

陈占国,陈振军,薛晨霞,等.基于单目相机的复杂场景深度估计网络[J].辽宁工程技术大学学报(自然科学版),2025,44(4):505-512.DOI:10.11956/j.issn.1008-0562.20240208
CHEN Zhanguo, CHEN Zhenjun, XUE Chenxia, et al. A complex scene depth estimation network based on monocular camera[J]. Journal of Liaoning Technical University(Natural Science), 2025, 44(4): 505-512. DOI: 10.11956/j.issn.1008-0562.20240208

基于单目相机的复杂场景深度估计网络

陈占国¹, 陈振军¹, 薛晨霞¹, 王国亮¹, 李金峰¹, 李玉廷¹, 于保才^{2*}

(1. 国家能源集团宝日希勒能源有限公司, 内蒙古 呼伦贝尔 021000;
2. 辽宁工程技术大学 鄂尔多斯研究院, 内蒙古 鄂尔多斯 017000)

摘要: 为提升复杂多变场景下深度估计的精度, 提出一种基于U型编码器-解码器的单目深度估计网络。采用Swin Transformer架构作为编码器核心, 实现对输入数据多层次、多尺度的精细化特征提取。采用逐层扩张卷积提取多尺度局部特征, 通过特征交互模块交互局部和全局特征, 实现对复杂场景的更全面理解。采用对称式Transformer解码器并结合图像块扩展层将相邻维度的特征图重塑为更高分辨率的特征图, 最终输出像素级深度预测结果。在NYU Depth v2数据集和KITTI数据集上进行定量实验。研究结果表明: 该网络在复杂多变场景中具有高效性和实用性。研究方法突破了传统方法在复杂多变场景下的局限性, 为深度估计的理论研究提供新的视角和方法。

关键词: 单目深度估计; U型编码器-解码器; 逐层扩张卷积; 特征交互模块; 对称式Transformer解码器

中图分类号: TP183

文献标志码: A

文章编号: 1008-0562(2025)04-0505-08

A complex scene depth estimation network based on monocular camera

CHEN Zhanguo¹, CHEN Zhenjun¹, XUE Chenxia¹, WANG Guoliang¹,
LI Jinyi¹, LI Yuting¹, YU Baocai^{2*}

(1. National Energy Group Baorixile Mineral Resources, Hulunbuir 021000, China;
2. Ordos Research Institute, Liaoning Technical University, Ordos 017000, China)

Abstract: To improve the depth estimation accuracy in complex and changeable scenes, a new monocular depth estimation network based on U-shaped encoder-decoder is proposed. The Swin Transformer architecture is adopted as the core of the encoder to realize fine-grained feature extraction of input data at multiple levels and scales. Multi-scale local features are extracted by using layer-by-layer dilated convolution. The local and global features are interacted through the feature interaction module to achieve a more comprehensive understanding of complex scenes. A symmetric transformer decoder is adopted and combined with an image patch expansion layer to reshape the feature map of adjacent dimensions into a feature map with higher resolution. Eventually, pixel-level depth prediction is output. Quantitative experiments are conducted on the NYU Depth v2 dataset and the KITTI dataset. The research results show that this network has high efficiency and practicability in complex and changeable scenes. The research conclusion breaks through the limitations of traditional methods in complex and changeable scenes and provides new perspectives and methodologies for the theoretical research of depth estimation.

Keywords: monocular depth estimation; U-shaped encoder-decoder; layer-by-layer dilated convolution; feature interaction module; symmetric transformer decoder

0 引言

单目深度估计是一种通过单个摄像头拍摄的

图像来推断场景中各物体深度信息的技术。从图像中估计深度信息是计算机视觉的一项基础且重要的功能, 可广泛应用于同步定位与建图

收稿日期: 2024-07-20 修回日期: 2024-08-26 接受日期: 2024-09-10 责任编辑: 黄琦

基金项目: 辽宁省教育厅科研项目(LJYL017)

作者简介: 陈占国(1976-), 男, 内蒙古呼伦贝尔人, 主要从事机电管理和矿山智能化方面的研究。E-mail: Lntu_songhuilun@163.com

