;α_{t-1}是扩散过程中预定义的系数(与噪声强度相关,通常是一个随步数变化的衰减因子);θ代表模型(如神经网络)的参数;$f(Z_t,t,c)$是模型学习到的去噪函数,输入包含当前步状态Z_t、时间步 t 及可能的条件信息 c(如文本提示),输出是对“干净状态”的预测;ε是服从预设分布(通常为标准正态分布)的噪声项。SD 模型的推理过程也体现了精妙的设计。采用去噪扩散隐式模型(denoising diffusion implicit model, DDIM[29])采样算法,通过非马尔可夫链的确定性过程,将采样步骤从传统的 1000 次减少到 50 次左右,同时保持生成质量。

2.2 Diffusion Perception 模型介绍

基于以上 SD 模型的整体架构与目前主流的视觉感知任务算法相比,提出了 Diffusion Perception 模型,基于生成模型的范式实现视觉感知任务的实现。基于 SD 模型的多任视觉感知任务的整体训练与推理框架如图 3 所示。

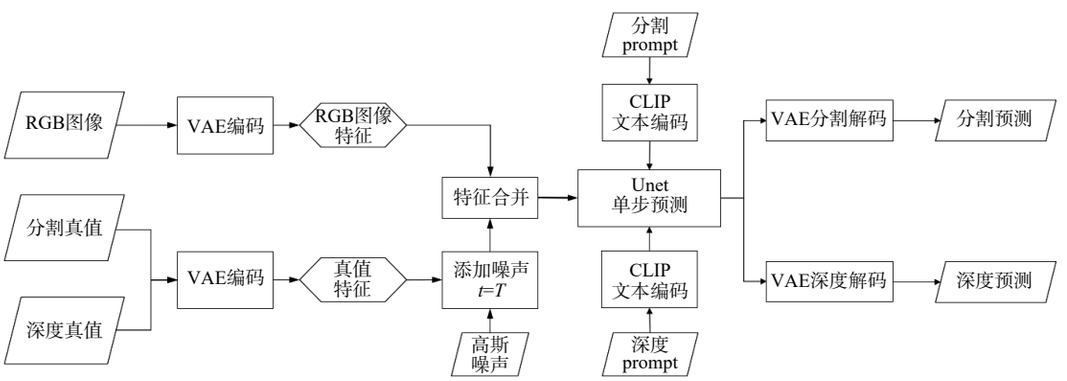

图 3 Diffusion Perception 算法训练流程图

模型分成 3 个核心处理阶段实现端到端的深度与分割图生成:多模态特征编码阶段、噪声输入预测阶段和文本控制特征提取与重建阶段。

1)多模态特征编码阶段。多模态学习(multi-modal learning)的核心挑战在于如何有效地融合不同模态的数据,使其互补优势并抑制噪声干扰。在视觉任务中,深度信息(depth)和分割(segmentation)能够提供彩色(RGB)图像所缺乏的几何与语义先验。

在保持 SD 模型范式不变的前提下,融入多模态深度分割信息存在显著挑战:简单拼接 RGB 与深度/分割图会因 VAE 编码器 3 通道输入要求引发架构冲突,重新训练 VAE 则耗费资源且破坏预训练权重优势;直接相加会导致 RGB 数据偏移,破坏潜在空间标准化特征。

在解决多模态数据融合的关键挑战时,本研究提出了一种创新的编码策略:将单通道辅助模态数据通过通道复制扩展为 3 通道伪 RGB 格式,经均值−方差归一化对齐数据分布,适配预训练 VAE 且规避重训成本。完成数据预处理后,模型采用双流编码的方式分别处理 RGB 图像和辅助模态数据。共享 VAE 编码器确保特征空间一致,对辅助模态潜在特征加噪以建立跨模态鲁棒关联。针对多模态维度变化,调整 U−Net 第 1 层输入通道数为 2 倍,以容纳来自 RGB 和辅助模态的双重信息,通过通道均分初始化权重继

承预训练能力。这种处理既保留了预训练模型的知识迁移优势,又为多模态学习提供了足够的容量。整个编码流程在保持 SD 核心范式不变的前提下,实现了深度和分割信息的高效融合,为视觉感知任务提供了更丰富的场景理解能力。

