智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于图像边缘相似性的室内自监督单目深度估计

寇旗旗, 陈飞宇, 张华强, 程德强, 韩成功

引用本文:
寇旗旗, 陈飞宇, 张华强, 等. 基于图像边缘相似性的室内自监督单目深度估计[J]. 智能系统学报, 2026, 21(3): 713-726.
KOU Qiqi, CHEN Feiyu, ZHANG Huaqiang, et al. Indoor self-supervised monocular depth estimation based on image edges similarity[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(3): 713-726.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202505005

DOI: 10.11992/tis.202505005
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20260205.1431.004

基于图像边缘相似性的室内自监督单目深度估计

寇旗旗[1], 陈飞宇[2], 张华强[2], 程德强[2], 韩成功[2]
(1. 中国矿业大学 计算机与科学技术学院, 江苏 徐州 221116; 2. 中国矿业大学 信息与控制工程学院, 江苏 徐州 221116)

摘　要: 本文针对室内单目深度估计中因结构复杂、边缘重叠严重及旋转分量大导致的深度推理不准确问题, 提出一种基于图像边缘相似性的自监督深度估计网络模型。1)引入图像边缘相似性损失函数, 作为形状先验约束, 缓解因遮挡和重叠引起的性能下降; 2)设计自适应特征聚合模块, 融合多尺度特征并保持上下文一致性, 增强弱相关场景的语义关联; 3)提出旋转量优化模块, 通过加权融合不同路径的向量来细化位姿估计中的旋转分量, 降低旋转误差。实验结果表明, 该方法在 NYU Depth V2 与 ScanNet 数据集上的深度预测精度分别达到 82.9% 与 78.0%, 优于现有先进方法, 能够恢复出细节丰富、边缘清晰平滑的深度图, 有效提升了室内场景的深度估计效果。

关键词: 自监督; 单目深度估计; 图像边缘相似性; 特征聚合; 位姿优化; 室内场景; 形状先验; 上下文一致性
中图分类号: TP391.4　　**文献标志码**: A　　**文章编号**: 1673−4785(2026)03−0713−14

中文引用格式: 寇旗旗, 陈飞宇, 张华强, 等. 基于图像边缘相似性的室内自监督单目深度估计 [J]. 智能系统学报, 2026, 21(3): 713–726.
英文引用格式: KOU Qiqi, CHEN Feiyu, ZHANG Huaqiang, et al. Indoor self-supervised monocular depth estimation based on image edges similarity[J]. CAAI transactions on intelligent systems, 2026, 21(3): 713–726.

Indoor self-supervised monocular depth estimation based on image edges similarity

KOU Qiqi[1], CHEN Feiyu[2], ZHANG Huaqiang[2], CHENG Deqiang[2], HAN Chenggong[2]
(1. College of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China; 2. College of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China)

Abstract: In this paper, we propose a self-supervised depth estimation network model based on image edge similarity to address the issue of inaccurate depth inference in indoor monocular depth estimation due to complex structures, severe edge overlapping, and large rotational components. First, we introduce an image edge similarity loss function as a shape prior constraint to mitigate performance degradation caused by occlusions and overlaps. Second, we design an adaptive feature aggregation module to fuse multi-scale features while maintaining contextual consistency, thereby enhancing semantic associations in weakly related scenes. Finally, we propose a rotation optimization module that refines the rotational components in pose estimation by weighted fusion of the vectors from different paths, reducing rotation errors. Experimental results show that our method achieves depth prediction accuracies of 82.9% and 78.0% on the NYU Depth V2 and ScanNet datasets, respectively, outperforming existing state-of-the-art methods. The proposed method can recover depth maps with rich details and clear, smooth edges, effectively improving depth estimation in indoor scenes.

Keywords: self-supervision; monocular depth estimation; image edges similarity; feature aggregation; pose optimization; indoor scenes; shape priors; contextual consistency

深度估计是近年来计算机视觉领域研究热点, 在各种三维感知任务如虚拟现实[1]、自动驾驶系统[2]、目标检测[3]中扮演着不可或缺的角色, 其最初主要是通过深度摄像头、激光雷达等传感器来获取各场景中的深度信息。由于传感器昂贵的生产成本以及相机校准困难等诸多问题, 使得深度估计技术的应用非常受限。因而结构简单易实现、计算开销小的基于单目图像的深度估计受到

收稿日期: 2025−05−13.　**网络出版日期**: 2026−02−05.
基金项目: 中央高校基本科研业务费专项资金项目(2024ZDPYCH1001); 国家自然科学基金项目 (52204177, 52304182).
通信作者: 程德强. E-mail: chengdq@cumt.edu.cn.

了众多研究者们的青睐，该方法旨在利用单幅图像，逐像素地推算出各场景目标至摄像机成像中心的距离，即深度，进而生成深度图像。同一幅2D平面图像都有可能是无穷的不同3D场景投影而成，并且缺少稳定的物理线索来约束这种不确定性，使得该技术在数学领域中成为了一项病态的议题[4]。早期的发展使用基于线索的传统深度估计方法进行研究，利用图像中一些辅助信息比如纹理、阴影、照明程度以及遮挡等特征信息来帮助确定像素点的深度值，然而这些传统方法对场景要求十分严格，对于复杂场景的深度信息预测收效甚微。随着机器学习的蓬勃发展，特别是卷积神经网络的出现，利用编−解码器的神经网络架构和深度真值标签的有监督单目深度估计取得了有效的突破。然而设备采集的为稀疏深度真值，使得许多像素点并没有对应的深度真值，不利于网络架构训练，此外大规模获取深度真值标签存在较高难度与高昂成本，传统有监督深度估计方法的发展受到显著制约。为此，无需依赖深度真值标签进行监督约束的自监督深度估计方法，逐渐成为一种高效且具有广泛应用前景的替代方案。

近几年的研究中，大部分单目深度估计模型均采用了编−解码器结构来预测深度。编码器从输入图像中提取多尺度特征，解码器通过通道拼接或元素叠加方法逐渐将多尺度特征聚合起来，虽然这些特征聚合操作在现有的室外数据集研究中已经取得了一定程度的有效性，当编−解码器结构迁移至室内场景时，其局限性逐渐凸显。与室外场景相比，室内环境具有深度变化剧烈、物体边缘交错重叠、场景布局复杂（如家具堆叠、墙角遮挡）的特性，传统固定的特征聚合方式难以自适应匹配不同尺度特征的上下文关联，易导致信息冗余或语义差距扩大，最终使预测深度图出现边缘模糊、细节丢失的问题；同时，室内单目视频序列多由手持设备（如 Kinect、手机）拍摄，其旋转分量远大于室外车载相机序列，现有位姿网络（如单路径卷积位姿估计器）对旋转分量的预测误差较大，进一步引发重投影失真，加剧深度估计精度下降[5]。尽管近年来研究人员针对室内场景提出了改进方案，如 StructDepth[6] 利用曼哈顿世界模型引入结构先验，解耦动态区域的模型[7]通过分割与光流融合优化动态物体深度，但现有方法仍存在三点不足：一是缺乏对室内场景“边缘结构”这一关键先验的有效利用，难以解决边缘重叠导致的深度不连续问题；二是现有固定聚合方式无法自适应匹配室内多尺度特征上下文关联的问题；三是对室内手持拍摄旋转分量鲁棒性不足。为解决上述问题，本文构建了一种融合图像边缘相似性约束的室内自监督单目深度估计网络，主要贡献如下：1）提出了用于室内深度估计的图像边缘相似性损失函数，为深度估计提供形状先验，通过计算目标图像与预测深度图像的边缘相似度作为额外的自监督信号约束网络，改善了因室内场景边缘重叠、深度变化较大带来的模型恶化问题。2）提出了一个自适应特征聚合模块，在聚合高、低尺度特征的同时自适应地保持其上下文一致性来增强场景间的弱相关性，缩小语义差距。3）提出了一个旋转量优化模块，在位姿网络中加权融合原始主路径与其他路径的前3维向量来细化旋转分量，改善了因室内序列旋转分量较大导致预测位姿误差大的问题，提高了预测相对位姿的准确性进而提升了整个网络预测深度性能。实验结果表明，本文所提方法在 NYU Depth V2 与 ScanNet 两大室内深度数据集上，均取得了优于当前主流方法的深度预测精度。

