

基于光流解耦与边缘补偿的室内动态场景 自监督单目深度估计方法

程德强^{1),2)} 张培媛¹⁾ 崔文¹⁾ 韩成功¹⁾ 江鹤¹⁾ 寇旗旗^{2),3)}

¹⁾(中国矿业大学信息与控制工程学院 江苏 徐州 221116)

²⁾(深地科学与工程云龙湖实验室 江苏 徐州 221116)

³⁾(中国矿业大学计算机科学与技术学院/人工智能学院 江苏 徐州 221116)

摘 要 现有自监督深度估计方法对室外场景的运动物体处理已较为成熟,但在面对室内场景中存在的大量无序且剧烈移动的动态像素时仍难以准确估计其深度。为此,本文提出一种基于光流解耦与边缘补偿的室内动态场景单目深度估计方法。首先,设计光流解耦建模网络,通过分解完整3D光流场为静态背景与动态前景,分别建模场景中的相机运动与动态目标运动,将破坏光度一致性的动态像素转化为可学习的运动实体,为深度估计网络提供鲁棒的自监督信号;其次,提出几何边缘补偿模块,向深度网络注入拉普拉斯算子提取的几何边缘先验,并融合上层深度预测进行边缘引导,补偿因动态物体运动模糊导致的深度边界信息退化。在Bonn、TUM和NYUv2三个室内数据集上的实验结果表明,本文方法在关键精度指标 $\delta < 0.125$ 上分别达到0.907、0.813与0.856,其中Bonn和TUM数据集的动态区域 $\delta < 0.125$ 分别达到0.774和0.807,验证了所提方法在室内动态场景深度估计中的有效性和鲁棒性。

关键词 室内深度估计;自监督学习;动态场景;光流估计;边缘补偿

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2026.01687

Self-Supervised Monocular Depth Estimation for Indoor Dynamic Scenes Based on Optical Flow Decoupling and Edge Compensation

CHENG De-Qiang^{1),2)} ZHANG Pei-Yuan¹⁾ CUI Wen¹⁾ HAN Cheng-Gong¹⁾

JIANG He¹⁾ KOU Qi-Qi^{2),3)}

¹⁾(School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, Jiangsu 221116)

²⁾(Yunlong Lake Laboratory, Xuzhou, Jiangsu 221116)

³⁾(School of Computer Science and Technology/ School of Artificial Intelligence China University of Mining and Technology, Xuzhou, Jiangsu 221116)

Abstract Existing self-supervised monocular depth estimation methods have shown considerable improvement in handling moving objects in outdoor scenes; however, when applied to indoor environments, they struggle to accurately estimate depth when confronted with densely distributed dynamic pixels exhibiting disordered and drastic movement in indoor scenes. In indoor environments, dynamic objects such as pedestrians frequently occupy large proportions of the image and exhibit complex, irregular motion patterns, causing widespread violations of the photometric consistency assumption that self-supervised methods fundamentally rely upon. This challenge is further compounded by the prevalence of motion blur, which degrades depth boundary

收稿日期:2025-08-15;在线发布日期:2026-04-01。本课题得到深地科学与工程云龙湖实验室项目资助(No. 104024005)、中央高校基本科研业务费专项资金资助(No. 2024ZDPYCH1001)、国家自然科学基金(Nos. 52204177, 52304182)资助。程德强,博士,教授,中国计算机学会(CCF)会员,主要研究领域为图像处理、机器视觉。E-mail: chengdq@cumt.edu.cn。张培媛,硕士研究生,主要研究领域为深度估计。崔文,硕士研究生,主要研究领域为图像恢复。韩成功,博士研究生,主要研究领域为深度估计。江鹤,博士,助理教授,主要研究领域为图像恢复。寇旗旗(通信作者),博士,副教授,主要研究领域为模式识别。E-mail: kouqiqi@cumt.edu.cn。

information and makes accurate structural estimation particularly difficult. To address this limitation, this paper proposes a self-supervised monocular depth estimation method based on optical flow decoupling and edge compensation for indoor dynamic scenes. First, an optical flow decoupling modelling network is designed to decompose the complete three-dimensional optical flow field into distinct static background and dynamic foreground components, thereby separately modelling camera motion and dynamic object motion in the scene. As a result, this approach transforms dynamic pixels that violate photometric consistency into learnable motion entities, which in turn provide robust self-supervised signals for the depth estimation network. Second, a geometric edge compensation module is proposed that injects Laplacian operators into the depth network to extract geometric edge priors. Building on this, these geometric edge priors are then fused with higher-level depth predictions for edge guidance, thus compensating for depth boundary information degradation caused by motion blur from dynamic objects. Experimental results on three indoor datasets—namely Bonn, TUM, and NYUv2—demonstrate that the proposed method achieves accuracy scores of 0.907, 0.813, and 0.856, respectively, on key precision evaluation metrics. Specifically, within dynamic regions, the proposed method reaches scores of 0.774 and 0.807 on the Bonn and TUM datasets, respectively, further validating the effectiveness and robustness of the proposed method for indoor dynamic scene depth estimation.

Keywords indoor depth estimation; self-supervised learning; dynamic scenes; optical flow estimation; edge compensation

1 引 言

深度信息作为场景理解的基础,在目标检测与识别^[1]、自动驾驶^[2]、机器人导航^[3]、三维重建^[4]等众多应用中发挥着至关重要的作用。准确的深度感知不仅能够帮助智能系统理解空间结构,还为后续的路径规划、物体检测等提供了必要的几何约束。

随着深度学习技术的快速发展,基于卷积神经网络(Convolutional Neural Network, CNN)的自监督单目深度估计方法^[5-6]凭借其仅需单个相机即可获取训练数据、无需昂贵的深度标注等优势,得到了广泛应用和发展。特别是在自动驾驶等室外场景中,这类方法通过挖掘视频序列中的多视图几何关系^[7],实现了高精度的深度估计。目前主流算法仍基于静态场景假设^[8],认为场景中所有物体在相邻帧间保持静止,但在现实场景中,行驶的车辆、行人等都会破坏多视图几何约束^[9],导致深度估计精度下降,现有研究通过引入运动掩码^[8]、语义模型^[7]或光流引导^[10]等较好地缓解了运动物体所带来的干扰。近年来,研究者们进一步探索了基于地面接触先验的粗到细训练策略^[11]以精确监督动态物体的深度学习,提出了自蒸馏框架^[12]通过教师-学生网络增强动态场景的泛化能力,以及设计了全局-局部信息

聚合的轻量化网络^[13]实现实时深度估计,这些方法从不同维度改善了动态场景下的深度估计性能。

然而,当这些运动像素出现在复杂的室内场景中时,其造成的干扰却难以通过简单地迁移上述方法得到有效解决。首先,室内动态物体的运动轨迹较室外驾驶场景等规则性差,人员的驻足、转身等行为完全脱离刚体运动约束^[14],其位移呈现高度无序性;同时,动态物体在相机曝光时间内的持续运动会在成像传感器上产生拖影效应,导致物体几何边界出现严重的运动模糊^[15],这种边界退化效应在深度估计过程中引发匹配错误,致使物体轮廓在最终深度图中严重模糊、扭曲;更为严峻的是,室内空间狭窄导致动态目标近距成像^[16],致使动态像素相较室外场景占比更高,相应干扰随之加剧,进一步放大了对深度网络带来的影响。

针对上述挑战,本文提出一种基于光流解耦与边缘补偿的室内动态场景自监督单目深度估计方法(Self-Supervised Monocular Depth Estimation for Indoor Dynamic Scenes based on Optical Flow Decoupling and Edge Compensation, FE-Depth),该方法由光流解耦建模网络(Flow Decoupling Modeling Network, FDMN)与几何边缘补偿模块(Geometric Edge Compensation Module, GECM)构成。其中,FDMN利用光流将场景中的动态目标显

示建模,通过解耦场景中静态背景与动态前景,将破坏光度一致性的动态像素转化为可学习的运动实体,为深度估计网络提供可靠自监督信号,提高网络在室内动态环境下的鲁棒性;GECM通过向深度网络注入零参拉普拉斯算子提取的几何边缘先验,利用原始图像的高频信息以及上层深度预测对物体边缘进行引导,补偿因动态性造成的边界信息损失,满足室内动态场景对高精度几何边缘的需求。在 Bonn, TUM, NYUv2 三个室内公开数据集上对 FE-Depth 进行实验,最关键指标 $\delta < 0.125$ 分别为 0.907、0.813、0.856,其中,FE-Depth 在 Bonn 和 TUM 数据集上动态区域 $\delta < 0.125$ 分别达到 0.774、0.807。

2 相关工作

2.1 室外自监督深度估计

自监督单目深度估计^[17]作为机器视觉领域的关键技术,凭借其仅需单目视频的训练特性,在自动驾驶等室外场景的环境感知中得到了广泛应用。为了缓解场景中行人、车辆等动态物体的干扰,研究者们提出了多种处理策略。Monodepth2^[8]使用自动掩膜机制通过比较光度误差自动识别并过滤动态区域,同时排除低纹理区域。PackNet-SfM^[7]采用两阶段训练策略,借助语义分割模型剔除包含动态物体的帧并重采样训练数据,以缓解动态物体造成的语义偏差。SGDepth^[18]根据光度差异判断物体的实际运动状态,从静止的动态类别中学习深度信息,Ranftl 等人^[19]采用 Vision Transformer 的全局上下文建模能力显著改善了物体边界处的深度估计质量。最近,Manydepth2^[10]将光流信息与深度估计结果相结合,构建了运动引导的代价体积框架,AFNet^[20]基于重投影置信度自适应选择单目/多目分支。尽管这些方法在室外场景成效显著,但其有效性依赖于驾驶场景的特定假设,如地面平面约束、相对稀疏的动态物体分布以及较大的相机运动幅度等。这些假设在室内环境中普遍失效,限制了现有方法在室内场景中的直接应用。