2)噪声输入预测阶段。传统的基于 SD 模型的深度估计与人像分割任务中,噪声预测过程通常采用多步迭代的方式,这一机制虽然在理论上能够渐进式地优化预测结果,但在实际应用中却暴露出诸多技术瓶颈。从理论层面分析,传统的噪声预测遵循以下公式

$$\varepsilon_\theta(x_t,t)=\sum_{i=1}^{N} w_i\cdot\varepsilon_\theta(x_t,t_i) \tag{3}$$

式中,$\varepsilon_\theta(x_t,t)$是模型的最终输出,$x_t$表示第 t 步的噪声图像,ε_θ为噪声预测网络,W_i为各去噪步骤的权重。这种多步预测机制虽然能够保证结果的渐进优化,但是会产生训练不稳定的问题。当迁移至视觉感知任务时,该特性会导致确定性预测的退化。这种预测不稳定性直接影响了感知任务的可靠性。

针对这一关键问题,提出了一种创新的单步噪声预测(single-step noise prediction, SNP)方法,从根本上重构了噪声预测的范式。SNP 方法的核心在于将传统的多步预测过程精简为具有决定性作用的单步预测,其数学表达为

$$\varepsilon_t^{\mathrm{SNP}}(x_t,t)=\varepsilon_\theta(x_1,1)\cdot\alpha(t) \tag{4}$$

式中,$\alpha(t)$是一个基于时间步 t 的动态缩放因子,用于保持预测噪声的强度一致性。SNP 方法采用了一种创新的噪声注入策略。与传统方法在训练时随机采样多个时间步不同,SNP 在每次训练迭代中仅添加单步高斯噪声,但同时完整保留了原有的时间步嵌入机制。

本研究在损失函数设计上进行了重要的范式创新,重新构建了生成范式与视觉感知范式之间的映射关系,将原本用于预测多步噪声的优化目标转化为直接预测潜在空间中的任务输出,建立从潜在空间到任务输出的端到端映射。为了维持模型的生成能力不被视觉感知任务削弱,在损失函数中创新性地引入了图像重建约束项。具体的损失函数表达式如下所示

$$L_t=\|z^x-f_\theta(z_t^y,z^x,t,s_x)\|_2^2+\lambda\|z^y-f_\theta(z_t^x,z^y,t,s_y)\|_2^2 \tag{5}$$

式中,L_t是模型的总损失;z^x代表模态 X(如图像)的特征表示,z^y代表模态 Y(如文本)的特征表示;$f_\theta(\cdot)$是模型(由参数θ控制)的预测函数,其输入包含另一模态的特征、当前模态的特征、时间步 t,以及模态专属的辅助信息s_x(模态X)或s_y(模态Y)。

3)文本控制特征提取与重建阶段。本研究提出了一种基于任务特定提示(task-specific prompt)的条件控制机制,通过将自然语言提示(prompt)作为任务描述符,动态引导 SD 模型的输出空间,从而实现任务间的精准区分与特征解耦。

Prompt 机制最初在文本到图像生成领域(如 DALL[30]、SD)中被用于控制生成内容。本研究创新性地将这一范式迁移至视觉感知任务,实现语义空间与视觉任务的对齐性。SD 的 U-Net 去噪网络通过交叉注意力层融合条件信息,其动态权重调整特性允许同一模型对不同 prompt 生成异构输出。本研究设计了 2 种类型的任务 prompt,对于深度估计任务。采用结构化描述(如“生成一张室外场景的深度估计图,要求边缘锐利,细节丰富”),通过强调深度图的场景与细节信息,引导模型学习场景的几何先验。同理,人像分割任务采用语义描述来增强语义信息的提取。

在推理阶段,本研究设计的 prompt 控制机制通过动态任务路由实现不同视觉感知任务的灵活切换,对应推理的过程如图 4 所示。

3 实验过程

3.1 实验环境

使用 PyTorch 实现 Diffusion Perception,并使用 Stable Diffusion v2 作为主干网络。设置文本条件并执行第 2.2 节中概述的步骤。在训练过程中,应用 DDIM 调度器[29],只采样最初的 1 步。训练方法使用 8 的批量大小进行 20 K 次迭代。使用学习率为 3×10^{-5} 的 Adam 优化器。此外,对训练数据应用随机水平翻转增强。在 8 张 Nvidia RTX V100 32G GPU 卡上训练方法收敛大约需要 7.5 d。