通信作者: 于保才(1993-), 男, 辽宁朝阳人, 博士, 讲师, 主要从事计算机视觉与模式识别方面的研究。E-mail: lntuybc@163.com

(simultaneous localization and mapping, SLAM)^[1]、三维重建^[2]、目标检测^[3]和语义分割^[4]等领域。

近年来,神经网络卓越的图像处理能力为单目深度估计领域注入了新的活力,推动了从单目图像中直接生成密集深度图的端到端方法的发展,多种神经网络架构已充分验证了它在单目深度估计方面的高效性和准确性。LAINA等^[5]开创性地将残差网络(residual network, ResNet)引入深度估计的研究中,构建编码器-解码器结构网络模型,获取信息更为丰富和细致的特征图,为精准实现单目深度估计奠定了坚实的基础。YIN等^[6]将VGG(visual geometry group network)模型中的网络架构替换为ResNet,提高了深度估计效果。印雅萌等^[7]采用四分支卷积替代原有单分支卷积,研究表明四分支卷积能够有效识别图片中的物体边界等区域。在卷积神经网络(convolutional neural network, CNN)中,卷积运算固有的局部接受域特性限制了它对远程全局信息的捕获。为突破这一局限,可构建更深的主干网络或设计更复杂的架构来扩大感受野。但是这种策略往往伴随梯度爆炸和梯度消失的风险,不仅会增加模型训练的难度,还会影响模型的稳定性和泛化能力。Transformer架构在解决图像分类、目标检测、深度估计和密集预测等计算机视觉任务方面效果较好。SUI等^[8]提出一种多深度可分离卷积头转置注意力模块,通过独特的注意力机制有效地捕获了图像的全局信息,深度估计结果极佳。IBRAHEM

等^[9]提出基于ViT(vision transformer)的具有编码器-解码器架构的深度估计网络,打破了传统卷积神经网络在图像处理中的垄断地位。郑宇航等^[10]在编码器中采用混合注意力机制(convolutional block attention module, CBAM)和ResNet网络提取了4个不同尺度的特征图,有效提高了深度估计精度。李滔等^[11]结合空间和通道注意力机制,在编码器和解码器特征之间建立了更加精细高效的信息传递路径,显著提高了深度估计的精度和细节表现力。

受上述研究启发,本文提出一种基于对称Transformer的U型深度估计网络,通过Transformer的全局建模能力和U型结构的层次化融合特征,有效捕捉与理解图像的局部细节和整体结构,强化特征信息,提高深度估计的精度。

1 方法

网络总体架构见图1,其中, H 、 W 、 C 分别表示图像的高度、宽度和通道数。单张RGB图输入编码器后分别形成不同的特征尺度,在编码不同阶段,引入自适应逐层扩张卷积来捕捉图像的局部细节信息。同时,特征交互模块融合全局特征并增强特征之间的关联性,最大限度地提取图像特征。基于对称Transformer的解码器可准确还原图像的细节和结构,获取更为精准深度估计结果。