2.2 室内自监督深度估计

相较于室外场景,室内自监督深度估计面临更高的动态复杂性。一方面,室内场景的相机运动通常幅度较小且非结构化,缺乏驾驶场景中前向主导模式^[21];另一方面,室内环境中物体几何结构复杂且深度突变现象更为频繁。针对这些特性,

MonoIndoor^[14]设计了专用于深度分解的 Transformer 架构,通过注意力机制应对室内场景的深度范围变化和复杂几何结构,改善了细节表征能力。SC-DepthV3^[9]引入外部单目深度估计模型生成伪标签,并设计动态区域精化和局部结构精化模块来约束自监督训练。StructDepth^[22]提取场景主方向提供平面先验,以改善平面区域的深度估计质量,IndoorDepth₁提出了增强的结构相似性 (Structural Similarity Index Measure, SSIM) 函数用于光度损失的测量,通过改进损失函数设计提升了深度估计的整体质量。这些方法在一定程度上提升了预测精度,但在处理动态物体干扰时缺乏对运动的显示建模,过度依赖动态区域掩膜或外部先验;且动态区域边缘模糊与细节丢失问题仍然突出,存在物体边缘模糊和细节丢失严重等问题。

2.3 光流辅助的单目深度估计

近年来,利用光流辅助的单目深度估计,已成为处理动态场景的关键范式。该类方法的核心在于利用光流建模物体运动,为深度估计网络提供更为鲁棒的监督信号。早期工作如 DF-Net^[23],通过建立深度与光流之间的几何一致性约束进行联合学习,为后续研究奠定了基础。为进一步提升对动态物体的处理精度,光流解耦技术逐渐成为研究焦点。Sun 等人的 FlowDepth₂提出了动态运动流模块,采用纯几何机制实现光流解耦,无需依赖额外的语义先验网络,展现了更好的可解释性与泛化能力。此外,Nguyen 等人^[24]则探索了从光流信息中挖掘动态区域的伪深度标签,以解耦动静区域的深度估计。近期,生成式模型也被引入该领域,通过扩散模型^[25]或 Flow Matching 等^[26]技术生成光流或深度,以应对估计任务中固有的模糊性与不确定性。现有光流辅助方法主要面向室外驾驶场景,对室内环境中动态物体运动轨迹无序且像素占比大等问题缺乏针对性处理。本文聚焦于室内动态场景的独特挑战,提出光流解耦与边缘补偿相结合的深度估计框架,通过对物体运动状态的建模以及运动边界的规范,有效提升室内场景的深度估计精度。

1 Fan C, Yin Z, Li Y, et al. Deeper into self-supervised monocular indoor depth estimation, <https://arxiv.org/abs/2312.01283>, 2023, 12, 3

2 Sun Y, Xu Z, Wang X, et al. FlowDepth: Decoupling Optical Flow for Self-Supervised Monocular Depth Estimation, <https://arxiv.org/abs/2403.19294>, 2024

3 网络模型

FE-Depth整体框架如图1所示,主要由三个网络构成:光流解耦建模网络(Flow Decoupling Modeling Network, FDMN)、集成了GECM模块的边缘补偿深度估计网络(Edge-Enhanced Depth Estimation Network, EE-DepthNet)以及位姿估计网络(PoseNet)。EE-DepthNet采用编码器-解码器架构,在解码器的四个跳跃连接层级($H \times W$ 至 $H/16 \times W/16$)中嵌入GECM模块,将拉普拉斯算

子提取的高频边缘特征作为先验知识注入到深度估计过程,提升对室内动态区域边界的深度估计精度。位姿估计网络以相邻图像对作为输入,估计连续帧之间的相机相对位姿变换,为后续几何约束和光流解耦提供基础。FDMN由全流场生成网络(Holistic Flow Generator, HFC)和运动掩码网络(Motion Mask Network, MMN)组成, HFC预测完整的3D光流场,捕捉场景中所有像素的运动信息; MMN输出流场调制掩码,通过概率映射标识独立运动区域。两个子网络协同工作,通过光流解耦以及重构准确建模物体的三维运动。

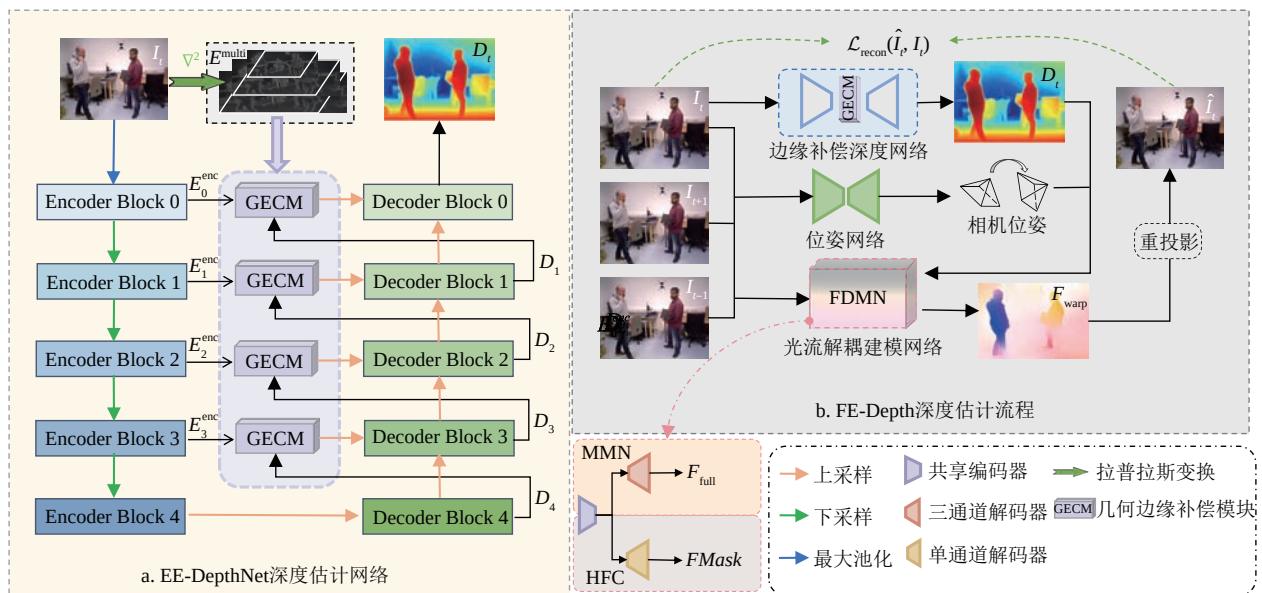


图1 FE-Depth框架图

3.1 光流解耦建模网络(Flow Decoupling Modeling Network, FDMN)

针对动态场景中静态世界假设失效问题,现有方法主要涵盖两类技术路径:一是使用动态屏蔽掩码^[8,27]直接剔除重建误差显著的像素区域以消除动态物体的干扰,但会导致相应区域失去有效监督信号,无法学习到准确的深度表示;另一种是引入额外的语义分割^[18]或物体检测模型,利用语义-运动关联性识别动态区域,但室内场景中语义静态物体会因环境扰动(如气流或人为接触)产生运动,引起语义-运动不一致^[27]问题,从而导致动态区域的误识别。因此,如何精确识别独立运动物体并建模其三维几何结构与运动状态,是动态场景理解的核心科学问题。

基于上述分析,本文将光流感知机制引入自监督深度估计框架,设计光流解耦建模网络(Flow

Decoupling Modeling Network, FDMN)作为解决室内动态场景挑战的关键策略。光流是帧间像素运动的总现象,同时编码相机位姿变化与独立物体运动引发的像素位移,能够显式建模场景中的真实运动场。尤其是对于室内场景中人员和物体的独立且不规律的运动,光流可以敏锐地捕捉此类非结构化、多方向的运动,为动态区域建模提供关键运动信息支撑。如图2a所示, FDMN由全流场生成网络和运动掩码网络两个核心构件组成。FDMN以在通道维度拼接的连续三帧图像 I_{t-1}, I_t, I_{t+1} 作为输入,采用 ResNet18 作为 HFC 与 MMN 的共享编码器。HFC(图2c)通过三通道解码器将相机位姿计算的刚性运动先验与层级特征进行残差学习、多尺度细

1 Slinko I, Vorontsova A, Konokhov F, et al. Scene motion decomposition for learnable visual odometry, <https://arxiv.org/abs/1907.07227>, 2019, 7, 16