3.2 实验数据集

深度数据集:研究采用 2 个互补合成数据集训练以适配多样场景。室内场景选用 Hypersim[31]数据集(含 461 个逼真室内环境),按官方标准筛选 365 个场景的 54000 个有效样本(剔除缺失样本),RGB 图与深度图统一调整为 480×640 分辨率。深度处理关键操

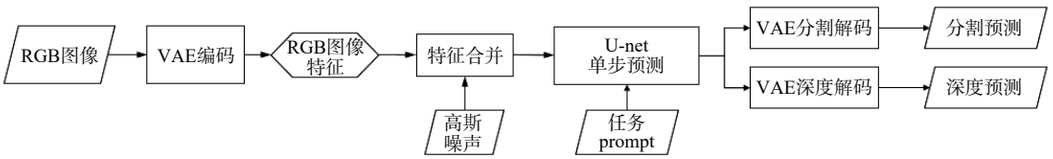

图 4 Diffusion Perception 算法推理流程

作:(1) 基于全局统计特性归一化深度值;(2) 通过透视投影逆变换,将原始相对焦点距离转换为视觉领域标准的相对焦平面深度表示。数学表达为

$$d' = \frac{f \cdot d}{d - f} \tag{6}$$

为扩展模型的室外场景适应能力,同时采用 Virtual KITTI 数据集[32]进行训练,该数据集包含 5 个典型街景场景的合成渲染。选取其中 4 个场景的约 20000 个样本构成训练集。考虑到实际应用场景的需求,所有图像均被裁剪至与真实 KITTI 基准一致的尺寸规格。在深度范围设定上,基于传感器性能分析将远平面阈值设置为 80 m,这一数值既覆盖了城市道路场景的典型观测需求(KITTI 数据统计显示 95% 的有效深度在 60 m 以内),又避免了过远距离带来的深度估计误差放大问题。

分割数据集:本研究构建人像分割高质量数据集。现有通用分割数据集存在人像标注粒度不足、复杂场景标注质量不稳定、人种多样性有限等缺陷。为此,设计系统采集与标注方案保障数据质量:采集采用双轨制,从 LAION−5B 等筛选 28000 张图像并采集真实生活人像;标注创新采用“SAM 预标注+三级人工校验”模式(SAM vit−huge 生成初始掩码,τ=0.85,经三级质控修正精细边界等关键细节);预处理实施标准化流程(512×512 双三次插值、光照增强等)。最终数据集含 65000 张高质量样本,详情见表 1。

表 1 深度估计与人像分割数据集

数据集类型	数据集名称	总数据量	训练集	测试集	场景覆盖
深度估计	Hypersim	54000	43200	10800	室内场景
	Virtual KITTI	20000	16000	4000	室外场景
人像分割	构建数据集	65000	52000	13000	室内外多光照

3.3 评估指标

本研究针对深度估计任务建立了多维度评价体系,重点考察几何关系保持、平面区域一致性和深度连续性 3 个关键特性。在深度估计任务中对应的评价指标如下所示。

相对关系正确率(RR)是评估深度图几何合理性的核心指标。该指标通过统计随机采样像素对的深度顺序一致性,量化模型对场景拓扑结构的理解能力。对应的计算公式为

$$RR = \frac{1}{N}\sum_{i=1}^{N} I\left(\mathrm{sign}(D_p^i - D_q^i) = \mathrm{sign}(\hat{D}_p^i - \hat{D}_q^i)\right) \tag{7}$$

式中,D_p^i和D_q^i表示采样点的真实深度,$\hat{D}_p^i$和$\hat{D}_q^i$表示对应的预测深度值,

平面区域符合度(plane)专门用于评估场景中主导平面(如墙面、地面)的深度估计质量。该指标通过 RANSAC 算法提取参考平面,计算预测深度与理想平面的偏离程度。在室内场景评估中,plane 指标与人类视觉对平面平整度的感知具有显著相关性。对应的计算公式如下

$$Plane = \frac{1}{M}\sum_{m=1}^{M}\left[\min_{\theta_m}\left(\frac{1}{|\Pi_m|}\sum_{p\in,\Pi_m}\left(D_p - \boldsymbol{\theta}_m^{\mathrm{T}}\phi_p\right)^2\right)\right] \tag{8}$$