1　自监督单目深度估计

现有的自监督单目深度估计模型根据训练数据类型可分为用立体图像对训练和用单目视频序列训练两种方式[8]。Garg 等[9] 将深度估计视为一种新的视图合成问题，提出输入左图像与合成右图像之间的最小光度损失，这种方法的监督信号来自输入立体图像对。Godard 等[10] 扩展了这一项工作，通过引入左右视差一致性损失实现了更高的预测精度。除了使用立体图像对外，监督信号也可以来自单目视频序列，Zhou 等[11] 通过训练独立的多视图姿态网络，实现了对连续两帧间姿态的有效估计，为了增强在处理遮挡和移动物体时的鲁棒性，还使用了可解释性预测网络来忽略违反视图合成假设的目标像素。为实现动态场景的有效建模，现有研究常采用多任务学习策略，例如联合光流估计[12] 与语义分割[13-14] 等任务，或引入不确定性估计[15] 等额外约束条件。Godard 等[16] 在不引入额外学习任务的情况下，通过简单地改进损失函数获得了有竞争力的结果，在 Monodepth2 中使用最小的重投影损失来缓解遮挡问题，并使用自动掩码过滤掉与相机速度相同的运动物体。此外，通过联合使用单目视频和额外的语义信息来学习深度的框架也在文献 [17] 中进行了研究，其他的一些工作被设计用于在具有挑战性的环境中处理自监督单目深度估计，例如

室内环境[18-19]和夜间环境[20]。Guo 等[21]探索了通过学习相对位姿信息监督多针深度学习来解决室内场景低纹理区域问题的方案。Cheng 等[22]为解决物体边界深度估计不一致的问题, 将语义约束纳入深度骨干网络中, 并协同位姿估计, 明确了深度边界。Ye 等[23]提出更加精确的重建约束损失函数, 来鼓励单目深度估计网络从多帧深度网络中获取有利信息。与本文最相关的 StructDepth[6]充分挖掘曼哈顿世界模型所蕴含的室内场景结构特性, 并将其应用于自监督单目深度估计任务中, 取得优异的性能表现。但是室内场景存在深度分布不一、变化较大、边缘重叠严重等复杂特性, 使得光度一致性损失无法很好地用来监督网络训练, 并且室内视频序列的旋转分量较大也会导致预测的位姿出现较大误差, 从而造成预测的深度信息边缘模糊、细节丢失严重, 预测精度太差。

2　算法与网络结构

2.1　训练网络结构

多数单目深度估计方法基于运动重构原理, 通过卷积神经网络实现深度网络与位姿网络的联合训练。该框架以视频序列中的连续多帧 RGB (red, green, blue)图像为输入, 由深度网络生成目标图像的深度预测图, 同时利用目标图像与相邻帧图像通过位姿网络估计相机运动的变换矩阵。基于上述两组输出构建重投影图像, 并将重投影误差与其他自监督损失项共同构成总损失, 通过反向传播迭代优化网络参数[24]。当前主流方法多采用编码器−解码器结构的 U-Net 架构, 常以 ResNet(residual network)坐标[25]、VGG(visual geometry group)坐标[26]、DenseNet[27]等作为骨干编码器。这些网络框架一定程度上实现了预测图像的深度功能, 但是在训练过程中对室内场景中重要的信息理解能力不够, 造成有用信息丢失或者出现较大误差, 使得网络预测的深度图像存在严重的场景结构边缘模糊、细节信息丢失等问题。本文为解决上述问题, 在 StructDepth[6]网络基础上提出了一种基于图像边缘相似性的室内自监督单目深度估计网络模型。该模型需要联合深度估计网络和位姿网络同时进行训练, 如图 1 所示, 本文采用单目视频序列作为输入的训练方式。

自监督深度估计是通过训练网络从另一个图像的视点预测目标图像的外观, 与文献 [11] 相似, 将该学习问题视为一个新的视图合成问题。给定目标图像和另一视角源图像, 利用预测深度图作为桥接变量对图像合成过程进行训练和约束, 训练网络既需要预测目标图像的深度图像, 也需要估计一对目标图像和源图像之间的相对姿态。因此光度重投影损失可以构造为

$$L_{\mathrm{pe}}=\sum_{\mathrm{s}}\rho\left(\boldsymbol{I}_{\mathrm{t}},\boldsymbol{I}_{\mathrm{s}\to\mathrm{t}}\right)$$

$$\boldsymbol{I}_{\mathrm{s}\to\mathrm{t}}=I_{\mathrm{s}}\left\langle\mathrm{proj}\left(\boldsymbol{D}_{\mathrm{t}},\boldsymbol{T}_{\mathrm{t}\to\mathrm{s}},\boldsymbol{K}\right)\right\rangle$$

式中ρ为光度重建误差, 它是L_1损失和图像结构化相似度 (structural similarity index measure, SSIM) 损失的加权组合, 定义为

$$\rho\left(\boldsymbol{I}_{\mathrm{t}},\boldsymbol{I}_{\mathrm{s}\to\mathrm{t}}\right)=\frac{\alpha}{2}\left(1-\mathrm{SSIM}\left(\boldsymbol{I}_{\mathrm{t}},\boldsymbol{I}_{\mathrm{s}\to\mathrm{t}}\right)\right)+(1-\alpha)\left\|\boldsymbol{I}_{\mathrm{t}},\boldsymbol{I}_{\mathrm{s}\to\mathrm{t}}\right\|_1$$

$\boldsymbol{I}_{\mathrm{s}\to\mathrm{t}}$是根据目标图像的深度将源图像扭曲到目标图像坐标帧, 本文将α设为 0.85, $\langle\cdot\rangle$表示中的局部次可微的双线性采样算子[11], proj(·)是将目标图像像素坐标p_{t}映射到源图像像素坐标p_{s}的变换函数, 计算公式为

$$p_{\mathrm{s}}\sim\boldsymbol{K}\boldsymbol{T}_{\mathrm{t}\to\mathrm{s}}\boldsymbol{D}_{\mathrm{t}}\left(p_{\mathrm{t}}\right)\boldsymbol{K}^{-1}p_{\mathrm{t}}$$

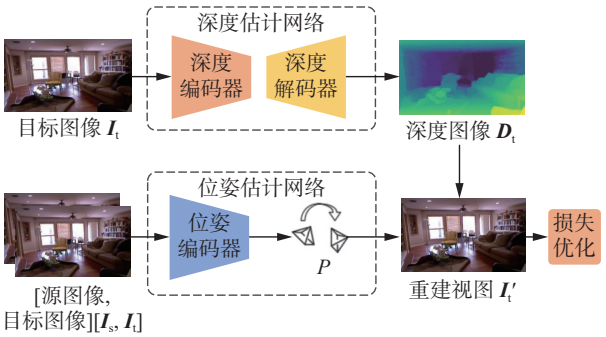


图 1　本文训练网络架构

Fig. 1　Training network architecture of this paper

假设所有图像的相机内参矩阵 $\boldsymbol{K}$ 相同, 当在室内图像出现低纹理区域时, 逐像素平滑损失函数能在光度重投影损失衰弱的情况下对深度值进行约束, 具体为

$$L_{\mathrm{Smooth}}=\left|\partial_x d_{\mathrm{t}}^*\right|\mathrm{e}^{-\left|\partial_x I_{\mathrm{t}}\right|}+\left|\partial_y d_{\mathrm{t}}^*\right|\mathrm{e}^{-\left|\partial_y I_{\mathrm{t}}\right|}$$

式中$d_{\mathrm{t}}^*=d/\bar{d}_{\mathrm{t}}$表示平均归一化逆深度[28], 在训练过程中, 本文采用自动掩码来处理静态像素。此外, 本文仍采用 StructDepth[6]中的法向量损失函数L_{nrom}和共平面损失约束L_{plane}, 因此这些损失函数可表式为

$$L=L_{\mathrm{pe}}+\lambda_1 L_{\mathrm{smooth}}+\lambda_2 L_{\mathrm{nrom}}+\lambda_3 L_{\mathrm{plane}}$$

根据实验效果, 超参数λ_1、λ_2、λ_3为权重参数, 分别设置为 0.001、0.05、0.1。

2.2　图像边缘相似性损失

计算机视觉领域中, 进行图像边缘检测主要目的是识别图像中像素值显著变化位置如场景深度不连续、物体颜色属性变化、场景照明的变化区域, 这恰好与室内场景存在着目标物体多、遮

挡重叠多以及亮暗区域交互多等复杂特点不谋而合。相较于室外图像，室内图像边缘更丰富，这为边缘结构相似性计算提供可靠数据支撑。此外，相比目前主流的图像相似性算法，计算室内图像边缘结构相似性能够在剔除不重要的区域信息、大幅度地减少计算数据量的同时保留图像重要的结构属性。本文提出的室内图像边缘相似性损失函数正是受到室内图像边缘检测启发而来，在深度估计网络中，如图 1 其预测的深度图像 $\boldsymbol{D}_{\mathrm{t}}$，能够清晰展现出物理三维世界的边缘结构，即使如图 2 所示，墙上相框与墙的深度差距较小，传统的深度估计网络预测两者深度时可能判定其一致，但通过引入边缘相似度来对比监督，能够增强深度网络对细节的深度预测。一般地，训练网络预测性能越好，其预测的深度图像在图像边缘结构方面应当与目标图像边缘结构越相似，应用上述原理来弥补深度估计网络由于下采样导致关键边缘信息的丢失问题，因此，计算预测深度图与目标图的边缘结构相似度，可作为有效的额外自监督约束信号作用于网络训练过程。训练阶段，网络通过前向传播计算图像边缘相似性损失值，再基于该损失误差反向传播，迭代更新网络所有权重参数，直至损失收敛至最小值，即两幅图像边缘相似度稳定收敛，通过图像边缘相似性损失函数的监督先验指导，有效地提高了训练网络预测室内图像深度的能力，进一步地提升了预测深度的精度。