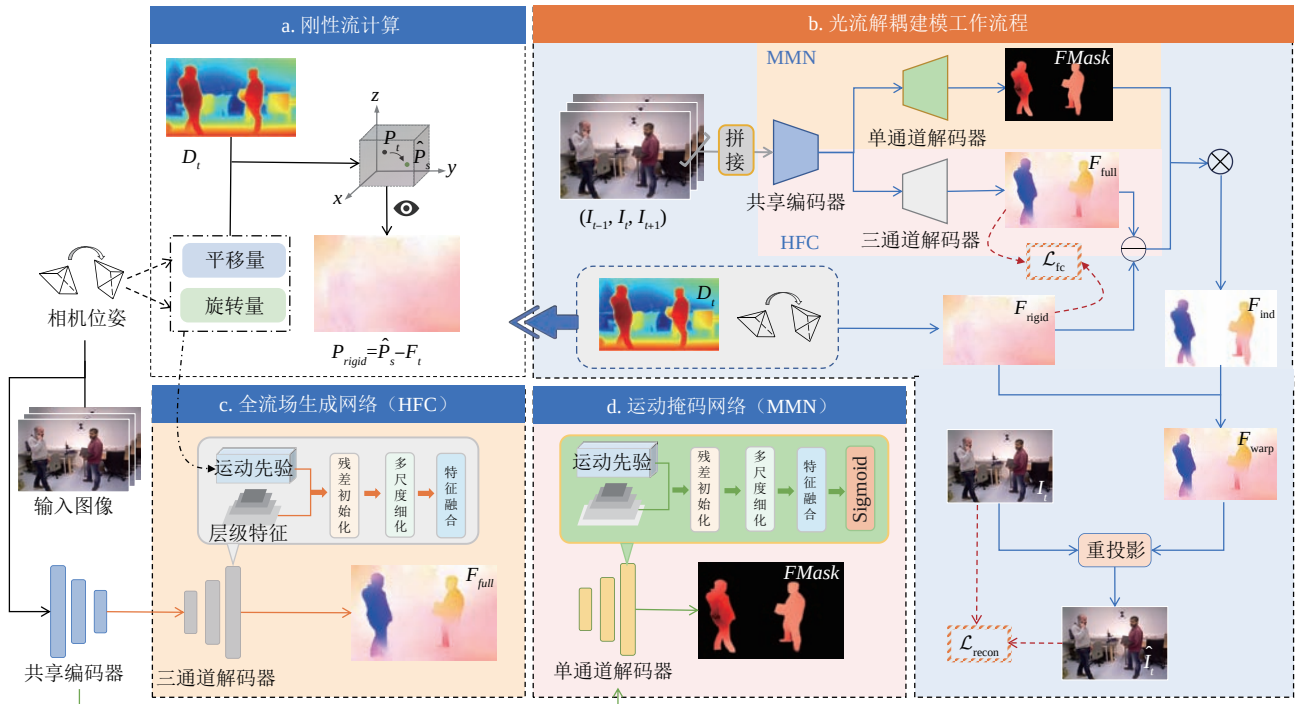


图2 FDMN框架图

化与特征融合,输出完整的3D光流场(Complete 3D Flow Field) F_{full} , 对应于三维空间中的 xyz 方向的运动分量,该过程旨在捕捉场景中像素的整体运动,为后续运动解耦提供基础光流表征。MMN(图2d)通过单通道解码器生成流场调制掩码 $FMask$ 。 $FMask$ 以二值化概率场形式体现像素属于独立运动物体的似然度。区别于传统的动态掩码的剔除机制, $FMask$ 作为刚性流(Rigid Flow) F_{rigid} 与 F_{full} 融合过程中的动态加权系数,通过精细的运动分解和自适应融合策略使网络能够理解并建模动态物体。 F_{rigid} (图2b)由深度信息与位姿信息计算得到,反映静态场景点的三维运动投影。具体而言,给定预测深度图和相机相对位姿,可通过三维几何投影计算:

$$P_t = D_t(p_t) K^{-1} \vec{p}_t \quad (1)$$

$$\hat{P}_s = T_{t \rightarrow s} \vec{P}_t \quad (2)$$

$$F_{rigid} = \hat{P}_s - P_t \in \mathbb{R}^3 \quad (3)$$

其中, P_t 为像素点 p_t 在目标帧 I_t 坐标系中的3D位置, $\vec{p}_t$ 为 p_t 的齐次坐标形式, $\vec{P}_t$ 为 P_t 在源帧 I_s 坐标系中的位置, F_{rigid} 定义为 P_t 与 $\hat{P}_s$ 之间的三维位移向量,严格表征了静态场景点在相机运动下的三维运动场。

如图2b所示,独立运动光流 F_{ind} 通过完整的3D光流场 F_{full} 与刚性流 F_{rigid} 残差计算得到,并通过流场调制掩码 $FMask$ 进行空间调制:

$$F_{ind} = FMask \otimes (F_{full} - F_{rigid}) \quad (4)$$

其中, $\otimes$ 表示逐元素相乘,这种分解方式显式解耦了场景流中的刚性运动与非刚性运动分量。

将代表物体独立运动的 F_{ind} 与代表场景中静态部分的 F_{rigid} 进行融合,构建综合场景中的所有三维运动信息的组合光流场(Warping Flow Field) F_{wrap} , 具体表示为:

$$F_{wrap} = F_{ind} + F_{rigid} \in \mathbb{R}^3 \quad (5)$$

基于此,可以计算目标帧中像素点 p_t 所对应三维点在源帧坐标系中的位置 $\hat{P}_s'$:

$$\hat{P}_s' = F_{wrap} + P_t \quad (6)$$

最后,组合光流场 F_{wrap} 用于计算源帧到目标帧的视图重建,为深度网络提供监督信号,重建过程通过三维点投影实现:

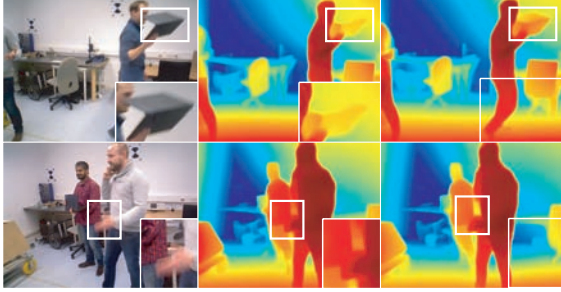
$$\hat{I}_t(p_t) = I_s(\pi(\hat{P}_s')) \quad (7)$$

其中 $\pi: \mathbb{R}^3 \rightarrow \mathbb{R}^2$ 实现三维至二维的映射,重建误差作为深度估计网络的自监督信号。同时研究设计静态流约束损失 $\mathcal{L}_{fc}$, 通过对齐 F_{full} 和 F_{rigid} 指导 HFC 的学习。

3.2 几何边缘补偿模块(Geometric Edge Compensation Module, GECM)

动态物体在曝光期内的位移导致成像过程产生运动模糊伪影,其边界呈现弥散型模糊。在室内等狭窄环境中,动态物体与相机的典型距离范围(1-3米)

显著放大了位移的视觉影响。如图3所示,移动的电脑、手势等动态区域边界出现明显的运动模糊,导致深度估计结果在物体轮廓处出现几何失真。



注. 第2列由 SC-DepthV3^[9]生成, 第3列由本文方法生成。

图3 动态区域边界退化现象

基于此, GECEM 采用物理驱动的边缘补偿机制, 利用拉普拉斯算子 L 的运动不变性来对抗动态模糊。 L 作为二阶微分算子, 对梯度突变敏感, 可以在模糊使边缘强度衰减的情况下保持梯度极值点位置的稳定, 有效捕获图像中的边缘和细节信息。且该方法具有零参鲁棒性, 在训练数据稀缺的情况下提供稳定和精确的边缘信息, 使得网络能够在解码器的每一层级获得可靠的边缘指导。

如图4所示, GECEM 接收多尺度边缘引导特征 E^{multi} 、编码器特征 E^{enc} 以及来自更深层的深度预测

D_{i+1} 三个关键输入。通过对输入 RGB 图像 I 应用拉普拉斯金字塔分解获得:

$$E^{\text{multi}} = \{L_0(I), L_1(I), L_2(I)\} \quad (8)$$

其中 L 表示拉普拉斯算子, $L_k(I)$ 表示输入图像 I 的 k 级拉普拉斯金字塔分解, 各层级分别对应不同分辨率下的边缘响应, 能够有效提取从局部细节到全局结构的多尺度边界信息, 为深度网络提供精确的边界定位。

在多尺度边缘特征融合阶段, 首先对来自拉普拉斯金字塔不同层级的 E^{multi} 进行空间尺寸归一化, 并在通道维度将 E^{multi} 堆叠为统一的张量表示, 随后, 通过一个轻量级的权重预测网络, 自适应地平衡不同尺度特征的贡献度, 得到尺度相关的注意力权重 A 。最终, 高频特征注意力 A_i^{high} 通过加权融合构建:

$$A_i^{\text{high}} = \text{Conv}_{3 \times 3}(E^{\text{multi}} \otimes A) \quad (9)$$

其中, $\otimes$ 表示逐元素相乘, A_i^{high} 利用 RGB 图像的多尺度边缘信息补偿深度预测的高频细节缺失, 实现跨模态的特征对齐。在自适应注意力生成阶段, A_i^{high} 对编码器特征 E^{enc} 执行边缘引导的特征调制操作, 通过逐像素的权重调节实现高频细节的自适应增强, 得到高频特征 E_i^{high} :

