

DOI: 10.19650/j.cnki.cjsi.J2615114

基于改进 Lite-Mono 的水下单目深度估计方法^{*}

傅荟璇, 王 炜, 王宇超

(哈尔滨工程大学智能科学与工程学院 哈尔滨 150001)

摘 要:水下无人航行器的自主对接回收是其长期自主作业的关键环节,精确的水下环境感知和深度信息的实时获取是实现该任务的核心基础。然而现有的轻量化单目深度估计网络,在复杂水下场景中存在特征表达能力不足和边缘细节模糊的问题,为此提出了一种面向水下场景的改进自监督单目深度估计方法。该方法以 Lite-Mono 为基础,设计了自适应解耦特征增强卷积模块,通过构建结构化特征增强机制,提升网络对水下场景多尺度特征的表达能力与边缘细节的感知精度。引入 Depth Anything V2 预训练模型生成高质量的伪标签辅助自监督训练,弥补水下真实标注数据不足的短板,并针对水下光照衰减与颜色偏移等场景特性对损失函数进行优化,设计了深度监督损失与边界加权损失,加强空间约束和边缘清晰度。以水下数据集 FLSea 为基准,通过消融实验验证各改进模块的贡献,结果表明,所提方法在水下深度估计精度上较基线 Lite-Mono 的绝对相对误差降低了 29.3%,精度指标提升了 27.9%,且边缘细节恢复效果更好,对复杂水下干扰的鲁棒性更强,验证了所提方法在水下场景单目深度估计的有效性与优越性。

关键词: 水下无人航行器;水下深度估计;自监督学习;特征增强;Lite-Mono

中图分类号: TH39 **文献标识码:** A **国家标准学科分类代码:** 520.20

Underwater monocular depth estimation method based on improved Lite-Mono

Fu Huixuan, Wang Wei, Wang Yuchao

(College of Intelligent Systems Science and Engineering, Harbin Engineering University, Harbin 150001, China)

Abstract: Autonomous docking and recovery of unmanned underwater vehicles is a key link for their long-term autonomous operation. Accurate perception of the underwater environment and real-time acquisition of depth information constitute the core foundation for accomplishing this task. However, existing lightweight monocular depth estimation networks suffer from insufficient feature representation capability and blurred edge details in complex underwater scenes. To address these issues, this paper proposes an improved self-supervised monocular depth estimation method tailored for underwater scenarios was proposed. Based on the Lite-Mono network, an adaptive decoupled feature enhancement convolution module is designed. By constructing a structured feature enhancement mechanism, this module improves the network's multi-scale feature representation capability and edge detail perception accuracy in underwater scenes. The Depth Anything V2 pre-trained model is introduced to generate high-quality pseudo-labels to assist self-supervised training, so as to make up for the shortage of real labeled data in underwater scenarios. In addition, the loss function is optimized according to the scene characteristics such as underwater light attenuation and color distortion. A depth-supervised loss and a boundary-weighted loss are designed to strengthen spatial constraints and improve edge sharpness. Based on the underwater dataset FLSea, ablation experiments are conducted to verify the effectiveness of each improved module. The results show that, compared with the baseline Lite-Mono, the proposed method reduces the absolute relative error metric for underwater depth estimation accuracy by 29.3% and improves the accuracy metric by 27.9%. In addition, it achieves better recovery of edge details and stronger robustness against complex underwater interferences, validating the effectiveness and superiority of the proposed method for monocular depth estimation in underwater scenarios.