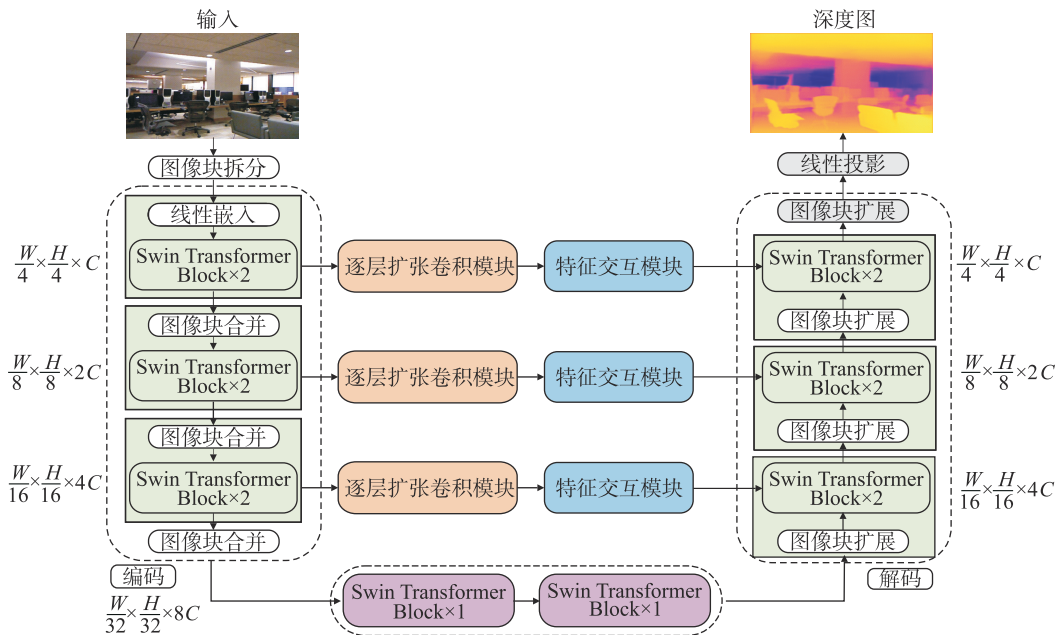


图1 网络总体架构

Fig.1 network overall architecture

1.1 U形编码器-解码器

本文提出的U型对称编解码网络结构以 Swin Transformer 模块为核心,包括编码器、图像块合并、瓶颈、图像块扩展以及解码器,各部分通过相互协作输出高质量的深度图像。

(1) Swin Transformer 模块。与传统的多头自注意力模块不同, Swin Transformer 模块基于移位窗口构建, 2个连续的 Swin Transformer 模块见图2, 其中, Z^L 和 Z^{L+1} 分别为 SW-MSA 和 MLP 模块在第 L 个 Block 的输出。

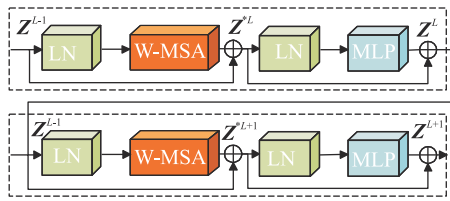


图2 2个连续的 Swin Transformer 模块

Fig.2 two consecutive Swin Transformer modules

每个 Swin Transformer 模块由 LayerNorm (LN) 层、多头自注意力模块、残差连接以及具有 GELU 非线性的多层感知机 (multi-layer perceptron, MLP) 组成。基于窗口的多头自注意力 (window based multi-head self-attention, W-MSA) 模块和基于移位窗口的多头自注意力 (shifted window based multi-head self-attention, SW-MSA) 模块被分别应用于 2 个连续的 Transformer 模块中。自注意力的计算式为

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{SoftMax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}} + B\right)\mathbf{V}, \quad (1)$$

式中: $\mathbf{Q}$ 、 $\mathbf{K}$ 、 $\mathbf{V}$ 分别为 query、key、value 向量, $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbf{R}^{M^2 \times d}$ 。其中, M^2 为 patch 在窗口中的数量; d 为键向量的维度; B 为偏置。

(2) 编码器。对于输入图像分辨率为原始图像分辨率 1/4 的图片, 虽然单个 Swin Transformer 模块能够提取图像的有效特征, 但由于图像的特征具有多层次性, 从低级的边缘、纹理到高级的形状、语义信息, 均须逐步抽象和深化, 单个 Swin Transformer 模块无法充分捕捉到这些不同层次的特征, 影响模型性能。因此, 采用 2 个连续的 Swin Transformer 模块进行深度表征学习, 促进特征融合, 增强泛化能力, 提高模型性能。

(3) 图像块合并。输入的图像块被分割为 4 部分, 并通过图像块合并技术进行高效融合。特征分辨率被有效地下采样至原始特征分辨率的一半, 由于 patch 合并操作, 特征维度增大 4 倍。为恢复特征维度并保持信息的一致性, 通过线性层对合

并后的特征进行变换, 将特征维度调整至原始维度的 2 倍, 在保持信息丰富度的同时, 实现特征的高效压缩。

(4) 瓶颈。针对 Transformer 模型堆叠多层时的收敛难题, 采用 2 个连续 Swin Transformer 模块构建瓶颈, 以实现高效深度表征学习。该设计在瓶颈区域维持结构特征尺寸和分辨率不变, 在保证性能的前提下实现网络训练稳定性。

(5) 图像块扩展。以解码器中的首个图像块扩展层为例, 在上采样操作之前, 通过线性变换层将输入特征的维度增大到原始维度的 2 倍。通过重排操作将输入特征的分辨率扩展至原始分辨率的 2 倍, 保持数据的完整性, 并将特征维数降低至原始维数的 1/4。此策略在保证信息丰富度的同时, 也能保证网络在处理过程中有效解码和处理图像数据。

(6) 解码器。为与编码器形成对称结构, 基于 Swin Transformer 模块构建相应的解码器。在解码器中, 采用图像块扩展层对编码器提取的深度特征进行上采样操作, 恢复原始分辨率。