$$E_i^{\text{high}} = E_i^{\text{enc}} \otimes A_i^{\text{high}} \quad (10)$$

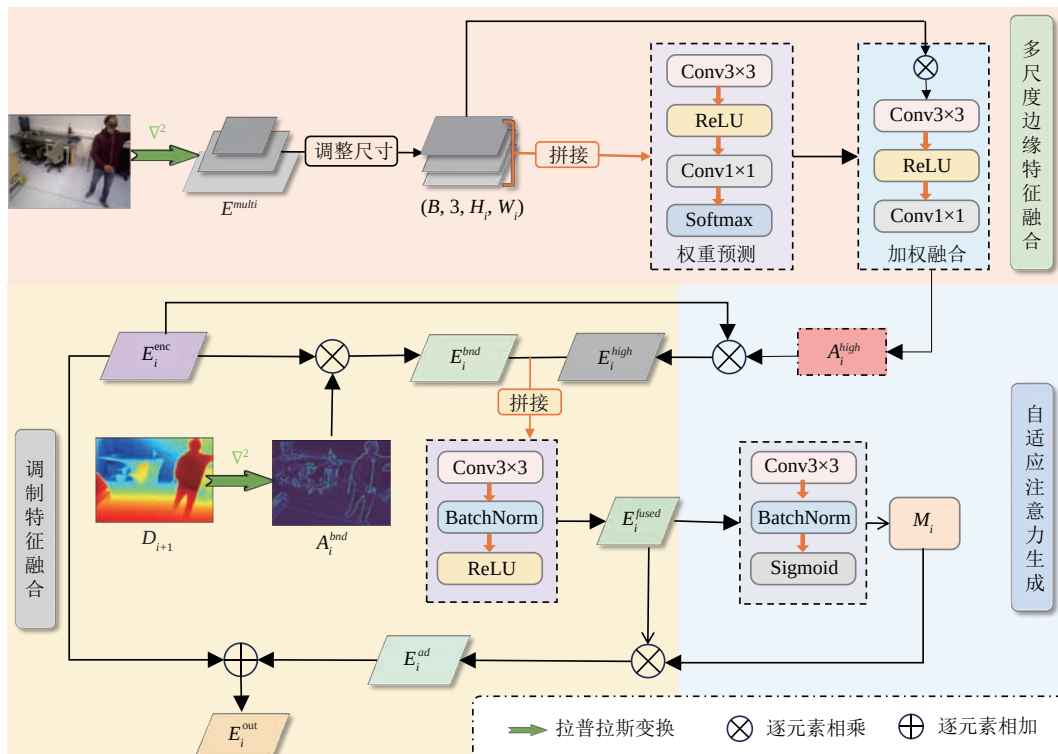


图4 GECEM 框架图

对更深层的深度预测 D_{i+1} 通过拉普拉斯算子提取深度边界, 计算边界注意力权重 A_i^{bnd} , 将 A_i^{bnd} 应用于 E_i^{enc} , 得到边界增强特征 E_i^{bnd} , 通过深度预测中的几何边界先验指导特征学习, 实现深度不连续性的显式建模和增强。 A_i^{bnd} 由深度信息驱动, 关注物理世界的几何结构, 而 A_i^{high} 由原始图像驱动, 关注视觉世界的纹理细节, 两者从不同维度为深度估计提供约束和指导。具体表示如下:

$$A_i^{bnd} = L_0(D_{i+1}) \quad (11)$$

$$E_i^{bnd} = E_i^{enc} \otimes A_i^{bnd} \quad (12)$$

将调制后的特征 E_i^{bnd} 与 E_i^{high} 进行初步整合以及维度压缩, 以消除特征冗余并提取融合特征的高层语义表示, 得到融合特征 E_i^{fused} :

$$E_{\text{contact}} = [E_i^{bnd}, E_i^{high}] \in \mathbb{R}^{B \times 3 \times H \times W} \quad (13)$$

$$E_i^{fused} = \text{ReLU}(\text{BN}(\text{Conv}_{3 \times 3}(E_{\text{contact}}))) \quad (14)$$

其中, ReLU 表示激活函数, BN 表示批量归一化。

为了抑制边缘信息中可能包含的噪声并突出关键区域, 本模块引入自适应注意力掩码 M_i , M_i 根据当前层特征分布自适应地调整边缘信息的增强程度, 从而避免固定权重可能导致的过度增强或增强不足问题。对 E_i^{fused} 应用 M_i 进行全局调制, 得到自适应特征 E_i^{ad} , 具体表示为:

$$M_i = \sigma(\text{BN}(\text{Conv}_{3 \times 3}(E_i^{fused}))) \quad (15)$$

$$E_i^{ad} = E_i^{fused} \otimes M_i \quad (16)$$

其中, σ 表示 Sigmoid 激活函数。

最终, 输出特征 E_i^{out} 通过残差连接计算, 并采用卷积块注意力模块^[28] (Convolutional Block Attention Module, CBAM) 对自适应特征进一步优化:

$$E_i^{in} = E_i^{ad} \oplus E_i^{enc} \quad (17)$$

$$E_i^{out} = \text{CBAM}(E_i^{in}) \quad (18)$$

其中, CBAM 包含通道注意力 (Channel Attention) 和空间注意力 (Spatial Attention) 两个子模块, 分别在通道维度和空间维度上重新校准特征权重, 确保重要的特征通道和关键的空间位置获得更高的响应。

3.3 损失函数

3.3.1 静态流约束损失

基于静态场景的刚体运动假设, 场景中静态区域的像素运动应完全由相机运动决定。为此, 本文提出静态流约束损失 $\mathcal{L}_{fc}$, 该损失强制要求, 在静态区域内, 网络预测的完整的 3D 光流场与基于相机姿态和深度估计计算所得的刚性流保持一致。其数学表达式如下:

$$\mathcal{L}_{fc} = \frac{1}{|\mathcal{S}_{\text{static}}|} \sum_{p \in \mathcal{S}_{\text{static}}} (1 - \text{FMask}(p)) \cdot \|F_{\text{full}}(p) - F_{\text{rigid}}(p)\|_1 \quad (19)$$

其中, $\mathcal{S}_{\text{static}}$ 为静态区域像素集合。

通过最小化静态区域内残差流的 L_1 范数, $\mathcal{L}_{fc}$ 强制网络学习物理上合理的运动分解, 确保静态背景区域的运动完全由相机姿态变化解释, 而独立物体运动仅在动态区域内存在。这一约束有效解决了动态场景深度估计中的运动歧义性问题, 为无监督学习提供了重要的几何先验约束。

3.3.2 运动场光度重建与几何一致性损失

运动场光度重建损失 $\mathcal{L}_{\text{recon}}$ 作为整个框架的核心监督信号, 基于光度一致性原理, 采用 SSIM 和 L_1 损失的加权组合, 通过最小化重建图像 $\hat{I}_t$ 与真实图像 I_t 之间的差异, 联合优化深度网络、位姿网络和光流解耦建模网络:

$$\mathcal{L}_{\text{recon}}(\hat{I}_t, I_t) = \frac{\alpha}{2} (1 - \text{SSIM}(I_t, \hat{I}_t)) + (1 - \alpha) \|I_t - \hat{I}_t\|_1 \quad (20)$$

其中, 重建图像 $\hat{I}_t$ 由完整 3D 运动场 F_{wrap} 三维几何投影生成, 使得 $\mathcal{L}_{\text{recon}}$ 不仅可以进行基础的深度-姿态约束, 还能够有效处理动态场景中的非刚性运动, 细化运动建模。几何一致性损失 $\mathcal{L}_{\text{geo}}$ 通过深度一致性约束, 确保多视图几何关系的合理性:

$$\mathcal{L}_{\text{geo}} = \frac{1}{|\mathcal{V}|} \sum_{p \in \mathcal{V}} \frac{|D_c(p) - D_p(p)|}{D_c(p) + D_p(p)} \quad (21)$$

其中, D_c 为计算深度, 表示目标帧像素对应的 3D 点在源帧相机坐标系中的几何深度值。 D_p 为投影深度, 表示在源帧深度图中, 对应投影位置处的深度预测值。 $\mathcal{V}$ 为有效像素集合, 排除遮挡区域、图像边界及深度不可靠区域。

3.3.3 深度平滑损失

深度平滑损失 $\mathcal{L}_s$ 是自监督单目深度估计中的重要正则化项, 基于自然场景中深度变化通常是连续且平滑的先验假设, 通过正则化约束确保预测结果的空间连续性和合理性, 防止过拟合并提高预测的鲁棒性, 在保留物体边界锐度的前提下, 约束深度图的局部平滑性:

$$\mathcal{L}_s = \sum_p |\nabla d(p)| \cdot e^{-|\nabla I(p)|} \quad (22)$$

最终, FE-Depth 的总损失为:

$$\mathcal{L}_{\text{total}} = \lambda_{fc} \mathcal{L}_{fc} + \lambda_{\text{recon}} \mathcal{L}_{\text{recon}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}} + \lambda_s \mathcal{L}_s \quad (23)$$

其中, λ_{fc} , λ_{recon} , λ_{geo} , λ_s 表示各项损失权重。

4 实 验

4.1 实验细节

本文模型基于PyTorch实现,训练过程使用单个Nvidia GeForce RTX 3090 GPU,在30个epoch内完成训练。批尺寸(batch size)设置为8,输入/输出分辨率为 256×320 。采用Adam优化器进行联合训练深度估计网络、位姿估计网络和光流解耦建模网络,其中 $\beta_1 = 0.9$, $\beta_2 = 0.999$,学习率设置为 1×10^{-4} ,训练过程中保持恒定。采用PyTorch Lightning框架进行训练管理,每个epoch包含1000个训练步骤。

数据增强:为提高训练的鲁棒性,对图像进行预处理。具体来说,以50%的概率执行以下增强操作:水平翻转、亮度调整(± 0.2)、对比度调整(± 0.2)、饱和度调整(± 0.2)和色调抖动(± 0.1)。同时采用随机缩放裁剪策略,缩放因子范围为1.0~1.15,增强模型的泛化能力。

深度估计网络:编码器基于ResNet18架构,使用ImageNet预训练权重进行初始化。GECM模块部署在解码器跳跃连接的四个不同层级中,分别对应 $H/2 \times W/2$ 、 $H/4 \times W/4$ 、 $H/8 \times W/8$ 、 $H/16 \times W/16$ 特征分辨率。

位姿估计网络:编码器基于ResNet18架构,解码器通过全局平均池化将特征降维至256维,随后通过三层卷积(步长为1的 3×3 卷积)和最终的 1×1 卷积输出6维位姿参数。位姿预测包括3维轴角表示的旋转和3维平移向量,通过0.01的缩放因子防止位姿估计的尺度退化问题。