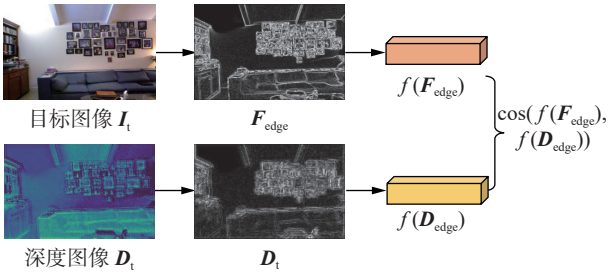


图 2　图像边缘相似性计算步骤

Fig. 2　Steps for calculating image edge similarity

图像边缘检测依赖于特定的检测算子，主要分为两类，一类是基于一阶微分算法的边缘检测算子，如 Roberts 和 Sobel 算子，它们通过将图像与微分滤波器卷积来识别图像边缘信息。另一类则是基于二阶微分算法的边缘检测算子，如 Laplacian[29] 和 Canny[30] 算子。本文使用 Canny 算子对室内图像进行边缘检测。如图 2 所示，图像边缘相似性损失计算步骤：1) 利用 Canny 算子分别检测出目标图像与预测深度图像的边缘特征信息图像 $\boldsymbol{F}_{\mathrm{edge}}$、$\boldsymbol{D}_{\mathrm{edge}}$；2) 将二维边缘信息图分别转换并拉伸为一维特征向量块 $f(\boldsymbol{F}_{\mathrm{edge}})$、$f(\boldsymbol{D}_{\mathrm{edge}})$；3) 计算两个特征向量块的余弦相似度。基于上述步骤，图像边缘相似性损失函数的计算公式为

$$L_{\mathrm{edge}} = 1 - \cos(f(\boldsymbol{F}_{\mathrm{edge}}), f(\boldsymbol{D}_{\mathrm{edge}}))$$

式中 $f(\cdot)$ 为将多维特征向量转为一维特征向量，为计算余弦相似度。因此本文总损失函数为

$$L_{\mathrm{all}} = L_{\mathrm{pe}} + \lambda_1 L_{\mathrm{smooth}} + \lambda_2 L_{\mathrm{nrom}} + \lambda_3 L_{\mathrm{plane}} + \lambda_4 L_{\mathrm{edge}}$$

根据实验效果，这里的超参数 λ_4 设置为 0.04。

2.3　自适应特征聚合的深度估计网络

本文模型的深度估计网络在 StructDepth[6] 网络模型基础上以 ResNet50 网络作为深度估计网络编码器，将目标图像 $\boldsymbol{I}_{\mathrm{t}}$ 作为输入，逐层下采样进行特征提取输出特征图像，基于自适应特征聚合的解码器将编码器中提取的多尺度特征逐层上采样，恢复至目标图像尺寸大小，最后通过 Sigmoid 激活函数映射处理输出预测估计的深度图像 $\boldsymbol{D}_{\mathrm{t}}$，如图 3 所示。图中 Layer1 ~ Layer5 表示编、解码器对应层级，$\boldsymbol{F}_1 \sim \boldsymbol{F}_5$、$\boldsymbol{F}'_1 \sim \boldsymbol{F}'_5$ 分别表示编、解码器各层输出特征图像，$\boldsymbol{C}_1 \sim \boldsymbol{C}_4$ 表示通过跳跃连接方式得到的编码器各层输出特征图像，作为自适应特征聚合 (adaptive feature aggregation, ADFA) 模块的输入之一，图中虚线表示编码器到解码器间的跳跃连接。本文模型的编码器网络部分对室内 RGB 图像 $\boldsymbol{I}_{\mathrm{t}} \in \mathbf{R}^{3\times H\times W}$ 逐层进行下采样操作提取不同尺度特征即各层级特征图像 $\boldsymbol{F}_1 \sim \boldsymbol{F}_5$。下采样过程中，各层级特征图像通道数由低到高变化分别为 64、256、512、1 024、2 056，其分辨率由高到低的变化分别为 $H/2\times W/2$、$H/4\times W/4$、$H/8\times W/8$、$H/16\times W/16$、$H/32\times W/32$。这种传统的编码器结构，随着编码网络层数的加深，提取的特征图像逐渐抽象化，其室内复杂场景中的几何空间的关联性愈来愈弱，丢失的特征细节信息逐渐变多，使得输出的特征图像表征能力降低，尽管传统的跳跃连接方法试图通过通道拼接或直接像素叠加逐渐将多尺度特征图像进行聚合来加强特征细节，一些现有研究已经一定程度上证明了其有效性，但它们忽略了室内多尺度特征图像间对应的区域应该包含相似场景的上下文信息这一重要特性，这极大地限制了深度网络预测室内图像深度的精度。

为了使网络对室内多尺度特征图像间的上下文信息一致性给予更多的关注，本文设计了如图 4 所示的自适应特征聚合模块，用于同时聚合一对低尺度和高尺度特征图像并自适应地保持其上下文一致性关系。ADFA 模块将解码器层级通过标

准双线性插值上采样以及反卷积操作得到的特征图像$\boldsymbol{F}'_{i+1}(1\leqslant i\leqslant 4)$与通过跳跃连接得到的上一层级编码器输出$\boldsymbol{C}_i(1\leqslant i\leqslant 4)$按照通道拼接，然后通过两个学习特征偏移映射分支来预测偏移量图$\Delta\boldsymbol{C}_i$、$\Delta\boldsymbol{F}'_{i+1}$，分别用于细化特征图像$\boldsymbol{F}'_{i+1}$和特征图像$\boldsymbol{C}_i$，细化函数$R$通过双线性插值进行细化，即通过偏移量图$\Delta(\boldsymbol{p})$在$\boldsymbol{p}=[x,y]^{\mathrm{T}}$中的位置处生成细化特征$\tilde{\boldsymbol{F}}(p)$，计算公式为

$$\tilde{\boldsymbol{F}}(\boldsymbol{p})=R(\boldsymbol{F},\Delta(\boldsymbol{p}))=\langle\boldsymbol{F}(\boldsymbol{p}+\Delta(\boldsymbol{p}))\rangle$$

式中：$\boldsymbol{p}$表示特征图中像素点，$\langle\cdot\rangle$表示双线性插值操作。因此图 4 中两个细化特征$\tilde{\boldsymbol{F}}'_{i+1}$、$\tilde{\boldsymbol{C}}_i$公式为

$$\tilde{\boldsymbol{F}}'_{i+1}=R\left(\boldsymbol{F}_{i+1},\Delta\boldsymbol{F}'_{i+1}\right)$$

$$\tilde{\boldsymbol{C}}_i=R(\boldsymbol{C}_i,\Delta\boldsymbol{C}_i)$$

最后将两个细化特征进行聚合得到$\boldsymbol{F}''_{i+1}$，公式为

$$\boldsymbol{F}''_{i+1}=\mathrm{conv}\left(\tilde{\boldsymbol{C}}_i\,\dot{+}\,\tilde{\boldsymbol{F}}'_{i+1}\right)$$

式中$\dot{+}$表示直接像素叠加。将得到聚合特征$\boldsymbol{F}''_{i+1}$作为解码器上一层级的输入，依次重复以上 4 次聚合处理，最后将解码器第一层级输出特征图像用 Sigmoid 函数映射处理得到预测深度图像$\boldsymbol{D}_{\mathrm{t}}$。本文通过解码器网络在解码恢复特征图像的同时自适应地聚合高低尺度特征图像，有效地保持了特征图像之间的上下文一致性，缩小了上下文特征的语义差距，使网络能够预测出边缘更清晰、平滑的深度信息。