1.2 逐层扩张卷积模块

逐层扩张卷积 (dilated convolution, LDC) 模块常用于提取多尺度的局部特征, 见图3。它采用 Swin Transformer 编码器, 通过层级结构提取输入图像的逐层特征。在特征提取过程的每个阶段, 特征图的分辨率会减半, 形成不同的特征尺度。在编码的每个阶段逐层插入连续的扩张卷积, 充分实现多尺度上下层聚合。

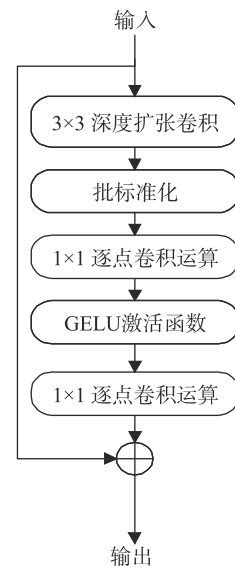


图3 逐层扩张卷积模块

Fig.3 layer by layer expansion convolution module

给定二维信号 $x[i]$, 则扩张卷积的输出 $y[i]$ 可定义为

$$y[i] = \sum_1^k x[i+k]w[k], \quad (2)$$

式中, $w[k]$ 表示长度为 k 的滤波器。

扩张卷积网络可在保证输出特征图大小固定的同时实现更大的感受野。尺寸为 $H \times W \times C$ 的输入特征 X 在LDC模块的输出为

$$\hat{X} = X + \text{Linear}_G(\text{Linear}(\text{BN}(\text{DWConv}(X))))), \quad (3)$$

式中: Linear_G 为逐点卷积运算; BN 为一个批量归一化层; $\text{DWConv}(\cdot)$ 为 3×3 的深度扩张卷积。

1.3 特征交互模块

LDC模块在图像处理中负责捕获丰富的多尺度局部特征, 然而, 由于它难以学习输入特征的全局表示, 从而难以对远程信息进行建模, 因此引入特征交互(feature interaction mechanism, FIM)模块, 见图4。FIM采用自注意力机制, 通过线性投影将长范围的全局信息编码到特征图中。

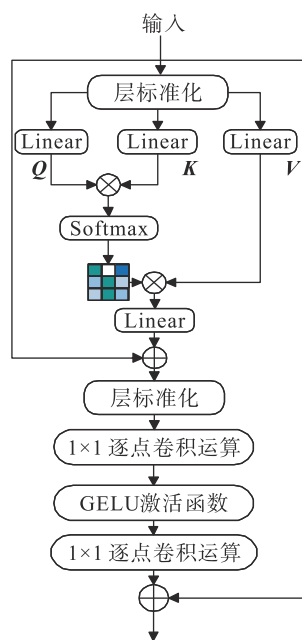


图4 特征交互模块

Fig.4 feature interaction module

输入特征 X 通过线性投影得到相同维度的查询 $Q=XW_q$ 、键 $K=XW_k$ 和值 $V=XW_v$, 其中, W_q 、 W_k 和 W_v 为权重矩阵。为减轻计算复杂度, 在信道维数上计算注意力, 得到增强输入 X^* 为

$$X^* = \text{Attention}(Q, K, V) + X. \quad (4)$$

将 X^* 与原始输入特征 X 逐元素相加, 实现特征的有效融合, 保留局部特征并增强全局信息表示, 使模型更有效地捕获图像的远程依赖关系。

为进一步增强特征表达能力, 引入 X' , 满足关系式

$$X' = X + \text{Linear}_G(\text{Linear}(L_N(X^*))). \quad (5)$$

采用层标准化操作来处理线性投影后的特征, 通过线性层和激活函数, 对经过标准化的特征进行非线性变换, 捕捉输入特征中更为抽象和复杂的信息, 提高模型的表征能力。

2 实验验证

2.1 数据集

NYU Depth V2数据集由RGB-D摄像机采集的室内场景视频序列构成, 包含1 449张RGB图像和1 449张深度图像。其中, 训练集包含249个场景的视频序列, 测试集包含215个场景的视频序列。

KITTI数据集有超过93 000张深度图像, 其中包含相应的原始激光雷达扫描图像和RGB图像。基准测试采用Eigen划分法, 训练集由来自28个不同场景的22 600张图像组成, 验证集包含888张图像, 测试集由来自28个不同场景的697张图像组成。