Keywords: UUV; underwater depth estimation; self-supervised learning; feature enhancement; Lite-Mono

0 引言

近年来,人工智能、深度学习与计算机视觉技术的快速发展,推动了水下无人航行器(unmanned underwater vehicle, UUV)的自主作业能力提升^[1]。水下无人航行器对接回收过程中,需要对周围环境进行实时感知以获取目标的位置与距离信息,精确的深度估计是完成该任务的前提。单目深度估计是场景三维感知的重要手段,依靠单目相机获取场景深度信息,具有成本低、部署方便等显著优势,在水下无人航行器定位^[2]、水下图像复原^[3]与水下目标检测^[4]等任务中具有重要的应用价值。然而,水下环境中光照衰减、颜色失真、悬浮粒子散射等复杂因素^[5],使得轻量化网络在特征表达^[6]与边缘重建上面面临挑战^[7],成为当前水下视觉感知领域的研究热点之一。

自监督单目深度估计方法是常用的水下深度估计方式,其利用连续图像帧之间的几何一致性构建重投影损失,不需要真实深度标签,能实现端到端的训练,具有较强的实用性。Zhou等^[8]提出了基于图像重投影损失的自监督单目深度估计框架,奠定了后续方法的基础训练范式。Yin等^[9]提出联合无监督框架 GeoNet,通过级联的刚性结构重建器与非刚性运动定位器,分别估计静态场景深度与相机位姿,并补偿动态物体的残余光流。Mahjourian等^[10]引入三维点云几何约束损失进一步提升了跨视角深度一致性。Johnston等^[11]则提出自注意力机制结合离散视差体结构以保留深度图的几何细节。Yan等^[12]在深度估计网络中引入通道注意力机制,通过结构感知模块捕获通道间的长距离依赖以增强场景结构理解,并设计细节强调模块突出物体边缘局部信息。Zhou等^[13]提出自蒸馏特征聚合模块,通过多分支学习特征偏移图,在自蒸馏方式下保持多尺度特征间的上下文一致性。HR-Depth通过新型跳跃连接与特征融合模块学习更具判别力的高分辨率深度特征^[14]。MonoViT将视觉Transformer与卷积网络结合以融入全局信息提升深度预测精度^[15]。Godard等^[16]提出左右一致性损失并设计自动掩码机制有效滤除静止区域的无效监督信号。谢昭等^[17]提出采样汇集网络(sampling aggregate network, SAN),通过尺度特征汇集与采样跨层汇集策略有效改善了复杂物体边界处的深度估计模糊问题。

水下环境的特殊性使得上述面向陆地场景设计的深度估计方法难以直接迁移应用。由于水体对光的吸收和散射导致图像质量下降,相邻帧之间的成像差异较大,破坏了自监督方法依赖的光度一致性假设。并且,水下有标注深度的数据集规模小、场景单一,导致了模型泛化能力较弱。

针对上述问题,水下深度估计有众多研究。Amitai等^[18]将自监督学习框架引入水下单目深度估计,验证重投影损失在水下环境的有效性,并针对水体背景深度误判问题提出局部掩码策略。Yu等^[19]通过引入水下光衰减特性的多通道输入空间,结合轻量级骨干网络与Transformer模块,实现水下单目深度估计。Zhang等^[20]利用 Stable Diffusion 与 ControlNet 生成高质量水下图像,构建三元组训练数据以缓解水下标注数据匮乏问题。在此基础上,Gu等^[21]提出结合多视图深度估计与图像增强的数据集构建方法,进一步丰富了适用于水下场景的监督训练数据。Petrovai等^[22]在自监督框架中引入伪标签辅助训练策略,验证了伪标签监督对深度估计精度的提升作用。Watson等^[23]利用半全局匹配生成深度提示作为额外监督信号,有效缓解了自监督信号不稳定的问题。Saunders等^[24]针对水下光散射、色偏等问题设计了鲁棒的自监督深度估计方法,为复杂场景适应提供了新思路。

尽管上述方法取得了一定进展,但轻量化网络在复杂水下场景中仍存在特征表达能力不足、边缘细节模糊及多尺度建模能力受限等问题,难以满足实际水下作业的高精度感知需求。

为此,本文以 Lite-Mono^[25]网络为基础,提出一种面向复杂水下场景的自监督单目深度估计方法。设计自适应解耦特征增强卷积模块(adaptive decoupled feature enhancement conv, ADFEC),构建结构化特征增强机制,以提升多尺度特征的表达和边缘感知能力。同时引入 Depth Anything V2^[26]预训练模型,生成高质量的伪标签辅助训练。针对水下光照衰减与颜色偏移特性,设计深度监督损失与边界加权损失,对损失函数进行了优化。实验结果验证了所提方法的有效性与优越性。