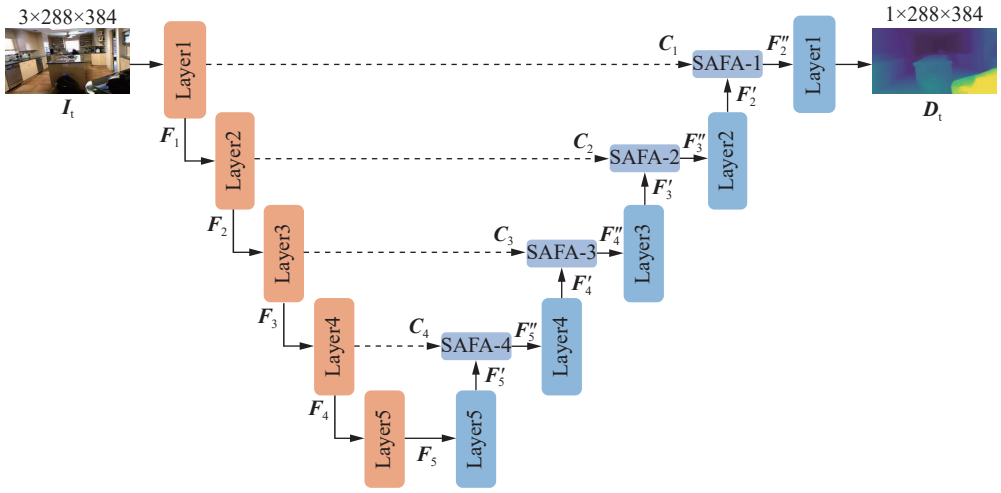


图 3　本文深度估计网络

Fig. 3　Depth estimation network in this paper

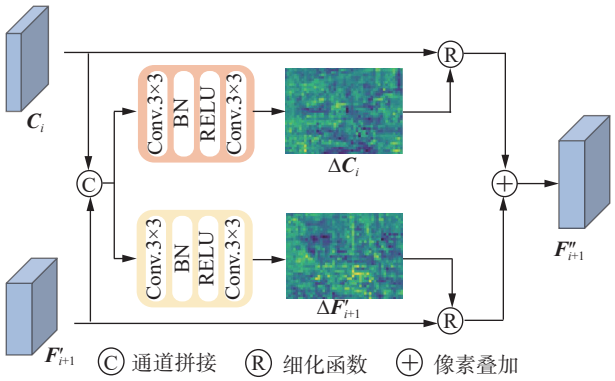


图 4　自适应特征聚合模块

Fig. 4　Adaptive feature aggregation module

2.4　旋转量优化的位姿估计网络

在单目深度估计中，位姿估计网络用来预测目标图像与源图像的相对位姿，而相对位姿的准确性影响着光度重投影损失函数是否在网络中监督有效，因为不准确的位姿可能导致目标图像和源图像像素之间的对应错误，从而导致深度预测出错，因此，它的性能将极大地影响着自监督网络模型的预测效果。现有的模型大多采用独立的位姿网络来估计两幅图像之间的 6 个自由度姿态。在室外场景，如 KITTI 数据集驾驶场景，其相机位姿非常简单，因为汽车大多向前移动，平移较大，但旋转较小，这意味着室外场景的姿势估计通常不那么具有挑战性。相比之下，在室内场景中，这些序列通常是用手持设备如 Kinect 记录的，不可避免地要经历频繁地旋转，包含着难以预测的复杂旋转运动，使得位姿网络预测准确的相机位姿变得更加困难，这不利于室内场景网络模型的自监督训练。基于上述问题，本文设计了旋转量优化 (rotation optimization, ROOP) 模块。如图 5 所示的传统位姿网络和改进位姿网络架构，都采用了编−解码器形式，编码器部分采用 ResNet18 网络对输入通道数为 3 的相邻帧图像进行 4 次下采样操作，最终输出 512 通道的大小的特征图像。解码器部分首先将得到的两个特征图像进行拼接，通过多层卷积运算与平均池化操作最后得到一个 6 维的特征向量即是相邻帧间图像

的 6 自由度位姿变换，其中前面欧拉角形式的 3 维特征代表旋转，后面 3 维特征代表平移。在具体应用时，需将这个向量转换成位姿矩阵，即

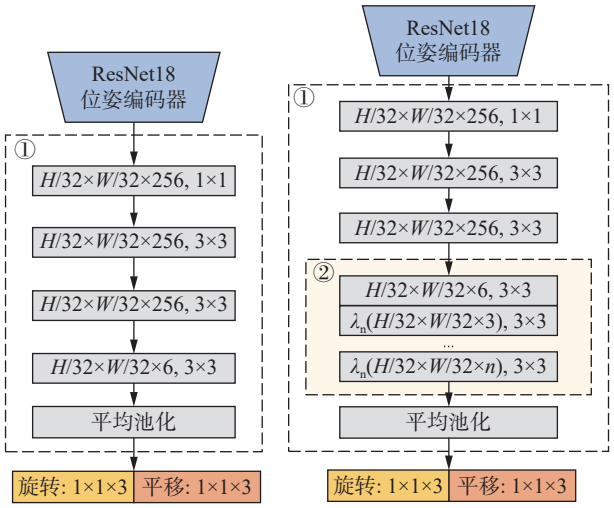
$$\boldsymbol{v}=(r,t)^{\mathrm{T}}\in\mathbf{R}^6,\boldsymbol{r}\in\mathbf{R}^3,\boldsymbol{t}\in\mathbf{R}^3$$

$$\boldsymbol{R}=\begin{bmatrix}1&0&0\\0&\cos r_0&-\sin r_0\\0&\sin r_0&\cos r_0\end{bmatrix}\begin{bmatrix}\cos r_1&0&\sin r_1\\0&1&0\\\sin r_1&0&\cos r_1\end{bmatrix}\begin{bmatrix}\cos r_2&-\sin r_2&0\\\sin r_2&\cos r_2&0\\0&0&1\end{bmatrix}$$

$$\boldsymbol{T}=\begin{bmatrix}\boldsymbol{R}&\boldsymbol{t}\\0&1\end{bmatrix}\in\mathbf{R}^{4\times4}$$

式中矩阵中的 r_0、r_1、r_2 表示旋转分量的 3 个欧拉角，即

$$\boldsymbol{r}=[r_0\quad r_1\quad r_2]^{\mathrm{T}}$$



(a) 原始位姿网络　　(b) 旋转量优化位姿网络

图 5　传统位姿网络与改进位姿网络架构比较

Fig. 5　Comparison of traditional pose network and improved pose network architecture

为了能够改善因室内旋转分量较大、难以预测而带来的相对位姿预测不精确的问题，本文改进了传统位姿解码器网络单路径预测 6 维向量方式，采取多路径融合细化方式预测 6 维向量，其中平移分量不融合细化。如图 5 (b) 改进位姿网络中②区的旋转量优化模块部分所示，在位姿解码器第 4 层中，将原始单层卷积层改成多层卷积层，原始主路径卷积层仍预测 6 维向量，其他路径卷积层分别预测 5 维向量、4 维向量等，最后将所有的前 3 维向量进行加权融合细化得到输出，计算公式为

$$\boldsymbol{r}=\lambda_3\boldsymbol{R}_3+\lambda_4\boldsymbol{R}_4+\lambda_5\boldsymbol{R}_5+\lambda_6\boldsymbol{R}_6+\lambda_7\boldsymbol{R}_7+\lambda_8\boldsymbol{R}_8$$

式中：超参数 λ 表示权重系数，$\boldsymbol{R}$ 表示预测多维向量的前 3 维向量，下方数字表示当前路径输出的向量维度。考虑到如果原始主路径的旋转分量与其他路径前 3 维向量直接相加，其他路径分量会干扰预测的旋转分量，因而本文使用较小的超参数以较少对原始主路径旋转分量的削弱，从而在提升旋转分量准确性的同时降低了对主路径数据的影响，根据实验效果，最终本文将超参数 λ_4 设为 0.4，λ_6 设为 1，其余设为 0。详见 3.5.1 节多种优化形式消融实验，通过旋转量优化模块，本文模型更准确地预测了相对位姿，提升了网络预测精度。

3　实验结果与分析

3.1　数据集

本文选取 NYU Depth V2[31] 和 ScanNet[32] 两种室内数据集进行训练与测试，其中，NYU Depth V2 数据集作为当前应用广泛的 RGB-D（RGB-Depth）数据集之一，由手持 RGB-D 相机以 640×480 分辨率相机拍摄的各种室内场景的视频序列组成，该数据集包含 464 个由 Kinect 传感器采集的室内场景，本文严格遵循官方划分方案拆分训练集与测试集，其中 249 个场景用于模型训练，剩余 215 个场景作为测试集。ScanNet 数据集则通过搭载于 iPad 的深度相机采集室内场景数据，包含 1 513 个视频序列，大约 2.5 万张图像，为了评估本文网络架构在 NYU Depth V2 上训练的深度模型的泛化性能，本文遵循官方发布的 ScanNet 数据集分割方法如 StructDepth[6] 进行实验验证。在训练和测试期间，图像的大小均被调整为 384×288。

3.2　实验细节

本文网络架构使用 Linux 系统 Ubuntu20.04.2 版本，所有实验均基于主流深度学习框架 PyTorch 1.8.1 搭建实现，训练与测试过程均在单卡 NVIDIA GeForce RTX 3090 GPU（显存 24 GB）上完成。训练阶段，将批次大小（Batch Size）设置为 16，初始学习率配置为 0.001。本文采用多步学习率降低策略，在第 26 个 Epoch 与第 36 个 Epoch，学习率乘以 0.1 进行衰减。网络采用 Adam 算法优化器进行参数优化，权重衰减参数分别设置为 β_1=0.9，β_2=0.999，共训练 50 个 Epoch。对于输入图像本文在边界裁剪 16 个像素后，缩放至 288×384 进行训练。为避免过拟合，本文对输入图像进行了随机翻转和颜色增强。同时，本文将深度限制为 10 m，并使用中位缩放策略避免单眼深度估计的尺度歧义。