光流解耦建模网络:FDMN由HFC和MMN两个子网络组成,共享ResNet18编码器结构。HFC通过三通道解码器输出完整的3D光流场,MMN输出单通道的运动概率掩码,通过Sigmoid激活函数转换为 $[0, 1]$ 范围的概率值。

4.2 数据集

FE-Depth旨在提高在具有复杂动态物体的室内场景中的深度估计精度,因此本文主要选取高动态性室内数据集进行评估:

Bonn数据集^[29]是专门为RGB-D SLAM设计的高度动态室内场景数据集,包含24个动态序列和2个静态序列。动态序列中包含人员执行各种任务,如操作箱子、玩气球等快速移动行为,为室内场

景的深度估计提供了极具挑战性的测试环境。数据集使用Optitrack Prime 13运动捕捉系统记录传感器的真实位姿,并通过Leica BLK360地面激光扫描仪提供静态环境的真实3D点云作为几何参考。本文实验参考SC-DepthV3的划分方法,在26个室内视频序列中选择4个包含快速移动人员的具有挑战性序列(共1735张图像)作为测试集,其余序列用于训练。所有图像均调整为 320×256 分辨率输入网络,深度范围限制在0米~10米,确保评估结果符合室内场景的实际深度分布特征。

TUM数据集^[30]提供了丰富的室内视频序列以及相应的Kinect捕获深度图作为真实标签。为了专注于动态场景下的深度估计性能评估,本文在属于Dynamic Objects类别的11个序列中选择最后2个包含移动人员的序列(共1375张图像)构成测试集,其余9个动态视频序列用于训练,Dynamic Objects类别包含多种室内动态行为,如人员走动、物体移动等非刚性运动模式,以确保模型在包含独立运动物体的复杂场景中得到充分验证。所有输入图像统一调整为 320×256 分辨率,深度评估范围设定为0米~10米。

NYUv2数据集^[31]包含464个室内视频序列,由手持Microsoft Kinect RGB-D相机以 640×480 分辨率拍摄。该数据集包含来自464个不同室内场景的原始视频序列,按照官方标准划分,795幅图像用于训练,654幅图像用于测试。在训练过程中,图像大小被调整为 320×256 。

4.3 对比实验

为评估所提方法在室内动态场景中的性能,本文参照SC-DepthV3^[9],将BONN和TUM数据集测试图像划分为动态区域与静态区域。其中动态目标主要为快速移动的人体,因此将所有的人形分割区域视为动态区域,其余视为静态背景。深度评估时,首先对齐预测深度与真实深度的全局尺度,随后分别在动态区域和完整图像上独立评估深度估计精度,以量化分析所提方法在处理动态物体时的性能提升。NYUv2数据集测试图像主要为静态室内场景,因此重点比较整体深度估计精度。

(1) 在BONN数据集上,将本文方法与Monodepth2^[8]以及SC-Depth系列^[9,32-33]进行对比,定量对比结果如表1所示,对比方法的数值为原文献结果,准确度值越高越好(下文同),误差值越低越好(下文同)。在动态区域评估中,本文方法的AbsRel为0.194,相比于最强基线SC-DepthV3^[9]降

表1 在 BONN 数据集上完整图像以及动态区域的深度估计结果对比

方法	完整图像					动态区域		静态区域	
	AbsRel	RMSE	$\delta < 0.125$	$\delta < 0.125^2$	$\delta < 0.125^3$	AbsRel	$\delta < 0.125$	AbsRel	$\delta < 0.125$
Monodepth2 ^[8]	0.565	2.337	0.352	0.591	0.728	0.474	0.172	0.594	0.383
IndoorDepth ₁	0.402	2.141	0.366	0.675	0.830	0.551	0.255	0.355	0.393
Manydepth2 ^[10]	0.264	0.685	0.671	0.866	0.947	0.720	0.153	0.166	0.762
SC-Depth ^[33]	0.272	0.733	0.623	0.858	0.948	0.704	0.166	0.180	0.714
SC-DepthV2 ^[32]	0.211	0.619	0.714	0.873	0.936	0.488	0.247	0.152	0.803
SC-DepthV3 ^[9]	0.126	0.379	0.889	0.961	0.980	0.220	0.720	0.102	0.931
本文	0.117	0.368	0.907	0.963	0.982	0.194	0.774	0.099	0.936

注:粗体表示最优值。

低 11.8%，准确度指标 $\delta < 0.125$ 相对提升 5.4%，并明显优于其他对比方法，表明所提出的方法能够更准确地估计动态区域的深度信息。在完整图像评估中，本文方法的 AbsRel 达到 0.117、 $\delta < 0.125$ 达到 0.907，为所有对比方法中的最优结果。4 种方法的定性比较结果如图 5 所示，可以看出，通过使用光流对运动进行显示建模，本文对场景内气球、手持电脑等动态物体响应更为敏锐，如第 1 行~第 3 行所示；同时，本文方法很好地解决了由运动模糊导致的深度边界泄露问题，如第 4 行~第 5 行所示。即使在快速移动的人与人、物与人之间发生交叠的情况下，本文方法仍能清晰区分其深度关系，提供准确的深度边界。

(2) 在 TUM 数据集上，将本文方法与 Monodepth2^[8] 以及 SC-Depth 系列^[9,32-33] 进行对比，定量对比结果如表 2 所示，本文方法在各项指标上较 SC-DepthV3^[9] 提升显著，动态区域准确度指标 $\delta < 0.125$ 从 0.796 相对提升至 0.807，提高 1.1%，证明了本文方法对运动物体处理的有效性；在完整

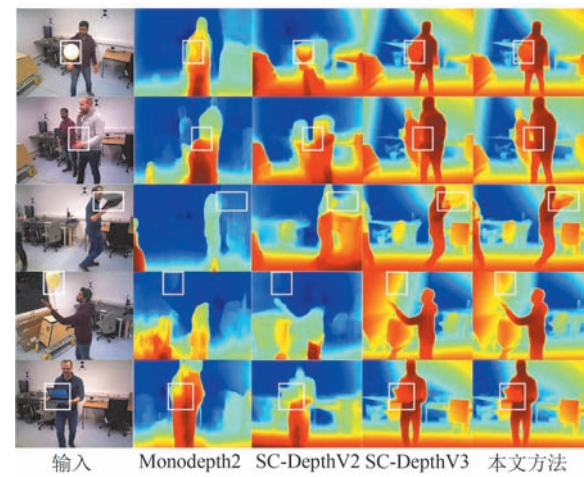


图5 在 BONN 数据集上 4 种方法的定性对比结果

图像评估中，AbsRel 从 0.163 降低至 0.146，相对误差降低 1.7%，RMSE 从 0.265 改善至 0.213，相对提升 4.3%。静态区域的性能保持稳定的改进，整体上验证了方法的普适性与有效性。4 种方法的定性比较结果如图 6 所示，可以观察到，本文方法在处理动态人物深度和保持复杂场景细节上均有明显优势。

表2 在 TUM 数据集上完整图像以及动态区域的深度估计结果对比

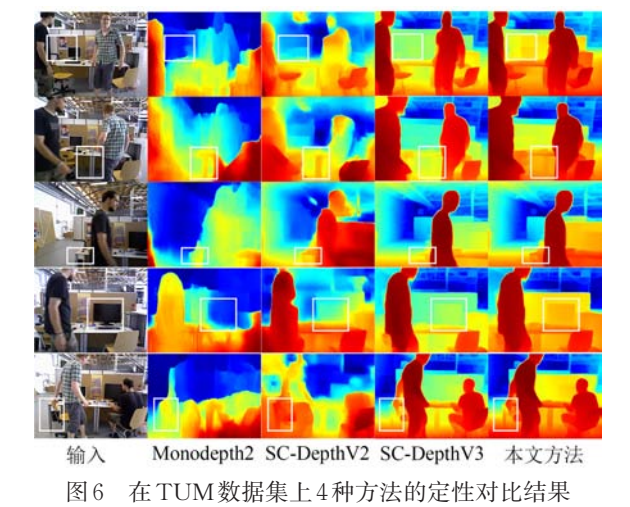
方法	完整图像					动态区域		静态区域	
	AbsRel	RMSE	$\delta < 0.125$	$\delta < 0.125^2$	$\delta < 0.125^3$	AbsRel	$\delta < 0.125$	AbsRel	$\delta < 0.125$
Monodepth2 ^[8]	0.312	1.408	0.474	0.793	0.905	0.431	0.348	0.262	0.526
IndoorDepth ₁	0.281	0.950	0.558	0.841	0.940	0.409	0.338	0.256	0.605
Manydepth2 ^[10]	0.297	1.862	0.547	0.767	0.858	0.523	0.166	0.227	0.664
DnD ^[27]	0.175	0.631	0.751	0.947	0.982	-	-	-	-
SC-Depth ^[33]	0.257	0.283	0.616	0.814	0.909	0.512	0.274	0.176	0.715
SC-DepthV2 ^[32]	0.223	0.282	0.643	0.862	0.932	0.283	0.494	0.206	0.686
SC-DepthV3 ^[9]	0.163	0.265	0.797	0.882	0.937	0.165	0.796	0.171	0.780
本文	0.146	0.214	0.813	0.932	0.984	0.158	0.807	0.149	0.793

注:粗体表示最优值。

(3) 在 NYUv2 数据集上，将本文方法与当前先进的方法进行对比，包括 SC-DepthV3^[9]、GAM-Depth^[34]、IndoorDepth₁ 等，定量对比结果如表 3 所示。其中，“监督”表示需要深度真值作为标签的监督学习。可以看出，本文方法在自监督算法中表