1 Lite-Mono 深度估计网络

1.1 Lite-Mono 网络模型

Lite-Mono 是一种面向自监督单目深度估计的轻量化混合架构,采用单目图像序列训练数据集和基础训练策略的深度估计方法,其由深度估计网络(depth estimation network, DepthNet)和位姿估计网络(pose estimation network, PoseNet)两部分组成。

在网络架构中,DepthNet 采用编码器-解码器架构:编码器包含4个阶段,其中第一阶段为卷积前置模块,由一个步长为2的3×3卷积和两个步长为1的3×3卷积串联,用于浅层特征提取。Concat表示沿通道维度的拼接操作,该机制使各阶段在下采样过程中始终保留来自浅层的细粒度语义信息,有效缓解了深层特征中空间细节丢失的问题。深度解码器末端采用S型激活函数 Sigmoid 约束网络原始输出。后3个阶段则依次通过连

续扩张的卷积模块 (consecutive dilated convolutions, CDC), CDC 用来提取多尺度局部特征。同时, Lite-Mono 结合局部-全局特征交互模块 (local-global features interaction, LGFI), 引入自注意力机制, 用于获取长距离全局的下文信息。解码器则通过双线性上采样逐步恢复

空间分辨率, 并融合编码器各阶段的跳跃连接特征, 最终输出多尺度逆深度图。PoseNet 以预训练的 ResNet-18 为骨干网络, 以连续图像序列的相邻帧对作为输入, 输出结果用于估计帧与帧之间的六自由度相对位姿。其结构如图 1 所示。