3.3　性能评价指标

为了与先前的工作保持一致的测试基准从而准确地评估本文网络的效果，本文采用 Eigen 等[4]

提出的度量标准作为深度估计精度的评价指标, 指标包括绝对相对误差 (absolute relative error, Abs Rel)、均方根误差 (root mean square error, RMSE)、对数空间下的均方根误差 (log-root-mean-square error, RMSE Log) 以及不同阈值的精确率 (Accuracies), 定义公式为

$$I_{\text{Abs Rel}}=\sqrt{\frac{1}{|N|}\sum_{p\in N}\frac{\left|d_p-d_p^*\right|}{d_p^*}}$$

$$I_{\text{sq Rel}}=\frac{1}{|N|}\sum_{p\in N}\frac{\|d_p-d_p^*\|^2}{d_p^*}$$

$$D_{\text{RMSE}}=\sqrt{\frac{1}{N}\sum_{p\in N}\|d_p-d_p^*\|^2}$$

$$D_{\text{RMSE Log}}=\sqrt{\frac{1}{|N|}\sum_{p\in N}\|\log(d_p)-\log(d_p^*)\|^2}$$

$$A_{\text{Accuracies}}=\max\left(\frac{d_p}{d_p^*},\frac{d_p^*}{d_p}\right)=\sigma<\text{thr}$$

式中d_p和d_p^*分别代表像素 p 的预测深度与真实深度, N为含真实深度标注的像素总数, thr为阈值, 通常取 1.25、1.25^2、1.25^3。对图像中各像素, 先计算预测深度与真实深度的比值并取最大值, 记为σ。再统计满足σ小于 thr 的像素占比, 即为对应阈值下的精确率, 当这个比值越接近于 1, 网络性能越好。

3.4　实验分析

3.4.1　NYU Depth V2 数据集实验结果

为了验证本文所提网络模型的优越性, 按照 3.1、3.2 节进行实验配置, 在 NYU Depth V2 室内数据集上进行实验测试, 并与现有一些先进的室内单目深度估计网络模型进行实验指标对比, 同时按照第 3 节所述, 本文在 StructDepth[6] 基础上进行改进, 将特征提取的编码器网络 ResNet18 换成 ResNet50, 并以此为基准模型 (Baseline), 将其表述为 StructDepth_50。实验定量结果如表 1 所示, 表中加粗的数值为最优实验结果。

表 1　本文网络模型与现有主要模型在 NYU Depth V2 数据集上的实验结果对比

Table 1　Comparison of experimental results between the proposed network and existing main models on the NYU Depth V2 dataset

模型	监督方式	误差指标(越低越好)			预测精确率(越高越好)		
		RMSE	Abs Rel	RMSE Log10	$\sigma<1.25$	$\sigma<1.25^2$	$\sigma<1.25^3$
DORN[33]	有监督	0.509	0.115	0.051	0.828	0.965	0.992
Hu 等[34]	有监督	0.530	0.115	0.050	0.866	0.975	0.993
Yin 等[35]	有监督	0.416	0.108	0.108	0.875	0.976	0.994
AdaBins[36]	有监督	0.364	0.103	0.044	0.903	0.984	0.997
Niklaus 等[37]	有监督	0.300	0.080	0.030	0.940	0.990	1.000
MovingIndoor[38]	自监督	0.712	0.208	0.086	0.674	0.900	0.968
TrainFlow[39]	自监督	0.686	0.208	0.086	0.701	0.912	0.978
Monodepth2[16]	自监督	0.600	0.161	0.068	0.771	0.948	0.987
SC-Depth[40]	自监督	0.608	0.159	0.068	0.772	0.939	0.982
P^2Net[41]	自监督	0.561	0.150	0.064	0.796	0.948	0.986
P^2Net(5framesPP)[41]	自监督	0.553	0.147	0.062	0.801	0.951	0.987
Bian 等[42]	自监督	0.536	0.147	0.062	0.804	0.950	0.986
PLNet(5frames)[43]	自监督	0.540	0.144	0.061	0.807	0.957	0.990
StructDepth[6]	自监督	0.540	0.142	0.060	0.813	0.954	0.988
SC-DepthV2[44]	自监督	0.532	0.138	—	0.820	0.956	0.989
TSD-Depth[45]	自监督	0.533	0.139	0.059	0.823	0.956	0.989
MonoIndoor[5]	自监督	0.526	0.134	—	0.823	0.958	0.989
SPDepth+PP[46]	自监督	0.579	0.157	0.066	0.781	0.947	0.986
F^2Depth+PP[47]	自监督	0.569	0.153	0.065	0.787	0.950	0.987
StructDepth_50	自监督	0.528	0.137	0.058	0.823	0.958	0.989
本文	自监督	**0.525**	**0.133**	**0.057**	**0.829**	**0.959**	**0.990**

注: 粗体表示最优值。

从表 1 可见，在目前主流的深度估计评价指标度量上，本文模型均优于现有先进的室内自监督单目深度估计模型。相比于基线模型 StructDepth_50，RMSE 指标下降了 0.3%，Abs Rel 指标下降了 0.4%，RMSE Log10 指标下降 0.1%，尤其是在衡量室内自监督单目深度估计网络预测性能最重要的指标精确度 σ<1.25 上提高了 0.6%，充分说明了本文网络模型拥有更优的预测性能。为了更加直观地说明本文的有效性，将预测的深度进行可视化处理，如图 6 所示，本文与 StructDepth[6]、Baseline 模型进行对比，其中 Ground truth 为深度真值。可以看出，本文方法预测的深度图像更接近真实深度，物体边缘更完整，同深度区域更平滑，成功改善了现有方法中边缘轮廓模糊、细节特征缺失的缺陷，使深度预测精度得到明显提升。

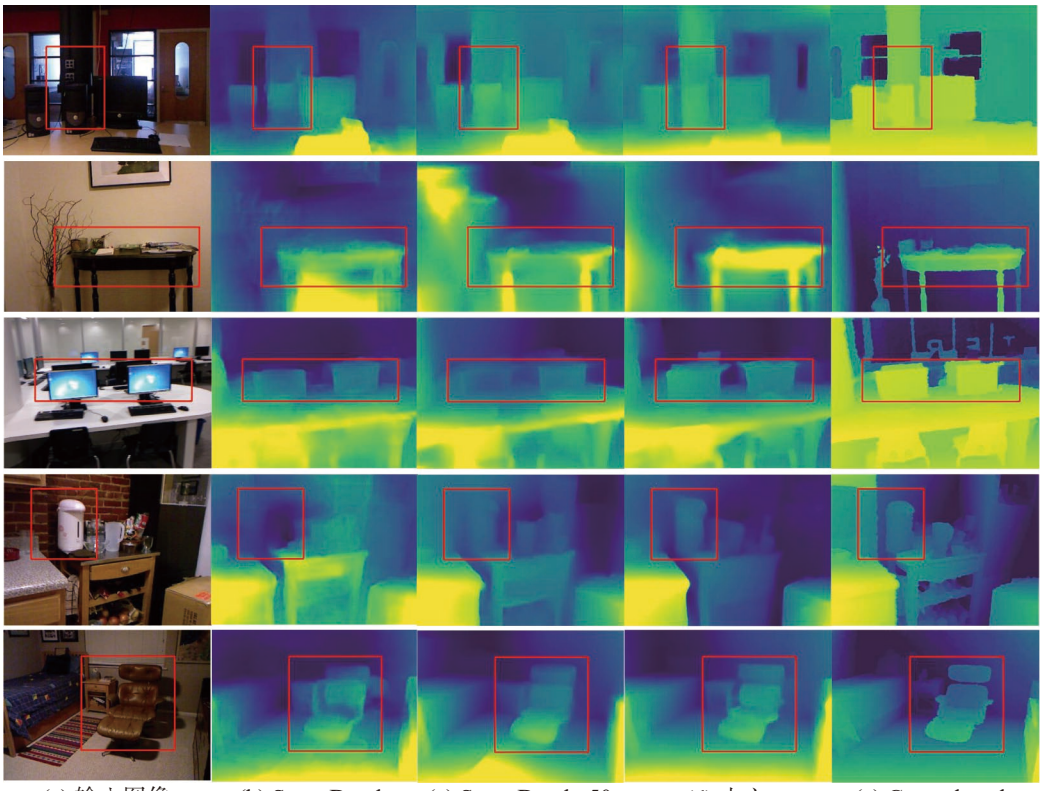


图 6　本文网络模型与现有主要模型在 NYU Depth V2 数据集上的预测深度图对比

Fig. 6　Comparison of predicted depth maps between the proposed network model and existing state-of-the-art models on the NYU Depth V2 dataset