式中,Π_m表示第 m 个平面区域,ϕ_p是对应像素的坐标,$\boldsymbol{\theta}_m^T$是平面参数向量。

深度一致性系数(consistence)用于量化局部区域的深度突变现象。该指标通过分析像素邻域内的最大深度差异,评估深度图的平滑连续性。优化 consistence 指标可以显著提升深度图的主观视觉效果,使物体边界过渡更加自然。

对于人像分割评价指标,交并比(IoU)作为分割任务的基础指标,直接反映了预测区域与真实标注的重叠程度。对应的计算公式如下

$$\mathrm{IoU}=\frac{1}{K}\sum_{k=1}^{K}\frac{\left|Y_k\cap\hat{Y}_k\right|}{\left|Y_k\cup\hat{Y}_k\right|} \quad (9)$$

式中, Y_k是第 k 个样本的真实分割掩码, $\hat{Y}_k$是预测的分割掩码。

多实例交并比(mIoU)是针对复杂多人场景设计的扩展指标。该指标通过最优匹配算法建立预测实例与真实实例的对应关系, 然后计算匹配对的平均IoU。相较于全局 IoU, mIoU 的优势在于能够公平评估模型对多个实例的分辨能力, 避免了密集人群场景中的评估偏差。具体的计算公式如下

$$\mathrm{mIoU}=\frac{1}{|M|}\sum_{(i,j)\in M}\frac{|Y_i\cap\hat{Y}_j|}{|Y_i\cup\hat{Y}_j|} \quad (10)$$

式中, mIoU 特别适用于评估模型对遮挡关系的处理能力, 是衡量实例感知性能的重要指标。Y_i为对应的真实实例, $\hat{Y}_i$是对应的预测实例。

3.4 实验结果

1) 深度估计算法性能对比。本研究选取了当前最具代表性的 5 种深度估计算法作为对比基准, 包括基于扩散模型的 Marigold[22]、Lotus[23]、大规模预训练的通用深度估计模型(universal depth estimation model, Depth Anything v2[33])、高精度深度重建模型(high-precision depth reconstruction model, DepthPro[34])以及最新优化版本 BetterDepth[35]。在不同的数据集上的深度评估指标如表 2 所示。

表 2 不同方法深度估计结果

模型	测试项目/%		
	RR	*Plane*	*Consistence*
Marigold	88.89	99.06	88.72
Lotus	83.91	97.30	87.29
Depth Anything v2	81.73	98.74	85.82
DepthPro	83.47	98.63	86.29
BetterDepth	85.42	99.38	86.67
Diffusion Perception	93.98	99.61	93.61

实验结果表明, 现有深度估计算法均存在显著性能局限: Marigold 基于 Stable Diffusion(SD)模型, 采用多步迭代过程导致推理速度远未满足实时性要求; Lotus 在 Virtual KITTI 等室外数据集上, 因难以适配大范围深度变化, 导致远距离区域估计效果不佳; Depth Anything v2 为基于 Transformer 架构的通用深度估计模型在大人脸数据上因深度变化范围过大, 深度一致性指标显著下降; DepthPro 在遮挡、光照变化场景下鲁棒性不足, 导致深度相对关系失真; BetterDepth 在保证实时性的同时优化估计精度, 但受限于庞大的网络结构, 推理耗时增加, 面临严峻的部署算力挑战。

Diffusion Perception 在多数据集上表现优异, 显著优于基线算法。量化分析表明, 其单目深度估计性能突出(表 2): 相对精度达 93.98%, 较最优 Marigold(88.89%)提升 5.09 个百分点, 优于传统 Depth 系列算法 8~12 个百分点; 平面估计精度 99.61% 接近理论极限, 规则平面平均误差小于 0.4%, 得益于几何先验约束模块; 跨场景一致性指标 93.61% 居首, 挑战性场景稳定率超 93%, 鲁棒性强劲。