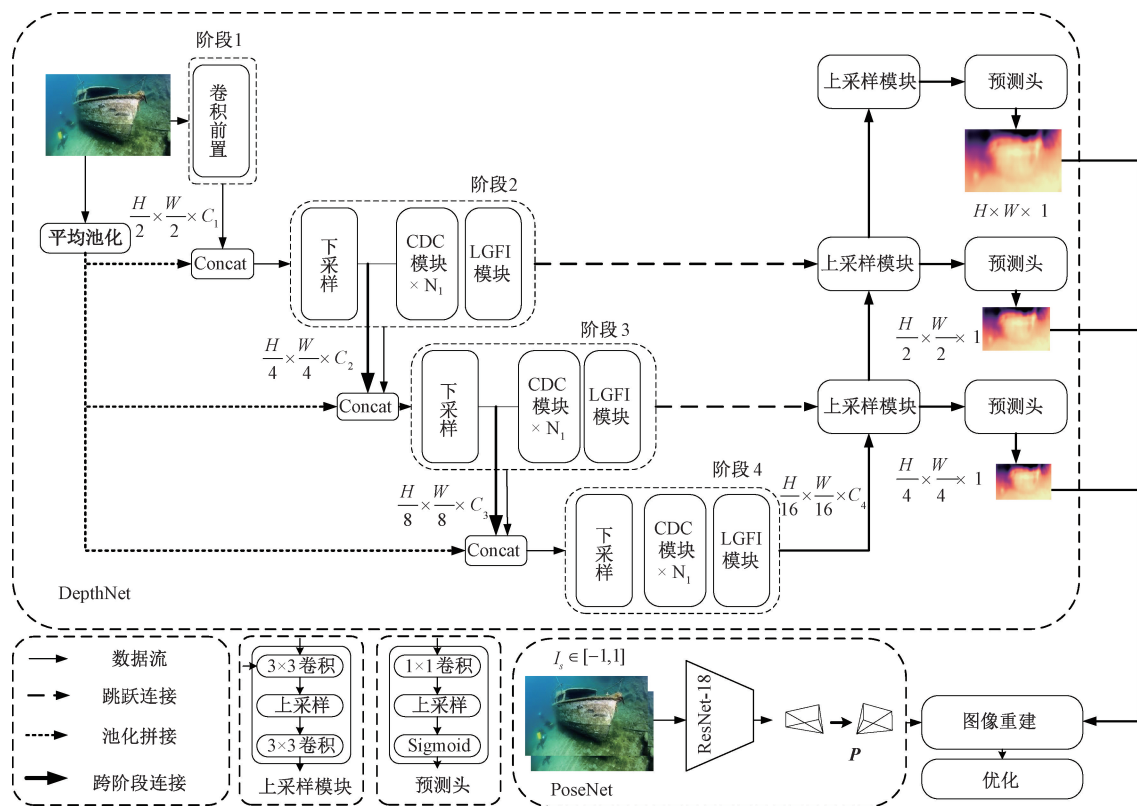


图 1 Lite-Mono 网络结构

Fig. 1 Lite-Mono network architecture

1.2 Lite-Mono 模型特性与水下适用性分析

Lite-Mono 标准模型的可训练参数量仅为 3.1 M, 模型规模较小, 计算复杂度低, 对存储资源和算力要求较低。单帧推理时间约为 4.5 ms, 响应速度快, 具备较高的运行效率与良好的实时处理能力^[25]。网络在保证深度估计精度的同时也能满足系统实时性要求。因此 Lite-Mono 模型符合在嵌入式设备等有限计算平台上的在线运行需求, 适合水下航行器执行水下对接回收等任务。

Lite-Mono 以自动驾驶场景下的 KITTI 数据集^[27]为基准, 其训练与优化均面向空气介质中的自然光照条件。所以该架构在应用于复杂的水下环境时会具有天然的缺陷。水下图像普遍存在颜色偏移、非均匀的光照衰减、显著的散射噪声以及对对比度低等问题, 而 Lite-Mono 的视图合成高度依赖于像素级的光度一致性假设, 网络容易受到诸多水下图像退化条件的影响, 导致其约束效果减弱。并且, 水体中悬浮粒子产生的前向散射会破坏局部纹理

结构, 进而影响 CDC 模块对多尺度局部特征的有效提取。因此, 直接将 Lite-Mono 迁移至水下深度估计任务, 会导致模型性能显著下降。基线 Lite-Mono 在水下图像的表现如图 2 所示。



图 2 Lite-Mono 的水下图像输出

Fig. 2 Underwater image output of Lite-Mono

图2中,第1列为原始图像,第2列为 Lite-Mono 输出的深度图。从图2中可以看出,基线 Lite-Mono 在水下场景中输出的深度图整体结构模糊,近景与远景的深度层次区分不足,目标边界处存在明显的深度扩散现象,难以准确反映真实场景的空间结构。

在 UUV 对接回收的近距离操作阶段,目标轮廓的深度边界精度直接决定定位姿估计与导引控制的可靠性,因此提升水下场景中深度估计的边缘细节感知能力具有重要的实际意义。

2 自适应解耦特征增强模块

ADFEC 模块主要由改进后的解耦-聚合卷积层(D decoupling-and-aggregating convolution, DAConv)^[28]、高斯误差线性单元激活函数(gaussian error linear unit, GELU)、归一化层以及倒残差瓶颈结构组成,并且通过残差连接把输入特征和卷积变换后的特征进行融合,在保留原始信息的同时缓解深层网络的梯度消失的问题,提高网络训练稳定性。

为了实现对水下图像关键结构信息的有效表达,本文引入自适应解耦特征增强模块,其整体结构及其嵌入位置如图3所示。

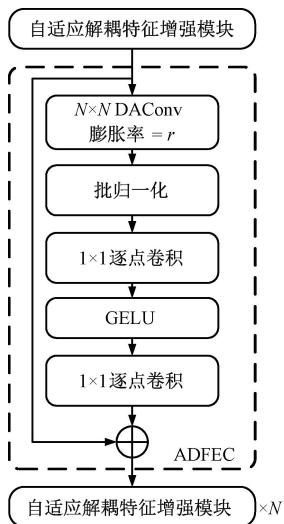


图3 自适应解耦特征增强模块结构

Fig. 3 Structure of the ADFEC

该模块将解耦-聚合卷积结构与倒残差瓶颈设计相结合,在保持较低的计算复杂度的同时,增强了网络对于多尺度信息的感知能力,从而提高网络在水下退化条件下对特征信息的提取效率。