3.4.2　ScanNet 数据集实验结果

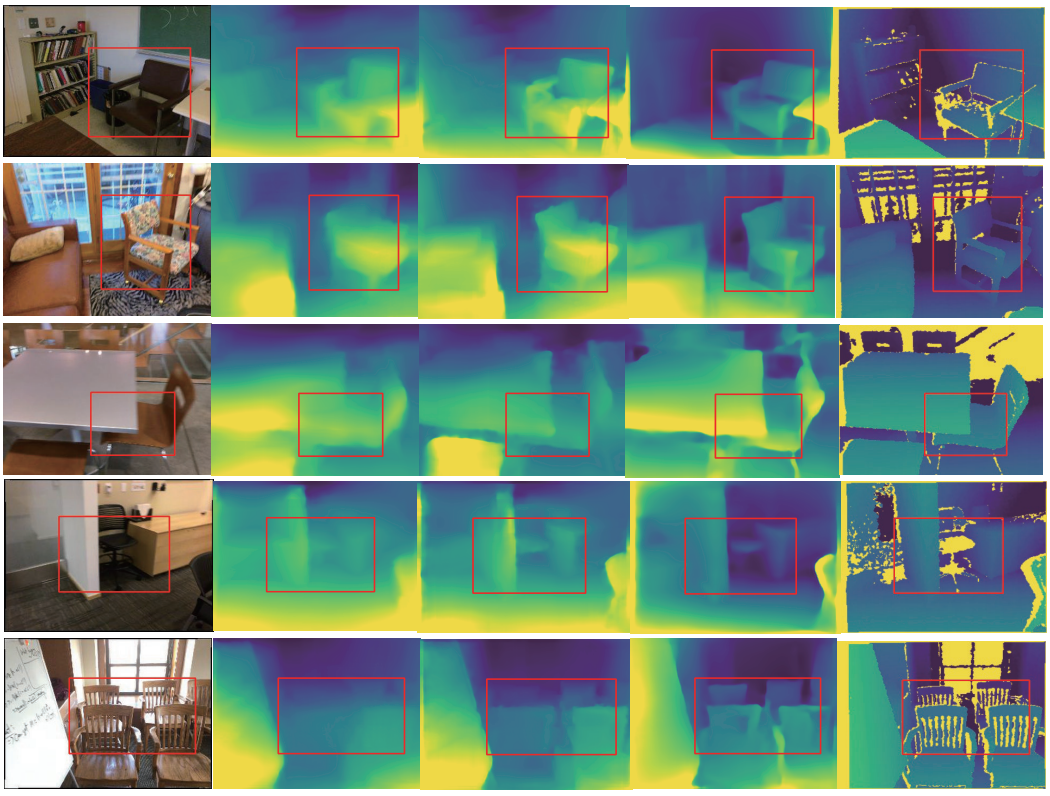
为了验证本文模型的迁移泛化能力，将 NYU Depth V2 数据集上训练的网络模型推广到其他室内数据集进行评估。同样地，按照 NYU Depth V2 数据集实验设定在 ScanNet 室内数据集上进行实验验证，并与现有一些先进的室内单目深度估计网络模型进行实验指标对比，实验定量结果如表 2 所示，表中加粗的数值为最优实验结果。从表 2 中实验结果可见，按照 3.3 节所述的深度估计评价指标度量，本文模型优于现有的一些先进的室内自监督单目深度估计模型。相比于 Baseline 模型，RMSE 指标下降了 0.9%，Abs Rel 指标下降了 0.2%，RMSE Log10 指标下降 0.2%，尤其是在精确度 σ<1.25 上，实验指标提高了 1.0%。将网络预测的深度进行更直观地可视化处理，如图 7 所示，本文与 StructDepth[6]、Baseline 模型进行对比，其中 Ground truth 为深度真值，可以看到本文预测出的深度图像相比 Baseline 模型，其预测的物体边缘更加完整、深度相同的区域更加平滑，同样有效地改善了场景结构边缘模糊、细节丢失严重等问题，提高了预测的精度。尤其是在图 7 红色框中，本文预测的椅子、桌角、墙角以及栅栏等外形轮廓深度相比 Baseline 模型更加准确、平滑，充分说明了本文网络模型不仅具有优越的性能而且也具有良好的迁移泛化能力。

表 2　本文网络模型与现有主要模型在 ScanNet 数据集上的实验结果对比

Table 2　Comparison of experimental results between the proposed network model and existing state-of-the-art models on the ScanNet dataset

模型	监督方式	误差指标（越低越好）			预测精确率（越高越好）		
		RMSE	Abs Rel	RMSE Log10	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
MovingIndoor[38]	自监督	0.483	0.212	0.088	0.650	0.905	0.976
Monodepth2[16]	自监督	0.451	0.191	0.080	0.693	0.926	0.983
P^{2}Net[41]	自监督	0.420	0.175	0.074	0.740	0.932	0.982
P^{2}Net-finetune[41]	自监督	0.412	0.172	0.073	0.743	0.935	0.984
StructDepth[6]	自监督	0.400	0.165	0.070	0.754	0.939	0.985
SC-DepthV2[44]	自监督	**0.361**	0.159	**0.066**	**0.781**	**0.947**	0.987
TSD-Depth[45]	自监督	0.406	0.172	0.071	0.754	0.939	0.985
IFMNet[48]	自监督	0.402	0.170	0.071	0.758	0.940	**0.989**
BiGeoDepth[49]	自监督	0.399	0.164	0.070	0.758	0.943	0.986
GeoDepth[50]	自监督	0.387	0.161	—	0.769	0.946	0.987
StructDepth_50	自监督	0.388	0.159	0.068	0.770	**0.947**	0.987
本文	自监督	0.379	**0.157**	**0.066**	0.780	**0.947**	0.988

注：粗体表示最优值。



(a) 输入图像　(b) StructDepth　(c) StructDepth_50　(d) 本文　(e) Ground truth

图 7　本文网络模型与现有主要模型在 ScanNet 数据集上的预测深度图对比

Fig. 7　Comparison of predicted depth maps between the proposed network model and state-of-the-art models on the ScanNet dataset

3.5　消融实验

3.5.1　旋转量优化消融实验

为了验证位姿旋转量优化模块对室内自监督单目深度估计网络模型性能提升的效果，本文以 StructDepth_50 作为 Baseline 模型对比实验结果，在 NYU Depth V2 数据集上针对位姿网络第 4 卷积层中添加旋转量优化模块时不同加权融合形式进行了消融实验。如表 3 所示，表中单条、双条

以及 3 条融合形式分别表示在融合优化旋转量时，在原始位姿网络第 4 层预测多维向量基础上添加 1 条、2 条、3 条加权融合路径，“路径”列中数值表示添加预测维数所对应的加权融合路径，原始主路径为预测 6 维向量，表中指标数值加粗为实验最优结果。

表 3　NYU Depth V2 数据集上多种旋转量优化形式消融实验

Table 3　Ablation study of various rotation optimization strategies on the NYU Depth v2 dataset

融合形式	路径	误差指标（越低越好）			预测精确率（越高越好）		
		RMSE	Abs Rel	RMSE Log10	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Baseline	/	0.528	0.137	**0.058**	0.823	0.958	0.989
单条路径	4	**0.526**	**0.136**	**0.058**	**0.826**	**0.959**	**0.990**
	5	0.527	0.138	**0.058**	0.825	**0.959**	0.989
	7	0.527	0.137	**0.058**	0.824	**0.959**	0.989
	8	0.529	0.139	0.059	0.823	**0.959**	**0.990**
双条路径	4+5	0.527	0.137	**0.058**	0.825	**0.959**	0.989
	4+7	0.527	**0.136**	**0.058**	**0.826**	**0.959**	0.989
	5+7	0.528	0.138	**0.058**	0.823	**0.959**	0.989
	5+8	0.529	0.139	**0.058**	0.822	**0.959**	0.989
三条路径	4+5+7	0.528	0.137	**0.058**	0.825	**0.959**	**0.990**
	4+5+8	0.528	0.138	**0.058**	0.823	**0.959**	0.989
	5+7+8	0.529	0.139	0.059	0.822	**0.959**	0.989

注：粗体表示最优值，“/”表示未使用本文所提出的加权融合路径。

从表 3 中可以看出，一方面，合理地添加单条路径或者多条路径进行加权融合时，能够有效地优化位姿旋转量，从而提升网络的预测深度的性能，其中原始主路径融合路径“4”最能够提升网络的预测深度性能。另一方面，越远离原始主路径维数的路径在加权融合时其预测深度性能变得越差，越多路径进行加权融合时，也会导致预测深度性能变差，甚至不如 Baseline 模型，说明在路径加权融合时越远离原始主路径的路径前 3 维向量和越多的路径融合后的前 3 维向量越不能代表旋转分量，会逐渐干扰原始旋转分量，从而影响网络预测深度性能。表 3 中融合距离原始主路径的路径“4”反而比路径“5”更能优化位姿旋转分量，可能是因为太靠近原始主路径反而一定程度地干扰了优化后旋转量的正确性，因此本文选择只融合路径“4”进行来细化旋转分量。