现优异，并超过了部分监督算法，与基线相比，在最关键指标 $\delta < 0.125$ 上相对提升 0.8。由于

1 Fan C, Yin Z, Li Y, et al. Deeper into self-supervised monocular indoor depth estimation, <https://arxiv.org/abs/2312.01283>, 2023, 12, 3



NYUv2数据集方法的核心优势在于动态物体处理,在BONN等高动态性室内数据集上相对提升幅度更为显著,在此数据集上的相对提升幅度虽有所收敛,但仍优于现有先进算法;4种方法的定性比较结果如图7所示,可以看出,在光流建模与几何边缘补偿的作用下,由运动引起的深度扩散效应被有效抑制,显著提升了复杂物体结构深度边缘的锐度与清晰度。

(4)在更具挑战性的复杂室内场景下,本文方法仍展现出较好的深度估计性能,如图8所示。在第1~2行光照不均以及弱光的条件下,本文方法通过边缘补偿保持了较好的几何结构完整性,虽然极端弱光导致部分细节损失,但仍能较好地保留门框、墙壁等关键边缘信息;在第3~4行多物体强交互的场景中,本文方法通过光流解耦有效分离前景与背景的深度关系,虽然仍存在轻微深度混淆,但相较对比方法在深度层次的准确性上具有明显提升;同时,为评估方法在不同动态场景下的泛化能力,本文进行了跨数据集零样本泛化实验。如表4所示,所有方法均在BONN数据集上训练并在TUM数据集上进行测试。实验结果表明,在零样本跨数据集测试场景下,本文方法表现优异,特别是在动态区域的评估中, $\delta < 0.125$ 达到0.582,相较SC-DepthV3^[9]提升4.0%,在完整图像以及静态区域也保持了稳定的性能,这表明本方法所采用的光流解耦建模与边缘补偿机制,能够有效学习具有强泛化能力的动态物体运动表征与边界结构特征,从而更好地适应不同室内环境中的动态场景变化。

表3 在NYUv2数据集上17种方法的深度估计结果对比

方法	监督	准确度↑			误差↓		
		$\delta < 0.125$	$\delta < 0.125^2$	$\delta < 0.125^3$	AbsRel	log10	RMSE
DORN ^[35]	✓	82.8	96.5	99.2	0.115	0.051	0.509
Hu等 ^[36]	✓	86.6	97.5	99.3	0.115	0.050	0.530
DPT ^[19]	✓	90.4	98.8	99.8	0.110	0.045	0.357
DnD ^[27]	✓	91.0	98.8	99.8	0.098	-	0.362
DepthFormer ^[37]	✓	92.3	98.9	99.7	0.094	0.040	0.329
Monodepth2 ^[8]	×	76.5	93.7	98.3	0.165	0.070	0.686
F ² Depth ^[15]	×	79.9	94.7	98.7	0.158	0.067	0.583
P ² Net ^[38]	×	79.6	94.8	98.6	0.150	0.064	0.561
Manydepth2 ^[10]	×	79.6	96.0	99.4	0.143	0.061	1.275
StructDepth ^[22]	×	81.3	95.4	98.8	0.151	0.060	0.540
IFMNet ^[39]	×	82.2	95.8	98.9	0.135	0.058	0.529
MonoIndoor ^[14]	×	82.3	95.8	98.9	0.134	-	0.526
TSD-Depth ^[40]	×	82.3	95.6	98.9	0.139	0.059	0.533
MonoIndoor++ ^[41]	×	83.4	96.1	99.0	0.132	-	0.517
LoFtDepth ^[4]	×	83.5	96.6	99.2	0.130	-	0.498
GAM-Depth ^[34]	×	83.6	96.0	99.0	0.131	-	0.507
IndoorDepth _i	×	84.5	96.5	99.1	0.126	0.054	0.494
SC-DepthV3 ^[9]	×	84.8	96.3	99.1	0.123	0.053	0.486
本文	×	85.6	96.6	99.2	0.120	0.052	0.466

注:粗体表示最优值。

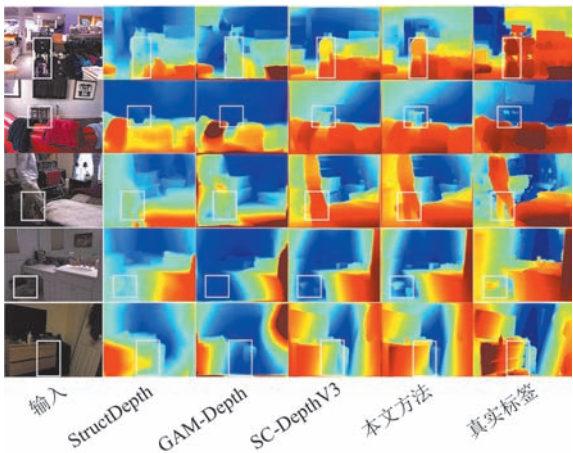


图7 在NYUv2数据集上4种方法的定性对比结果

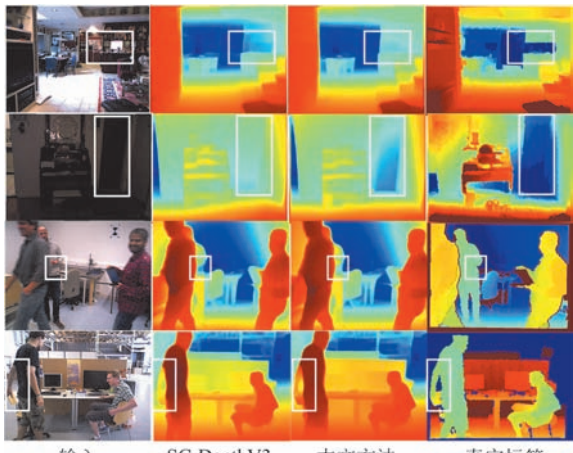


图8 在复杂场景中的定性对比结果

4.4 消融实验

4.4.1 FDMN和GECM

为验证FDMN和GECM的有效性,在BONN数据集上进行详细的消融实验。

如表5所示,当仅引入FDMN模块时,在动态区域, $\delta < 0.125$ 从0.720大幅提升至0.768(相对提

升4.8%),AbsRel从0.220显著降低至0.199;在完整图像上, $\delta < 0.125$ 从0.889提升至0.904(相对提升1.5%),AbsRel从0.126降低至0.121,静态区域的性能保持稳定。这表明FDMN通过光流建模动态物体的独立运动,在不影响静态区域的前提下,有效缓解了动态区域深度估计的歧义性问题。

表4 跨数据集零样本泛化性能评估

方法	完整图像					动态区域		静态区域	
	AbsRel	RMSE	$\delta < 0.125$	$\delta < 0.125^2$	$\delta < 0.125^3$	AbsRel	$\delta < 0.125$	AbsRel	$\delta < 0.125$
Monodepth2 ^[8]	0.578	2.568	0.257	0.458	0.624	0.906	0.140	0.483	0.293
SC-Depth ^[33]	0.457	1.892	0.394	0.671	0.841	0.608	0.360	0.392	0.416
SC-DepthV2 ^[32]	0.282	1.553	0.525	0.797	0.905	0.355	0.489	0.253	0.541
SC-DepthV3 ^[9]	0.204	1.646	0.706	0.856	0.888	0.264	0.542	0.189	0.746
本文	0.200	1.651	0.719	0.854	0.887	0.251	0.582	0.188	0.747

注:粗体表示最优值。

表5 BONN数据集上不同方法的消融实验结果

方法	FDMN	GECM	完整图像			动态区域		静态区域		复杂度		
			AbsRel	RMSE	$\delta < 0.125$	AbsRel	$\delta < 0.125$	AbsRel	$\delta < 0.125$	参数成本	计算量成本	耗时/ms
基线	×	×	0.126	0.379	0.889	0.220	0.720	0.102	0.931	14.33M	5.36G	7
	✓	×	0.121	0.376	0.904	0.199	0.768	0.102	0.935	36.43M	16.31G	20
本文	×	✓	0.120	0.378	0.904	0.199	0.762	0.102	0.935	14.55M	8.21G	15
	✓	✓	0.117	0.368	0.907	0.194	0.774	0.099	0.936	36.65M	19.16G	26

注:粗体表示最优值。

单独引入GECM时,在动态区域, $\delta < 0.125$ 从0.720提升至0.762(相对提升4.2%),AbsRel同样降至0.199;完整图像的 $\delta < 0.125$ 从0.889提升至0.904(相对提升1.5%),AbsRel降至0.120;静态区域性能保持稳定。这表明GECM通过注入几何边缘先验,有效补偿了运动模糊导致的边界信息退化,在整体和动态区域均带来性能提升。

当FDMN和GECM同时作用时,性能达到最

佳。在动态区域, $\delta < 0.125$ 达到0.774(相对提升7.5%),AbsRel降至0.194;在完整图像上, $\delta < 0.125$ 达到0.907(相对提升1.8%),AbsRel进一步降至0.117;各项指标均优于单独使用任一模块的结果,这表明动态解耦建模与边缘补偿机制形成有效互补,消除动态干扰的同时修复边界退化,协同作用下实现了最优性能。

本文完整方法的参数量为36.65M,计算量为

19.16 G FLOPs,推理耗时约 26 ms,所有测试均经过五次实验并取平均值。尽管各项计算开销有所增加,但推理速度仍可达到约 38 FPS,能够满足实时应用的需求。因此,性能与效率之间的权衡是合理且可接受的。图 9 展示了各模块在 Bonn 数据集上的训练损失曲线对比,可以观察到,虽然本文方法相较于基线在训练过程中有轻微震荡,但仍呈现了较好的收敛趋势,最终收敛至与基线持平的损失,并取得了更好的深度估计精度。这表明 FDMN 和 GECM 模块通过解耦动态物体运动和补偿边缘信息,有效提高了模型对室内动态场景的理解能力,而非简单地降低训练损失。