2.2 实验设备与参数设置

采用NVIDIA RTX 3090 GPU台式机开展实验, 基于PyTorch深度学习开发框架搭建网络模型。采用自适应学习率优化Adam算法进行网络训练, 超参数 β_1 、 β_2 分别设置为0.9、0.999。NYU Depth v2数据集共进行60轮训练, 耗时42 h, KITTI数据集共进行70轮训练, 耗时78 h。批量大小设置为8, 实验的初始学习率设置为0.000 2, 每5个周期将它降低10%, 权重衰减设置为0.02。

2.3 评价指标

采用7个常用指标来评估深度估计的准确度, 分别为绝对相对误差 E_{Abs} 、相对误差 E_{sq} 、均方根误差 E_{RMSE} 、对数误差 E_{lg10} , 以及阈值准确率 δ_1 、 δ_2 、 δ_3 。其中

$$E_{\text{Abs}} = \frac{1}{|N|} \sum_{d_i^* \in N} \frac{|d_i - d_i^*|}{d_i^*}, \quad (6)$$

$$E_{\text{sq}} = \frac{1}{|N|} \sum_{d_i^* \in N} \frac{\|d_i - d_i^*\|^2}{d_i^{*2}}, \quad (7)$$

$$E_{\text{RMSE}} = \sqrt{\frac{1}{|N|} \sum_{d_i^* \in N} \|d_i - d_i^*\|^2}, \quad (8)$$

$$E_{\text{lg10}} = \frac{1}{|N|} |\lg 10d_i^* - \lg 10d_i|, \quad (9)$$

$$\delta_i = \frac{1}{|N|} \left\{ d_i^* \in N \left| \max \left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i} \right) < 1.25^i \right. \right\} \times 100\% \quad (10)$$

$i = 1, 2, 3,$

式(6)~式(10)中: N 为每张图片的像素总数; d_i 为真实深度信息; d_i^* 为预测深度。

2.4 指标对比及可视化

不同方法在KITTI数据集上的评估结果见表1。与其他方法相比,本文方法在绝对误差、相对误

差和均方根误差方面表现出色,误差最小、预测最准。部分对比方法需采用额外数据来提升性能,而本文方法仅依赖KITTI数据集即可进行训练,且评估效果较好,进一步验证其高效性和泛化能力。KITTI数据集上的定性结果见图5。在光线不足环境下,本文方法较其他方法更能清晰勾勒目标轮廓;面对移动物体近摄等复杂场景,可准确预测物体的深度信息;在不同遮挡程度的性能测试中,预测表现稳定,展现出强鲁棒性。

表1 不同方法在KITTI数据集上的评估结果

Tab.1 evaluation results of different methods on the KITTI dataset

方法来源	编码器	误差指标评估结果				阈值准确率评估结果		
		E_{Abs}	E_{RMSE}	E_{lg10}	E_{sq}	δ_1	δ_2	δ_3
文献[12]	-	0.190	7.156	0.270	-	0.692	0.899	0.967
文献[13]	R101	0.072	2.727	0.120	0.307	0.932	0.984	0.995
文献[15]	R101	0.060	2.519	0.095	-	-	0.959	0.994
文献[16]	MViT	0.058	2.360	0.088	0.190	0.964	0.995	0.999
文献[18]*	ViTB	0.059	2.315	0.088	-	0.959	0.995	0.999
文献[19]	SwinL	0.052	2.129	-	0.155	0.974	0.997	0.999
文献[22]	SwinL+R-50	0.052	2.143	-	0.158	0.975	0.997	0.999
本文方法	SwinL	0.050	2.102	0.075	0.150	0.976	0.997	0.999