2) 人像分割算法性能对比。在人像分割任务评估中, 本研究选取了 Segment Anything Model[36]、自监督视觉特征学习模型(self-supervised visual feature learning model, DINOv2[37])和 ViT[2]作为对比算法, 这些模型代表了从通用分割到特定任务优化的最新进展。分割一切模型(segment anything model, SAM)的通用性设计导致在专业人像分割任务中, 对发丝级细节和复杂遮挡场景的处理精度不足。DINO v2 在肤色和光照变化较大时的鲁棒性较差, 特别是在低光照条件下的分割质量下降明显。ViT 的两阶段处理流程引入了额外的计算开销, 实时性能受限。

本方法在人像分割任务中展现出显著优势。如表 3 所示, 在自建测试集上, 本算法的 IoU 达到 96.98%, 较 SAM 提升 1.13 个百分点; 在多实例场景下的 mIoU 为 91.98%, 比 SAM 提高 1.38 个百分点。这些性能提升主要来源于 3 个方面: 首先, 专业构建的训练数据集提供了更精确的标注边界; 其次, 任务特定的 Prompt 设计有效引导模型关注人体特征; 最后, 与深度估计任务的联合优化增强了模型对三维结构的理解能力。

表 3 深度估计与人像分割数据集

模型	测试项目/%	
	IoU	mIoU
SAM	95.85	90.60
Dino v2	94.78	84.70
ViT	93.33	80.76
Diffusion Perception	96.98	91.98

4 结论

综上所述,本研究提出的 Diffusion Perception 算法创新性地将 Stable Diffusion 的生成范式迁移至视觉感知任务,通过多模态特征编码机制实现了深度估计与人像分割任务的协同优化。该方法的核心贡献体现在 3 个方面:首先,设计了一种跨模态的特征融合架构,通过潜在空间对齐和动态权重分配,使不同感知任务间形成互补优势;其次,提出了改进的单步噪声预测范式(SNP),显著提升了模型的训练效率和推理速度;最后,开发了任务特定的 Prompt 控制机制,利用语义嵌入引导模型聚焦任务关键特征。但是 Diffusion Perception 目前也只是将其应用于深度估计与分割任务中,可以将这种范式应用于其他的视觉任务中,如法线估计,采用更大的生成模型来进一步提升模型性能也是未来的研究方向。

参考文献(References)

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems (NeurIPS). Long Beach, CA, USA: Neural Information Processing Systems Foundation, Inc, 2017: 5998−6008.

[2] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 1 6x16 words: Transformers for image recognition at scale[J]. arXiv: 2020: 2010.11929.

[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems (NeurIPS). Lake Tahoe, Nevada, USA: Neural Information Processing Systems Foundation, Inc, 2012: 1097−1105.

[4] Simonyan K, Zisserman A. Very deep convolutional networks for large−scale image recognition[J]. arXiv: 2014: 1409.1556.

[5] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 770−778.

[6] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 2261−2269.

[7] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. California: PMLR, 2019: 6105−6114.

[8] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2021: 9992−10002.

[9] Ren S Q, He K M, Girshick R, et al. Faster R−CNN: Towards real−time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137−1149.

[10] He K M, Gkioxari G, Dollár P, et al. Mask R−CNN[C]//Proceedings of IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 2980−2988.

[11] Cheng B W, Collins M D, Zhu Y K, et al. Panoptic−DeepLab: A simple, strong, and fast baseline for bottom−up panoptic segmentation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2020: 12475−12485.

[12] He K M, Chen X L, Xie S N, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2022: 15979−15988.

[13] Caron M, Touvron H, Misra I, et al. Emerging properties in self−supervised vision transformers[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2021: 9630−9640.

[14] Geirhos R, Jacobsen J H, Michaelis C, et al. Shortcut learning in deep neural networks[J]. Nature Machine Intelligence, 2020, 2: 665−673.

[15] Donahue J, Krähenbühl P, Darrell T. Adversarial feature learning[J]. arXiv: 2016: 1605.09782.

[16] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems (NeurIPS), 2020, 33: 6840−6851.

[17] Xie S N, Girshick R, Dollár P, et al. Aggregated residual transformations for deep neural networks[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 5987−5995.

[18] Wang W H, Xie E Z, Li X, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2021: 548−558.

[19] Carion N, Massa F, Synnaeve G, et al. End−to−end object detection with transformers[M]//Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 213−229.

[20] Rombach R, Blattmann A, Lorenz D, et al. High−resolution image synthesis with latent diffusion models[C]//Proceedings of IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2022: 10674−10685.