4.4.2 GECM模块部署策略比较

为确定 GECM 模块的最优架构配置,评估了在深度解码器不同层级中嵌入不同数量 GECM 模块的效果。由于 GECM 模块需要更深层的深度预测信息作为引导,其最大可部署层级为 4 层。如表 6 所示,测试了 7 种不同的 GECM 嵌入策略。实验结果表明:单层嵌入时,尽管第 4 层嵌入 GECM 在完整

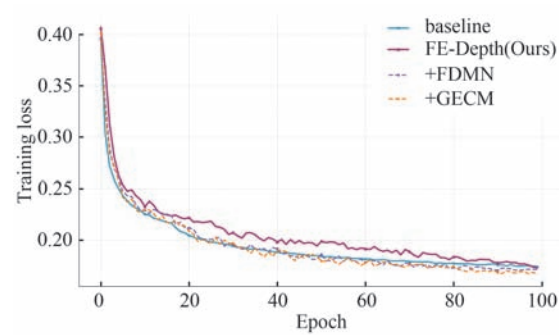


图9 各模块在 Bonn 数据集上的训练损失曲线

图像上取得了相对较好的 RMSE(0.375),但动态区域性能提升较低($\delta < 0.125$ 仅为 0.759);在浅层(1~2,1~3)嵌入 GECM 能够捕获更多细节信息,却易受局部噪声干扰导致性能波动。本文采用的全层级嵌入策略(1-4)综合了各层级的优势,在动态区域达到了最佳的 $\delta < 0.125$ (0.762)。浅层 GECM 捕获精细的边缘细节,深层 GECM 提供全局的结构约束,通过多尺度特征融合实现了最优的边缘补偿效果。

表6 不同几何边缘补偿模块数量的模型结果对比

嵌入层数	完整图像			动态区域		静态区域	
	AbsRel	RMSE	$\delta < 0.125$	AbsRel	$\delta < 0.125$	AbsRel	$\delta < 0.125$
1	0.123	0.392	0.895	0.199	0.727	0.107	0.928
2	0.120	0.384	0.892	0.201	0.730	0.104	0.924
3	0.124	0.393	0.896	0.200	0.741	0.108	0.928
4	0.122	0.375	0.901	0.207	0.759	0.101	0.933
1-2	0.124	0.392	0.894	0.202	0.730	0.107	0.927
1-3	0.122	0.387	0.900	0.204	0.747	0.105	0.931
1-4	0.120	0.378	0.904	0.199	0.762	0.102	0.935

注:粗体表示最优值。

4.4.3 静态流约束损失

静态流约束损失 $\mathcal{L}_f$ 是确保 FDMN 正确进行运动分解的关键约束。表 7 展示了不同权重 λ_f 对性能的影响。实验结果表明,在无约束情况下($0.0\lambda_f$),网络缺乏对静态区域运动一致性的约束,动态区域 $\delta < 0.125$ 仅为 0.725。权重过小时,静态流一致性约束作用有限,动态区域 $\delta < 0.125$ 仅为 0.737,权重过大会过度限制网络的运动建模能力,影响对动态物体独立运动的捕捉,导致动态区域性能下降;当 $\lambda_f=5.0$ 时达到最优权重,静态流约束损失有效约束了静态区域的运动一致性,确保静态背景的运动完全由相机姿态变化解释。动态区域 $\delta < 0.125$ 达到最高的 0.768, AbsRel 为 0.199,实现了运动分解的最佳平衡。

4.4.4 HFC与MMN

为明确 FDMN 的核心组件 HFC 和 MMN 对于网络的作用,进行消融实验。HFC 负责估计场景中所有像素的三维运动矢量,MMN 通过自适应权重分配区分刚体运动与非刚性运动分量。表 8 展示了消融实验结果,结果表明,当 HFC 单独作用时,相比基线方法,在动态区域 $\delta < 0.125$ 从 0.720 相对提升至 0.742, AbsRel 从 0.220 降低至 0.215;在完整图像上, $\delta < 0.125$ 从 0.889 相对提升至 0.895,表明 HFC 提升了网络对于动态像素的理解能力;引入 MMN 后,网络在所有关键指标上均达到最优:在动态区域的 $\delta < 0.125$ 从 0.742 相对提升至 0.768, AbsRel 从降低至 0.199,在完整图像的评估上, $\delta < 0.125$ 从 0.895 相对提升至 0.904。完整光流场与运动掩码的

表 7 静态流约束损失超参数影响							
方法	完整图像			动态区域		静态区域	
	AbsRel	RMSE	$\delta < 0.125$	AbsRel	$\delta < 0.125$	AbsRel	$\delta < 0.125$
0. $0\mathcal{L}_{fc}$	0.122	0.373	0.899	0.214	0.725	0.100	0.937
1. $0\mathcal{L}_{fc}$	0.123	0.392	0.892	0.197	0.737	0.107	0.923
3. $0\mathcal{L}_{fc}$	0.124	0.373	0.902	0.217	0.742	0.102	0.938
5. $0\mathcal{L}_{fc}$	0.121	0.376	0.904	0.199	0.768	0.102	0.935
7. $0\mathcal{L}_{fc}$	0.121	0.374	0.899	0.206	0.746	0.102	0.931

注:粗体表示最优值。

表 8 HFC与MMN消融实验结果									
方法	HFC	MMN	完整图像			动态区域		静态区域	
			AbsRel	RMSE	$\delta < 0.125$	AbsRel	$\delta < 0.125$	AbsRel	$\delta < 0.125$
基线	×	×	0.126	0.379	0.889	0.220	0.720	0.102	0.931
本文	✓	×	0.125	0.380	0.895	0.215	0.742	0.103	0.929
	✓	✓	0.121	0.376	0.904	0.199	0.768	0.102	0.935

注:粗体表示最优值。

协同作用实现了对刚性运动与独立运动的精确建模,有效提升了对动态区域深度估计的准确性。

4.4.5 骨干网络选型

为验证 ResNet18 作为 FDMN 模块编码器的合理性,本节在 BONN 数据集上对比了 4 种不同深度的 ResNet 网络^[42],实验结果如表 9 所示。实验结果表明,四种骨干网络在完整图像上的性能接近,AbsRel 指标在 0.117~0.121 之间,差异小于 5%,表明 ResNet18 已能提供充足的特征表达能力。在动

态区域,ResNet18 表现最优,AbsRel 为 0.194,优于更深的网络,说明过深网络在处理运动目标时可能存在过拟合风险。在静态区域,各网络性能较为一致, $\delta < 0.125$ 指标均在 0.934~0.941 之间。从参数效率来看,ResNet18 的参数量(11.7M)仅为 ResNet101(44.5M)的 26%,但性能相当甚至更优。综合考虑性能和效率,本文选择 ResNet18 作为骨干网络,在保证深度估计精度的同时大幅降低了模型复杂度和计算成本,更适合实际应用部署。

表 9 BONN 数据集上不同骨干网络的消融实验结果								
骨干网络	参数量	完整图像			动态区域		静态区域	
		AbsRel	RMSE	$\delta < 0.125$	AbsRel	$\delta < 0.125$	AbsRel	$\delta < 0.125$
ResNet18	11.7 M	0.117	0.368	0.907	0.194	0.774	0.099	0.936
ResNet34	21.8M	0.121	0.372	0.903	0.204	0.760	0.101	0.936
ResNet50	25.6M	0.121	0.375	0.901	0.207	0.760	0.101	0.934
ResNet101	44.5M	0.119	0.365	0.906	0.210	0.751	0.098	0.941

注:粗体表示最优值。

5 结束语

本文提出一种基于光流解耦和边缘补偿的室内动态场景自监督单目深度估计方法,通过光流解耦建模网络准确建模动态物体运动,有效分离静态背景与动态前景,将破坏光度一致性的运动像素转化为可学习的实体;设计几何边缘补偿模块,引入拉普拉斯算子提取运动模糊不变的几何边缘先验,缓解由于运动模糊产生的边界退化现象。在 Bonn、

TUM、NYUv2 等室内数据集上均取得了显著的性能提升,特别是在动态区域的深度估计准确率方面表现优异。但在处理高度复杂的多物体交互运动场景以及强运动模糊时仍存在局限性。未来工作将探索基于光流的更加鲁棒的多物体运动建模框架,将其应用于更复杂的室内场景中。

参 考 文 献

[1] Cheng De-Qiang, Qian Jian-Sheng, Guo Xing-Ge, et al. Review on key technologies of AI recognition for videos in coal