注:“-”表示未开展对应实验;“*”表示采用了额外数据来提升性能。

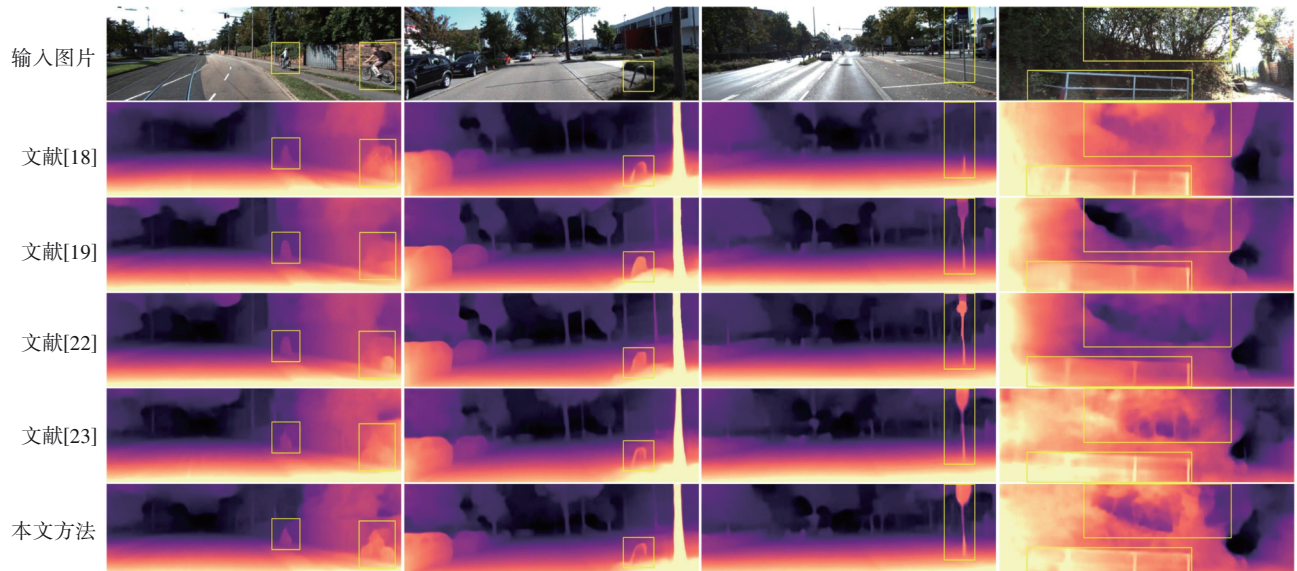


图5 KITTI数据集上的定性结果

Fig.5 qualitative results on KITTI

为全面验证本文方法的泛化能力和性能优势,在广泛使用的NYU Depth v2数据集上开展对比实验,评估结果见表2。当图片分辨率为540×198时,本文方法各项指标均优于其他近似分辨率的单目

深度估计方法。其中,与文献[23]所用方法相比,本文方法的均方根误差和相对误差分别降低了1.24%和13.95%,表明本文方法能有效提高深度估计的质量。

表2 不同方法在NYU Depth v2数据集上的评估结果

Tab.2 evaluation results of different methods on the NYU Depth v2 dataset

方法来源	误差指标评估结果				阈值准确率评估结果		
	E_{Abs}	E_{RMSE}	E_{lg10}	E_{sq}	δ_1	δ_2	δ_3
文献[12]	0.158	0.641	-	-	0.769	0.950	0.988
文献[13]	0.115	0.509	0.051	-	0.828	0.965	0.992
文献[14]	0.139	0.502	0.047	-	0.836	0.966	0.993
文献[15]	0.104	0.356	0.043	-	0.898	0.981	0.996
文献[16]	0.103	0.364	0.044	-	0.903	0.984	0.997
文献[17]	0.099	0.357	0.042	-	0.913	0.987	0.998
文献[19]	0.095	0.334	0.041	0.045	0.922	0.992	0.998
文献[20]	0.101	0.377	-	-	0.890	0.982	-
文献[21]	0.105	0.374	0.045	-	0.892	0.985	0.997
文献[22]	0.096	0.339	-	-	0.921	0.989	0.998
文献[23]	0.090	0.322	0.039	0.043	0.929	0.991	0.998
本文方法	0.089	0.318	0.036	0.037	0.930	0.993	0.998

注：“-”表示未开展对应实验。

NYU Depth v2 数据集的定性结果见图 6。由图 6 可知，在捕捉和再现较小空间物体的锐利边缘与细微纹理时，本文方法的精确度较高。对于灯杆类细长且纹理单一的物体，其他方法因分辨率或光照条件限制，出现深度估计模糊或失真问题，而本文方法注重轮廓细节还原，能清晰地勾勒灯

杆轮廓、准确捕捉桌子边缘结构，并精细还原桌面上的细微纹理和阴影变化。针对椅子脚等带黑色弱纹理待估计对象，本文方法通过良好学习能力推断补全椅子脚的深度信息，展现出强大的深度估计能力。