3.5.2　创新模块消融实验

为了评估所提各个创新模块对室内自监督单目深度估计网络的有效性，本文在 NYU Depth V2 数据集上进行消融实验，以 StructDepth_50 作为 Baseline 模型对比实验结果。如表 4 所示，表中“EDGE” “ADFA” “ROOP”分别代表图像边缘相似性损失函数、自适应特征聚合模块以及旋转量优化模块，表中√代表在消融实验中使用此创新模块，×表示没有使用此创新模块，加粗的数值为最优实验结果。从表 4 中“模型 1、2、3”实验结果可以见到，在衡量室内自监督单目深度估计网络预测性能最重要的指标 上，相比于 Baseline 模型，使用图像边缘相似性损失函数能够提高 0.3%，使用自适应特征聚合模块能够提高 0.2%，使用位姿旋转量优化模块能够提高 0.3%，均能不同程度地提升网络预测深度的性能，有效提高预测深度的精确率。此外，从表 4 可见，不论是两两组合地使用本文创新模块还是同时使用 3 个创新模块都能优于单独使用时的实验结果，尤其是同时使用 3 个创新模块的实验结果表现最佳，在指标上提高了 0.6%。

表 4　NYU Depth V2 数据集上不同创新模块的消融实验

Table 4　Ablation study of different innovative modules on the NYU Depth V2 dataset

模型	EDGE	ADFA	ROOP	误差指标（越低越好）			预测精确率（越高越好）		
				RMSE	Abs Rel	RMSE Log10	$\sigma < 1.25$	$\sigma < 1.25^2$	$\sigma < 1.25^3$
Baseline	×	×	×	0.528	0.137	0.058	0.823	0.958	0.989
1	√	×	×	0.526	0.134	0.058	0.826	**0.959**	**0.990**

续表 4

模型	EDGE	ADFA	ROOP	误差指标(越低越好)			预测精确率(越高越好)		
				RMSE	Abs Rel	RMSE Log10	$\sigma<1.25$	$\sigma<1.25^2$	$\sigma<1.25^3$
2	×	√	×	0.527	0.137	0.058	0.825	**0.959**	0.989
3	×	×	√	0.526	0.136	0.058	0.826	**0.959**	**0.990**
4	√	√	×	0.526	0.136	0.058	0.827	**0.959**	**0.990**
5	√	×	√	**0.525**	0.134	**0.057**	0.828	**0.959**	**0.990**
6	×	√	√	0.527	0.136	0.058	0.826	**0.959**	0.989
7	√	√	√	**0.525**	**0.133**	**0.057**	**0.829**	**0.959**	**0.990**

注: 粗体表示最优值。

3.5.3 计算复杂度分析

为了更精确地评估本文模型的复杂度, 本文参照 Han 等[51] 设定, 对模型参数量(parameters, Params)、计算量(浮点运算次数, floating point operations, FLOPs)以及推理速度与其他方法进行对比, 结果见表 5。实验结果为在 NVIDIA 3090Ti GPU 上运行 100 次后取其平均值, 图像大小为 288×384。从表中可见, 本文为提升模型的表征能力在基线模型基础上增加了一定程度的参数量, 推理速度有所牺牲, 但整体与主流方法仍处于相近水平, 满足实时视频分析需求, 同时实现了最佳的深度估计性能。

表 5　不同方法参数量、计算量和推理速度对比

Table 5　comparison of the Params, FLOPs, and speed of different methods

模型	编码器		解码器		完整模型		速度/ms
	Params/10^6	FLOPs/10^9	Params/10^6	FLOPs/10^9	Params/10^6	FLOPs/10^9	
Monodepth2[16]	11.200	4.500	3.100	3.500	14.300	8.000	14.300
P^2Net[41]	11.690	**4.019**	3.151	3.201	14.841	7.220	**6.653**
SC-DepthV2[44]	11.690	**4.019**	3.153	3.215	14.843	7.234	7.696
TSD-Depth[45]	25.720	5.180	7.120	4.040	32.840	9.220	17.571
StructDepth[6]	**4.794**	6.146	**0.005**	**0.159**	**4.799**	**6.305**	17.481
本文	11.177	**4.019**	3.151	3.201	14.328	7.220	17.628

注: 粗体表示最优值。

4　结束语

针对现有室内自监督单目深度估计模型由于室内场景具有深度变化较大、边缘重叠严重等复杂特性, 并且室内单目视频序列具有比室外序列更大的旋转分量, 导致位姿网络预测旋转分量时存在较大误差等问题。本文提出了一种基于图像边缘相似性的室内自监督单目深度估计网络模型。该模型提出了图像边缘相似性损失函数, 将其作为额外的自监督信号约束网络, 改善了因室内场景边缘重叠严重带来的模型恶化问题; 提出自适应特征聚合模块, 在聚合高、低尺度特征的同时自适应地保持其上下文一致性来增强场景间的弱相关性, 缩小语义差距; 提出了旋转量优化模块, 在位姿网络中加权融合原始主路径与其他路径的前 3 维向量来细化旋转分量, 改善了因室内序列旋转分量较大导致预测位姿误差大的问题。实验结果表明, 本文所提出的网络模型相比 Baseline 模型很好地缩小了预测深度图像与深度真值之间的语义差距, 有效地提高了网络预测深度的性能, 使得预测的深度细节更多, 场景轮廓更加清晰、平滑, 并具有良好的泛化性能。下一步工作将考虑进一步提高深度估计网络的泛化能力并降低网络参数量, 提高网络预测深度的效率。

参考文献:

[1] 胡海洋, 陈超平, 高天沐, 等. 单/双目深度估计研究进展与应用综述[J]. 红外与激光工程, 2025, 54(7): 35−48.
HU Haiyang, CHEN Chaoping, GAO Tianmu, et al. Recent progress in research and applications of monocular and binocular depth estimation[J]. Infrared and laser engineering, 2025, 54(7): 35−48.

[2] 李乐, 张茂军, 熊志辉, 等. 基于内容理解的单幅静态街景图像深度估计[J]. 机器人, 2011, 33(2): 174−180.

LI Le, ZHANG Maojun, XIONG Zhihui, et al. Depth estimation from a single still image of street scene based on content understanding[J]. Robot, 2011, 33(2): 174−180.

[3] 张楠, 程德强, 寇旗旗, 等. 基于随机遮挡和多粒度特征融合的行人重识别[J]. 北京航空航天大学学报, 2023, 49(12): 3511−3519.
ZHANG Nan, CHENG Deqiang, KOU Qiqi, et al. Person re-identification based on random occlusion and multi-granularity feature fusion[J]. Journal of Beijing University of Aeronautics and Astronautics, 2023, 49(12): 3511−3519.

[4] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[J]. Advances in neural information processing systems, 2014, 27.

[5] JI Pan, LI Runze, BHANU B, et al. MonoIndoor: towards good practice of self-supervised monocular depth estimation for indoor environments[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2022: 12767−12776.

[6] LI Boying, HUANG Yuan, LIU Zeyu, et al. StructDepth: Leveraging the structural regularities for self-supervised indoor depth estimation[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2022: 12643−12653.

[7] ZHOU Kaichen, BIAN Jiawang, ZHENG Jianqing, et al. Manydepth2: motion-aware self-supervised monocular depth estimation in dynamic scenes[J]. IEEE robotics and automation letters, 2025, 10(7): 6704−6711.

[8] 程德强, 范舒铭, 钱建生, 等. 基于坐标感知注意的多帧自监督单目深度估计[J]. 北京航空航天大学学报, 2025, 51(7): 2218−2228.
CHENG Deqiang, FAN Shuming, QIAN Jiansheng, et al. Coordinate-aware attention-based multi-frame self-supervised monocular depth estimation[J]. Journal of Beijing University of Aeronautics and Astronautics, 2025, 51(7): 2218−2228.

[9] GARG R, B G V K, CARNEIRO G, et al. Unsupervised CNN for single view depth estimation: geometry to the rescue[C]//Computer Vision–ECCV 2016. Cham: Springer, 2016: 740−756.

[10] GODARD C, MAC AODHA O, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6602−6611.

[11] ZHOU Tinghui, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6612−6619.

[12] YIN Zhichao, SHI Jianping. GeoNet: unsupervised learning of dense depth, optical flow and camera pose[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1983−1992.

[13] CASSER V, PIRK S, MAHJOURIAN R, et al. Unsupervised monocular depth and ego-motion learning with structure and semantics[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach: IEEE, 2020: 381−388.

[14] 程德强, 徐帅, 韩成功, 等. 基于视觉注意的自监督单目深度估计[J]. 计算机辅助设计与图形学学报, 2024, 36(12): 1920−1931.
CHENG Deqiang, XU Shuai, HAN Chenggong, et al. Visual attention-based self-supervised monocular depth estimation[J]. Journal of computer-aided design & computer graphics, 2024, 36(12): 1920−1931.