在 ADFEC 中以 3×3 解耦-聚合卷积模块作为核心的特征提取单元。将该卷积优化为可调节膨胀率的卷积操作(图4),在扩大了感受野的同时提高了网络对不同

尺度特征的建模能力。基于局部平滑性假设,卷积过程分为两条并行分支:对比度感知单元(contrast aware, CA)和细节感知单元(detail aware, DA)。

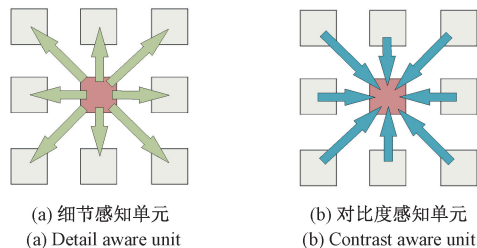


图4 可变膨胀率的 DA 和 CA 卷积核

Fig. 4 DA and CA convolution kernels with variable dilation rates

假设中心像素为 p_c , 亮度为 x , 则其他位置的像素亮度可表示为中心像素亮度与偏差值 n_i 的和:

$$x(p_i) = x(p_c) + n_i \quad (1)$$

根据解耦-聚合的原理,可将其整理为:

$$y_h(p_c) = \sum_{p_i \in R} w(p_i) \cdot (x(p_i) - x(p_c)) = \sum_{p_i \in R} w(p_i) \cdot n_i \quad (2)$$

$$y_l(p_c) = \sum_{p_i \in R} w(p_i) \cdot (x(p_i) + x(p_c)) \quad (3)$$

式中: $y_h(p_c)$ 表示中心像素处的高频特征响应; $y_l(p_c)$ 表示低频特征响应; $w(p_i)$ 为邻域像素 p_i 的聚合权重; R 为卷积核的采样邻域范围; $x(p_i)$ 与 $x(p_c)$ 分别表示邻域像素与中心像素的特征值; n_i 为邻域像素相对于中心像素的偏差值。

在以 Lite-Mono 为基础架构的网络中,ADFEC 模块主要用于编码器阶段的多尺度特征提取,如图5所示。

网络包含3个主要编码阶段,每个阶段由多个 ADFEC 模块堆叠而成。本文根据实际实验,将 ADFEC 在网络中的模块数量配置成[4, 4, 10],对应的通道数依次为[48, 80, 128]。逐层递增的结构使得网络可以在浅层阶段提取丰富的局部特征信息,而在深层阶段获取到更高维度的语义特征。

在每个编码阶段的末端引入 LGFI 模块用于全局上下文建模。ADFEC 模块负责局部特征提取,而 LGFI 模块通过注意力机制实现全局特征交互,形成了局部-全局协同的特征提取与表示的结构。