图6 NYU Depth v2数据集上的定性结果

Fig.6 qualitative results on NYU Depth v2

通过以上分析可知,本文方法性能明显优于目前的主流网络模型,特别是在深度图成像质量方面,表现出色。

2.5 消融实验

为验证本文方法的有效性,采用 EIGEN 等^[12]提出的训练和测试分割法,从 46 976 张图像中选取 697 张图像进行测试,去除网络中的逐层扩张卷积(LDC)模块,并在 KITTI 数据集上开展不同模块的消融实验,实验结果见表 3。由 1~3 组实验结果可知,不同模块组合下各精度和误差指标变化趋势各异。与第 5 组实验相比,第 6 组实验各误差

指标的评估结果显著降低,表明去除 LDC 模块会显著降低网络的特征提取能力。与第 1 组实验相比,第 4 组实验的均方根误差和对数误差分别降低了 0.33% 和 2.56%,这是由于特征交互模块可在相邻帧之间建立充足的空间,实现位置信息交互,增强特征提取能力。对比第 4、6 组实验结果可知,U 型编码器-解码器对平衡网络结构、改善特征提取质量具有重要作用。在 LDC、FIM 和 U 型编码器-解码器模块的共同作用下,各指标均达到最优状态,表明本文方法可获取更加精确的深度信息,验证了不同模块在深度估计中的有效性。

表 3 在 KITTI 数据集上的消融实验结果
Tab.3 results of ablation experiments on KITTI dataset

组别	模块			误差指标评估结果			阈值准确率评估结果	
	LDC	U-Encoder-Decoder	FIM	E_{RMSE}	E_{lg10}	E_{Abs}	δ_1	δ_2
1	√	×	×	2.124	0.078	0.051	0.973	0.997
2	×	√	×	2.120	0.076	0.051	0.974	0.997
3	×	×	√	2.119	0.077	0.051	0.973	0.997
4	√	×	√	2.117	0.076	0.052	0.974	0.997
5	×	√	√	2.110	0.077	0.051	0.975	0.997
6	√	√	√	2.102	0.075	0.050	0.975	0.997

注:“√”表示采用对应模块;“×”表示不采用对应模块。

3 结论

(1) 提出一种面向复杂多样性场景的单目深度估计增强网络。通过连续扩张卷积和特征交互处理,保留目标物体更多的细节特征;在解码阶段引入对称的 Swin Transformer 来优化初始深度估计结果,提高预测结果的一致性和准确性。

(2) 本文网络在深度估计中的误差更小,能够更准确地捕捉边缘和轮廓信息,在复杂场景中表现出更强的鲁棒性。

对于极端天气以及特殊动态场景,本文方法的适应性有限,未来可结合光流估计、运动预测等技术,进一步提高网络对动态场景的建模能力。

参考文献(References):

[1] 刘冬,于涛,丛明,等.基于深度学习图像特征的动态环境视觉 SLAM 方法[J].华中科技大学学报(自然科学版),2024,52(6):156-163.
LIU Dong, YU Tao, CONG Ming, et al. Visual SLAM method for

dynamic environment based on deep learning image features[J]. Journal of Huazhong University of Science and Technology (Natural Science Edition),2024,52(6):156-163.
[2] 郭宝云,姚玉凯,李彩林,等.改进的 3D-BoNet 算法应用于点云实例分割与三维重建[J].测绘通报,2024(6):30-35.
GUO Baoyun, YAO Yukai, LI Cailin, et al. Application of improved 3D-BoNet to segmentation and 3D reconstruction of point cloud instances[J]. Bulletin of Surveying and Mapping, 2024(6):30-35.
[3] 张志佳,范莹莹,邵一鸣,等.基于改进 YOLO v3 模型的多类交通标识检测[J].沈阳工业大学学报,2023,45(1):66-70.
ZHANG Zhijia, FAN Yingying, SHAO Yiming, et al. Detection of multi-type traffic signs based on improved YOLO v3 model[J]. Journal of Shenyang University of Technology, 2023, 45(1):66-70.
[4] 刘经纬,周彦.基于局部特征聚合网络的三维语义分割[J].计算技术与自动化,2024,43(2):170-176.
LIU Jingwei, ZHOU Yan. Local feature aggregation networks for 3D semantic segmentation[J]. Computing Technology and Automation, 2024, 43(2):170-176.
[5] LAINA I, RUPPRECHT C, BELAGIANNIS V, et al. Deeper depth