- mine. Coal Science and Technology, 2023, 51(2): 349-365 (in Chinese)
(程德强, 钱建生, 郭星歌, 等. 煤矿安全生产视频 AI 识别关键技术综述. 煤炭科学技术, 2023, 51(2): 349-365)
- [2] Lyu Chen, Han Cheng-Gong, Yu Ting, et al. Multi-domain structural information guided indoor self-supervised depth estimation. Journal of Computer-Aided Design & Computer Graphics, to appear (in Chinese)
(吕晨, 韩成功, 余婷, 等. 多域结构信息引导的室内自监督深度估计. 计算机辅助设计与图形学学报, 待发表)
- [3] Jiang Rong-Kun, Gai Shao-Yan, Da Fei-Peng. Self-supervised monocular depth estimation with multi-single attention and scale refinement. Journal of Beijing University of Aeronautics and Astronautics, 2024: 1-12 (in Chinese)
(蒋容坤, 盖绍彦, 达飞鹏. 基于注意力机制和尺度优化的自监督单目深度估计. 北京航空航天大学学报, 2024: 1-12)
- [4] Ai Hao-Jun, Zhang Feng, Lyu Peng-Fei, et al. Improving self-supervised monocular indoor depth estimation with local feature guidance. Journal of Computer Research and Development, 2026, 63(2):338-351 (in Chinese)
(艾浩军, 张锋, 吕鹏飞, 等. 局部特征引导的室内自监督单目深度估计方法. 计算机研究与发展, 2026, 63(2):338-351)
- [5] Cheng De-Qiang, Xu Shuai, Lyu Chen, Han Cheng-Gong, Jiang He and Kou Qiqi. Lightweight self-supervised monocular depth estimation method with enhanced direction-aware. Journal of Electronics and Information Technology, 2024, 46(9): 3683-3692 (in Chinese)
(程德强, 徐帅, 吕晨, 韩成功, 江鹤, 寇旗旗. 方向感知增强的轻量级自监督单目深度估计方法. 电子与信息学报, 2024, 46(9): 3683-3692)
- [6] Han C, Lv C, Huang X, Kou Q, Cheng D, Jiang H. PRDepth: pose refinement enhancement-based monocular depth estimation for indoor scenes. IEEE Transactions on Instrumentation and Measurement, 2025, 74(3): 1-16
- [7] Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A. 3D packing for self-supervised monocular depth estimation//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 2482-2491
- [8] Godard, C., Aodha, O. M., Firman, M., et al. Digging into self-supervised monocular depth estimation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, the republic of Korea, 2019: 3827-3837
- [9] Sun L B, Bian J W, Shen R C. SC-DepthV3: robust self-supervised monocular depth estimation for dynamic scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(1): 497-508
- [10] Zhou K, et al. Manydepth2: motion-aware self-supervised monocular depth estimation in dynamic scenes. IEEE Robotics and Automation Letters, 2025, 10(7): 6704-6711
- [11] Moon J, Bello J L G, Kwon B, et al. From-ground-to-objects: coarse-to-fine self-supervised monocular depth estimation of dynamic objects with ground contact prior//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 10519-10529
- [12] Hu H, Feng Y, Li D, et al. Monocular depth estimation via self-supervised self-distillation. Sensors, 2024, 24(13): 4090
- [13] Zhao B, He H, Xu H, et al. RTIA-Mono: real-time lightweight self-supervised monocular depth estimation with global-local information aggregation. Engineering Applications of Artificial Intelligence, 2025, 156: 104769
- [14] Ji P, Li R, Bhanu B, et al. Monoindoor: towards good practice of self-supervised monocular depth estimation for indoor environments//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 12787-12796
- [15] Guo, X., Zhao, H., Shao, S., et al. F2Depth: self-supervised indoor monocular depth estimation via optical flow consistency and feature map synthesis. Engineering Applications of Artificial Intelligence, 2024, 133: 108391
- [16] Lv C, Han C, Song T, et al. MMAindoor: patched MLP and multi-dimensional cross attention based self-supervised indoor depth estimation. Neurocomputing, 2024, 596: 127972
- [17] KOU Qiqi, WANG Weichen, HAN Chenggong, et al. Multi-frame self-supervised monocular depth estimation with multi-scale feature enhancement. Optics and Precision Engineering, 2024, 32(24): 3603-3615 (in Chinese)
(寇旗旗, 王伟臣, 韩成功等. 多尺度特征增强的多帧自监督单目深度估计. 光学精密工程, 2024, 32(24): 3603-3615)
- [18] Klingner M, Termöhlen J A, Mikolajczyk J, et al. Self-supervised monocular depth estimation: solving the dynamic object problem by semantic guidance//Vedaldi A, Bischof H, Brox T, et al. Computer Vision - ECCV 2020. Cham: Springer International Publishing, 2020: 582-600
- [19] Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 12179-12188
- [20] Liu R, Mi L, Chen Z. AFNet: adaptive fusion network for remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(9): 7871-7886
- [21] Han C, Lv C, Kou Q, et al. DCL-depth: monocular depth estimation network based on iam and depth consistency loss. Multimedia Tools and Applications, 2025, 84(8): 4773-4787
- [22] Li B, Huang Y, Liu Z, et al. StructDepth: leveraging the structural regularities for self-supervised indoor depth estimation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 12663-12673
- [23] Zou Y, Luo Z, Huang J B. DF-Net: unsupervised joint learning of depth and flow using cross-task consistency//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 36-53
- [24] Nguyen H C, Wang T, Alvarez J M, et al. Mining supervision for dynamic regions in self-supervised monocular depth estimation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA,

- 2024: 10446-10455
- [25] Saxena S, Herrmann C, Hur J, et al. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 2023, 36: 39443-39469
- [26] Gui M, Schusterbauer J, Prestel U, et al. DepthFM: fast generative monocular depth estimation with flow matching// *Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 2025, 39(3): 3203-3211
- [27] Jung D, Choi J, Lee Y, et al. DnD: dense depth estimation in crowded dynamic indoor scenes//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal, Canada, 2021: 12797-12807
- [28] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module//*Proceedings of the European Conference on Computer Vision*. Munich, Germany, 2018: 3-19
- [29] Palazzolo E, Behley J, Lottes P, et al. ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals//*Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. Macao, China, 2019: 7855-7862
- [30] Sturm J, Engelhard N, Endres F, et al. A benchmark for the evaluation of RGB-D SLAM systems//*Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. Vilamoura, Portugal, 2012: 573-580
- [31] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from RGBD images//*Fitzgibbon A, Lazebnik S, Perona P, et al. Computer Vision - ECCV 2012*. Berlin, Germany: Springer, 2012: 746-760
- [32] Bian J W, Zhan H, Wang N, et al. Auto-rectify network for unsupervised indoor depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 44(12): 9802-9813
- [33] Bian J, Li Z, Wang N, et al. Unsupervised scale-consistent depth and ego-motion learning from monocular video// *Proceedings of the Annual Conference on Neural Information Processing Systems*. Vancouver, Canada, 2019
- [34] Cheng, A., Yang, Z., Zhu, H., Mao, K. GAM-Depth: self-supervised indoor depth estimation leveraging a gradient-aware mask and semantic constraints//*Proceedings of the 2024 IEEE International Conference on Robotics and Automation*. Yokohama, Japan, 2024: 5367-5374
- [35] Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D. Deep ordinal regression network for monocular depth estimation//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City, USA, 2018: 2002-2011
- [36] Hu J, Ozay M, Zhang Y, et al. Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries//*Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision*. Waikoloa, USA, 2019: 1043-1051
- [37] Li, Z., Chen, Z., Liu, X., et al. DepthFormer: exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, 2023, 20(6): 837-854
- [38] Yu Z, Jin L, Gao S. P2net: patch-match and plane-regularization for unsupervised indoor depth estimation// *Vedaldi A, Bischof H, Brox T, et al. Computer Vision-ECCV 2020*. Cham: Springer, 2020: 206-222
- [39] Wei, Y., Guo, H., Lu, J., Zhou, J. Iterative feature matching for self-supervised indoor depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022, 32(6): 3839-3852
- [40] Wu C Y, Wang J, Hall M, Neumann U, Su S. Toward practical monocular indoor depth estimation//*Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, USA, 2022: 3804-3814
- [41] Li, R., Ji, P., Xu, Y., Bhanu, B. MonoIndoor++: towards better practice of self-supervised monocular depth estimation for indoor environments. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, 33(2): 830-846
- [42] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, USA, 2016: 770-778



CHENG De-Qiang, Ph. D., professor. His main research areas include image processing and machine vision.

KOU Qi-Qi, Ph. D., associate professor. His main research area is pattern recognition.

ZHANG Pei-Yuan, M. S. candidate. His main research area is depth estimation.

CUI Wen, M. S. candidate. His main research area is image restoration.

HAN Cheng-Gong, Ph. D. candidate. His main research area is depth estimation.

JIANG He, Ph. D., assistant professor. His main research area is image restoration.

Background

Self-supervised monocular depth estimation represents a fundamental problem at the intersection of computer vision and machine learning, aiming to accurately acquire scene depth information using only monocular video sequences. In outdoor driving scenarios, the presence of dynamic objects violates multi-view geometric constraints, leading to photometric consistency failures. Internationally, methods such as Monodepth2 and PackNet-SfM have partially mitigated this issue through dynamic masking and other strategies, enabling outdoor scene depth estimation to reach practical-level accuracy.

However, depth estimation in indoor dynamic scenes faces even more severe challenges. The confined nature of indoor spaces results in complex motion patterns of dynamic objects with close-range imaging, where motion pixels exhibit characteristics of extensive disordered and drastic movement. Currently, specialized methods targeting indoor dynamic scenes are relatively scarce internationally. While existing approaches such as SC-DepthV3 and MonoIndoor have demonstrated improvements, they still possess limitations in handling complex dynamic motion and boundary preservation, with depth

estimation accuracy in dynamic regions generally remaining relatively low.

This paper proposes FE-Depth, a method based on optical flow decoupling and edge compensation that explicitly models dynamic object motion while introducing geometric edge priors, elevating indoor dynamic scene depth estimation accuracy to unprecedented levels. In dynamic regions of the Bonn and TUM datasets, our method achieves $\delta < 0.125$ accuracies of 0.774 and 0.807 respectively, significantly surpassing existing state-of-the-art approaches.

This research is supported by the Yunlong Lake Laboratory for Deep Earth Science and Engineering, the Fundamental Research Funds for the Central Universities, and the National Natural Science Foundation of China, holding significant importance for advancing intelligent perception technologies in complex environments. Our research team has previously published multiple high-impact papers in the depth estimation domain, accumulating extensive experience in multi-scale feature enhancement, direction-aware enhancement, and pose optimization, thereby establishing a solid foundation for this study.