3 伪标签辅助自监督训练策略

3.1 总体思路与伪标签生成

纯自监督的单目深度估计方法通常依赖相邻帧之间的光度一致性约束进行训练,其核心假设是,当前场景所

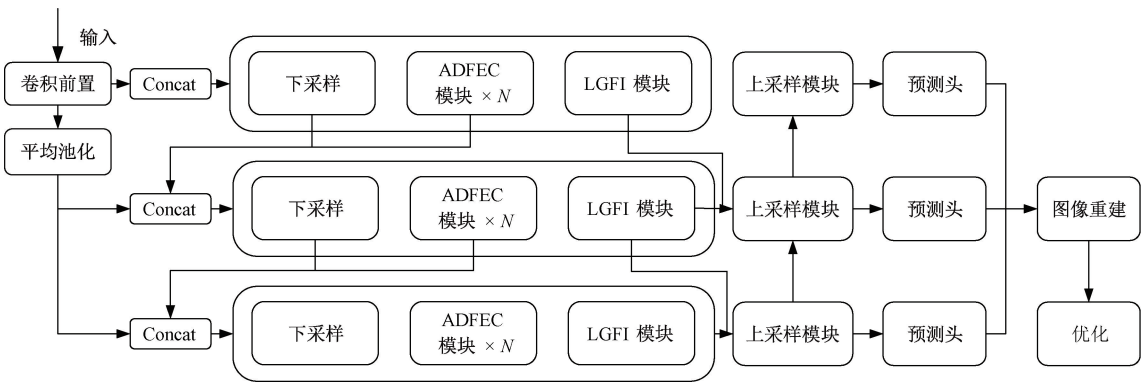


图 5 嵌入 ADFEC 模块的 Lite-Mono 网络结构

Fig. 5 Lite-Mono network structure embedded with the ADFEC module

有像素满足朗伯反射模型^[29]。然而在复杂多变的水下环境中,上述假设由于水体散射等因素影响不能成立,自监督信号的可靠性会受到明显的影响。

为弥补水下环境中自监督信号的不稳定性,本文采用一种伪标签辅助的自监督网络训练策略,利用预训练深度估计模型生成的伪标签,为深度估计网络模型提供额外的几何结构约束,提升模型在复杂水下环境中的深度估计能力,总体思路如图 6 所示。

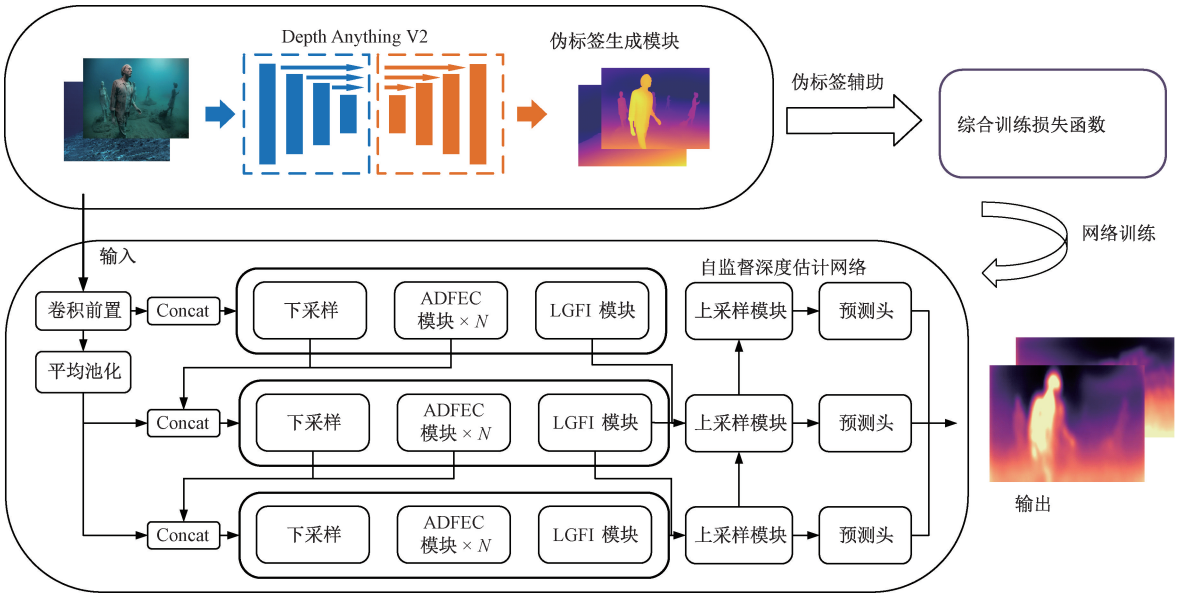


图 6 伪标签辅助自监督训练策略

Fig. 6 Pseudo-label assisted self-supervised training strategy

在伪标签生成阶段,本文采用离线生成策略。利用预训练深度估计模型对 FLSea 数据集^[30]中的训练图像进行批量处理,并且生成对应的深度图预测伪标签,建立伪标签与 RGB 图像一一对应的关系,在训练阶段与图像数据同步读取到 Lite-Mono 中。图 7 为 FLSea 数据集部分训练样本及其对应的伪深度标签可视化结果。