- prediction with fully convolutional residual networks[C]//2016 Fourth International Conference on 3D Vision (3DV). October 25-28, 2016, Stanford, CA, USA. IEEE, 2016: 239-248.
- [6] YIN Z C, SHI J P. GeoNet: unsupervised learning of dense depth, optical flow and camera pose[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-23, 2018, Salt Lake City, UT, USA. IEEE, 2018: 1983-1992.
- [7] 印雅萌, 周嘉麒, 王指辉. 基于多分支卷积的单目深度估计方法研究[J]. 计算机与数字工程, 2023, 51(12): 2966-2970.
- YIN Yameng, ZHOU Jiaqi, WANG Zhihui. Research on monocular depth estimation based on diverse branch block[J]. Computer & Digital Engineering, 2023, 51(12): 2966-2970.
- [8] SUI X, GAO S, XU A G, et al. Lightweight monocular depth estimation using a fusion-improved transformer[J]. Scientific Reports, 2024, 14(1): 22472.
- [9] IBRAHEM H, SALEM A, KANG H S. RT-ViT: real-time monocular depth estimation using lightweight vision transformers[J]. Sensors, 2022, 22(10): 3849.
- [10] 郑宇航, 曹雏清. 基于多尺度特征混合注意力的连续帧深度估计[J]. 重庆工商大学学报(自然科学版), 2024, 41(4): 104-111.
- ZHENG Yuhang, CAO Chuqing. Continuous frame depth estimation based on multi-scale feature mixed attention mechanism[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2024, 41(4): 104-111.
- [11] 李滔, 胡婷, 武丹丹. 结合金字塔结构和注意力机制的单目深度估计[J]. 图学学报, 2024, 45(3): 454-463.
- LI Tao, HU Ting, WU Dandan. Monocular depth estimation combining pyramid structure and attention mechanism[J]. Journal of Graphics, 2024, 45(3): 454-463.
- [12] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, Massachusetts: MIT Press, 2014: 2366-2374.
- [13] FU H, GONG M M, WANG C H, et al. Deep ordinal regression network for monocular depth estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-23, 2018, Salt Lake City, UT, USA. IEEE, 2018: 2002-2011.
- [14] RAMAMONJISOA M, LEPETIT V. SharpNet: fast and accurate recovery of occluding contours in monocular depth estimation[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop. October 27-28, 2019, Seoul, Korea (South). IEEE, 2019: 2109-2118.
- [15] PATIL V, SAKARIDIS C, LINIGER A, et al. P3Depth: monocular depth estimation with a piecewise planarity prior[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-24, 2022, New Orleans, LA, USA. IEEE, 2022: 1600-1611.
- [16] FAROOQ BHAT S, ALHASHIM I, WONKA P. AdaBins: depth estimation using adaptive bins[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 20-25, 2021, Nashville, TN, USA. IEEE, 2021: 4008-4017.
- [17] BHAT S F, ALHASHIM I, WONKA P. LocalBins: improving depth estimation by Learning local distributions[C]//Computer Vision-ECCV2022. Cham: Springer Nature Switzerland, 2022: 480-496.
- [18] RANFTL R, BOCHKOVSKIY A, KOLTUN V. Vision transformers for dense prediction[C]//2021 IEEE/CVF International Conference on Computer Vision. October 10-17, 2021, Montreal, QC, Canada. IEEE, 2021: 12159-12168.
- [19] YUAN W H, GU X D, DAI Z Z, et al. Neural window fully-connected CRFs for monocular depth estimation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-24, 2022, New Orleans, LA, USA. IEEE, 2022: 3906-3915.
- [20] LONG X X, LIN C, LIU L J, et al. Adaptive surface normal constraint for depth estimation[C]//2021 IEEE/CVF International Conference on Computer Vision. October 10-17, 2021, Montreal, QC, Canada. IEEE, 2021: 12829-12838.
- [21] LEE S, LEE J, KIM B, et al. Patch-wise attention network for monocular depth estimation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 1873-1881.
- [22] LI Z Y, CHEN Z H, LIU X M, et al. DepthFormer: exploiting long-range correlation and local information for accurate monocular depth estimation[J]. Machine Intelligence Research, 2023, 20(6): 837-854.
- [23] AGARWAL A, ARORA C. Attention attention everywhere: monocular depth prediction with skip attention[C]//2023 IEEE/CVF Winter Conference on Applications of Computer Vision. January 2-7, 2023, Waikoloa, HI, USA. IEEE, 2023: 5850-5859.