[15] 柴国强, 薄祥仕, 刘海军, 等. 基于不确定性单目图像自监督场景深度估计[J]. 北京航空航天大学学报, 2024, 50(12): 3780−3787.
CHAI Guoqiang, BO Xiangshi, LIU Haijun, et al. Self-supervised scene depth estimation for monocular images based on uncertainty[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(12): 3780−3787.

[16] GODARD C, MAC AODHA O, FIRMAN M, et al. Digging into self-supervised monocular depth estimation[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 3827−3837.

[17] 宋霄罡, 胡浩越, 宁靖宇, 等. 联合语义分割的自监督单目深度估计方法[J]. 计算机研究与发展, 2024, 61(5): 1336−1347.
SONG Xiaogang, HU Haoyue, NING Jingyu, et al. Self-supervised monocular depth estimation method for joint semantic segmentation[J]. Journal of computer research and development, 2024, 61(5): 1336−1347.

[18] 沙浩, 刘越, 王涌天, 等. 基于二维图像和三维几何约束神经网络的单目室内深度估计方法[J]. 光学学报, 2022, 42(19): 47−57.
SHA Hao, LIU Yue, WANG Yongtian, et al. Monocular indoor depth estimation method based on neural networks with constraints on two-dimensional images and three-dimensional geometry[J]. Acta optica sinica, 2022, 42(19): 47−57.

[19] 程德强, 张华强, 寇旗旗, 等. 基于层级特征融合的室内自监督单目深度估计[J]. 光学精密工程, 2023, 31(20): 2993−3009.
CHENG Deqiang, ZHANG Huaqiang, KOU Qiqi, et al. Indoor self-supervised monocular depth estimation based

on level feature fusion[J]. Optics and precision engineering, 2023, 31(20): 2993−3009.

[20] 姚广顺, 孙韶媛, 方建安, 等. 基于红外与雷达的夜间无人车场景深度估计[J]. 激光与光电子学进展, 2017, 54(12): 121003.
YAO Guangshun, SUN Shaoyuan, FANG Jian'an, et al. Depth estimation of night driverless vehicle scene based on infrared and radar[J]. Laser & optoelectronics progress, 2017, 54(12): 121003.

[21] GUO Xiaotong, ZHAO Huijie, SHAO Shuwei, et al. SIM-MultiDepth: self-supervised indoor monocular multi-frame depth estimation based on texture-aware masking[J]. Remote sensing, 2024, 16(12): 2221.

[22] CHENG Anqi, YANG Zhiyuan, ZHU Haiyue, et al. GAM-depth: self-supervised indoor depth estimation leveraging a gradient-aware mask and semantic constraints [C]//2024 IEEE International Conference on Robotics and Automation. Yokohama: IEEE, 2024: 5367−5374.

[23] YE Xinchen, OU Yuxiang, WU Biao, et al. Self-supervised monocular depth estimation from videos via adaptive reconstruction constraints[J]. IEEE transactions on circuits and systems for video technology, 2025, 35(3): 2161−2172.

[24] 陈莹, 王一良. 基于密集特征融合的无监督单目深度估计[J]. 电子与信息学报, 2021, 43(10): 2976−2984.
CHEN Ying, WANG Yiliang. Unsupervised monocular depth estimation based on dense feature fusion[J]. Journal of electronics & information technology, 2021, 43(10): 2976−2984.

[25] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770-778.

[26] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014−09−04)[2025−05−13]. https://arxiv.org/abs/1409.1556.

[27] HUANG Gao, LIU Zhuang, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2261−2269.

[28] WANG Chaoyang, BUENAPOSADA J M, ZHU Rui, et al. Learning depth from monocular videos using direct methods[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2022−2030.

[29] MARR D, HILDRETH E. Theory of edge detection[J]. Proceedings of the royal society of London series B biological sciences, 1980, 207(1167): 187−217.

[30] KOSCHAN A, ABIDI M. Detection and classification of edges in color images[J]. IEEE signal processing magazine, 2005, 22(1): 64−73.

[31] SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor segmentation and support inference from RGBD images [C]//Computer Vision–ECCV 2012. Berlin: Springer, 2012: 746−760.

[32] DAI A, CHANG A X, SAVVA M, et al. ScanNet: richly-annotated 3D reconstructions of indoor scenes[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 2432−2443.

[33] FU Huan, GONG Mingming, WANG Chaohui, et al. Deep ordinal regression network for monocular depth estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2002−2011.

[34] HU Junjie, OZAY M, ZHANG Yan, et al. Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries[C]//2019 IEEE Winter Conference on Applications of Computer Vision. Waikoloa Village: IEEE, 2019: 1043−1051.

[35] YIN Wei, LIU Yifan, SHEN Chunhua, et al. Enforcing geometric constraints of virtual normal for depth prediction[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 5683−5692.

[36] FAROOQ BHAT S, ALHASHIM I, WONKA P. AdaBins: depth estimation using adaptive bins[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 4008−4017.

[37] NIKLAUS S, MAI Long, YANG Jimei, et al. 3D Ken Burns effect from a single image[J]. ACM transactions on graphics, 2019, 38(6): 1−15.

[38] ZHOU Junsheng, WANG Yuwang, QIN Kaihuai, et al. Moving indoor: unsupervised video depth learning in challenging environments[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 8617−8626.

[39] ZHAO Wang, LIU Shaohui, SHU Yezhi, et al. Towards better generalization: joint depth-pose learning without PoseNet[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 9148−9158.

[40] BIAN Jiawang, ZHAN Huangying, WANG Naiyan, et al. Unsupervised scale-consistent depth learning from video[J]. International journal of computer vision, 2021, 129(9): 2548−2564.

[41] YU Zehao, JIN Lei, GAO Shenghua. P2net: patch-match and plane-regularization for unsupervised indoor depth estimation[C]//Computer Vision–ECCV 2020. Cham:

Springer International Publishing, 2020: 206−222.

[42] BIAN Jiawang, ZHAN Huangying, WANG Naiyan, et al. Unsupervised depth learning in challenging indoor video: weak rectification to rescue [EB/OL]. (2020−06−04)[2025−05−13]. https://arxiv.org/abs/2006.02708v1.

[43] JIANG Hualie, DING Laiyan, HU Junjie, et al. PLNet: plane and line priors for unsupervised indoor depth estimation[C]//2021 International Conference on 3D Vision. London: IEEE, 2021: 741−750.

[44] BIAN Jiawang, ZHAN Huangying, WANG Naiyan, et al. Auto-rectify network for unsupervised indoor depth estimation[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 44(12): 9802−9813.

[45] LYU Chen, HAN Chenggong, CHEN Junhui, et al. TSD-Depth: Using transformers and self-distilling for self-supervised indoor depth estimation[J]. Optik, 2023, 288: 171219.

[46] GUO Xiaotong, ZHAO Huijie, SHAO Shuwei, et al. SP-Depth: enhancing self-supervised indoor monocular depth estimation via self-propagation[J]. Future Internet, 2024, 16(10): 375.

[47] GUO Xiaotong, ZHAO Huijie, SHAO Shuwei, et al. F2Depth: self-supervised indoor monocular depth estimation via optical flow consistency and feature map synthesis[J]. Engineering applications of artificial intelligence, 2024, 133: 108391.

[48] WEI Yi, GUO Hengkai, LU Jiwen, et al. Iterative feature matching for self-supervised indoor depth estimation[J]. IEEE transactions on circuits and systems for video technology, 2022, 32(6): 3839−3852.

[49] DONG Li, REN Qingji, SHI Jianyang, et al. BiGeoDepth: leveraging bi-geometric priors for unsupervised monocular depth estimation in indoor environments[J]. IEEE transactions on consumer electronics, 2025, 71(2): 2988−2998.

[50] WU Haifeng, GU Shuhang, DUAN Lixin, et al. Geo-Depth: from point-to-depth to plane-to-depth modeling for self-supervised monocular depth estimation[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2025: 11525−11535.

[51] HAN Chenggong, LV Chen, HUANG Xiaolin, et al. PRDepth: pose refinement enhancement-based monocular depth estimation for indoor scenes[J]. IEEE transactions on instrumentation and measurement, 2025, 74: 5028216.

作者简介：

寇旗旗，副教授，主要研究方向为图像处理、智能检测与模式识别、图像增强与复原，主持国家自然科学基金项目 1 项，发表学术论文 80 余篇，获得发明专利授权 13 项。E-mail：kouqiqi@cumt.edu.cn。

陈飞宇，硕士研究生，主要研究方向为深度估计、图像质量评价。E-mail：TS23060160P31@cumt.edu.cn。

程德强，教授，主要研究方向为智能传感与控制、图像处理与计算机视觉。主持国家自然科学基金项目 3 项，发表学术论文 120 余篇，出版专著 2 部。E-mail：chengdq@cumt.edu.cn。

[责任编辑：丁钰]