[21] Ronneberger O, Fischer P, Brox T. U−Net: Convolutional networks for biomedical image segmentation[M]//Medical Image Computing and Computer−Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing, 2015: 234−241.

[22] Ke B X, Obukhov A, Huang S Y, et al. Repurposing diffusion−based image generators for monocular depth estimation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 9492−9502.

[23] He J, Li H, Yin W, et al. Lotus: Diffusion−based visual foundation model for high−quality dense prediction[J]. arXiv: 2024: 2409.18124.

[24] Zhao W L, Rao Y M, Liu Z Y, et al. Unleashing text−to−image diffusion models for visual perception[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2023: 5706−5716.

[25] Wu W, Zhao Y, Chen H, et al. Datasetdm: Synthesizing data with perception annotations using diffusion models[C]// Advances in Neural Information Processing Systems (NeurIPS). New Orleans, Louisiana, USA: Neural Information Processing Systems Foundation, Inc, 2023: 54683−54695.

[26] Kondapaneni N, Marks M, Knott M, et al. Text−image alignment for diffusion−based perception[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 13883−13893.

[27] Esser P, Rombach R, Ommer B. Taming transformers for high−resolution image synthesis[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021: 12873−12883.

[28] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. Virtual Event: PMLR, 2021: 8748−8763.

[29] Song J, Meng C, Ermon S. Denoising diffusion implicit models[J]. arXiv: 2020: 2010.02502.

[30] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text−conditional image generation with clip latents[J]. arXiv: 2022: 2204.06125.

[31] Roberts M, Ramapuram J, Ranjan A, et al. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding[C]//Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2021: 10892−10902.

[32] Cabon Y, Murray N, Humenberger M. Virtual kitti 2[J]. arXiv: 2020: 2001.10773.

[33] Feng J S, Huang Z L, Kang B Y, et al. Depth anything V2[C]//Proceedings of Advances in Neural Information Processing Systems 37. Vancouver: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024: 21875−21911.

[34] Bochkovskii A, Delaunoy A Ã Ç, Germain H, et al. Depth pro: Sharp monocular metric depth in less than a second[J] arXiv: 2024: 2410.02073.

[35] Zhang X, Ke B, Riemenschneider H, et al. Betterdepth: Plug−and−play diffusion refiner for zero−shot monocular depth estimation[J]. arXiv: 2024: 2407.17952.

[36] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]// Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2023: 3992−4003.

[37] Oquab M, Darcet T, Moutakanni T, et al. Dinov2: Learning robust visual features without supervision[J]. arXiv: 2023: 2304.07193.

Research on depth estimation and portrait segmentation based on diffusion models

DONG Zongbo, WANG Yifan, WANG Lijun*, LU Huchuan

Dalian University of Technology School of Future Technology, Dalian 116024, China

Abstract While diffusion models have demonstrated remarkable capabilities in generative tasks, their application to visual perception tasks such as depth estimation and portrait segmentation remains underexplored. This paper proposes Diffusion Perception, a unified framework based on diffusion models for high−quality depth estimation and portrait segmentation. By reformulating traditional perception tasks as conditional generation problems, the framework leverages the denoising characteristics of latent diffusion models (LDMs) to optimize prediction results in latent space. The innovative design incorporates three core processing stages: multimodal feature encoding, noise input prediction, and text−controlled feature extraction and reconstruction, enabling the transition of diffusion models from generative paradigms to visual perception task paradigms. Experimental results demonstrate that on our custom depth estimation dataset, the proposed method achieves evaluation metrics of 93.98% Relative Accuracy (RR), 99.61% Plane Estimation Accuracy (Plane), and 93.61% Scene Consistency (Consistence), outperforming existing state−of−the−art depth estimation methods. Furthermore, in portrait segmentation tasks, the method achieves Intersection over Union (IoU) and mean IoU (mIoU) scores of 96.98% and 91.98% respectively, surpassing existing segmentation algorithms. This study provides novel insights into applying diffusion models in visual perception, where their generative paradigm naturally handles prediction uncertainty and is well−suited for robust perception in dynamic environments.

Keywords diffusion models; depth estimation; portrait segmentation; fully convolutional networks; deep learning

（责任编辑　王微）