图 7 第 1 行为原始水下 RGB 图像。第 2 行为经 SeaErra 软件增强处理后的图像。第 3 行为真实深度图。第 4 行为 Depth Anything V2 对原始图像生成的伪深度标签,经尺度归一化处理后以热力图形式可视化。

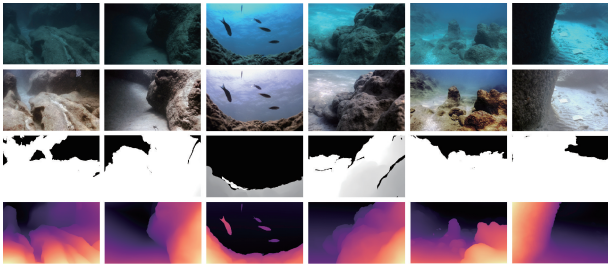


图 7 FLSea 数据集及其伪标签

Fig. 7 FLSea dataset and its pseudo-labels

与在线生成方式相比,离线生成策略具有更高的训练效率和更好的稳定性。考虑到本文提出的深度估计方法最终需要部署在水下无人航行器平台上,嵌入式设备通常计算资源有限,难以在实际运行过程中同时执行教师模型推理与深度估计任务。因此采用离线生成策略预先构建伪标签库,可在不增加推理阶段计算负担的前提下充分利用伪标签的几何监督信息,更适合资源受限的嵌入式部署场景。

通过在训练阶段预先生成伪标签,可以在推理阶段不增加计算负担的前提下,加入额外的几何监督信息,提升模型在水下场景中的深度估计能力。

3.2 伪深度标签的尺度归一化

Depth Anything V2 作为在大规模数据集上预训练的单目深度估计模型,其输出为相对深度图,反映的是场景内各像素之间的深度大小关系,不包含绝对物理尺度信息。不同场景、不同帧之间的深度输出值域差异较大。例如,开阔水域场景的输出值可能集中在较大数值区间,而近距离拍摄的珊瑚礁场景则集中在较小区间。若直接以原始输出值作为监督信号,批次内不同样本之间的深度量级不一致,会导致损失函数在不同样本上的梯度量级失衡,进而影响训练稳定性。

为解决上述问题,本文在数据加载阶段,对每帧的伪深度图进行尺度归一化处理。归一化流程分为3步:1)确定当前帧的有效像素集合,排除深度值缺失的无效区域;2)基于有效像素的第95百分位数估计当前帧的深度尺度;3)以该尺度对全图进行归一化,将有效深度值域约束至 $(0, 1]$ 附近的统一范围。

假设 Depth Anything V2 对第 i 帧输出的原始伪深度图为 D_{raw} ,定义有效像素集合为 Ω_{valid} ,则其关系可定义为:

$$\Omega_{\text{valid}} = \{p \mid D_{\text{raw}}(p) > 0\} \quad (4)$$

其中,深度值为零的像素对应传感器的无效测量区域,以及纯水体远端等深度缺失区域,不参与后续的尺度估计与监督计算。在有效像素集合上计算第95百分位数作为当前帧的深度尺度估计量:

$$p_{95} = P_{95}(\{D_{\text{raw}}(p) \mid p \in \Omega_{\text{valid}}\}) \quad (5)$$

为了抑制离群极值对尺度估计的干扰,此处选用第95百分位数而不是最大值。水下图像中常出现因反射、散射或遮挡产生的局部极大深度值,若以最大值归一化,少数离群点会将绝大多数有效像素的归一化值压缩至接近零的区间,严重削弱监督信号的有效性。以第95百分位数为基准,可保证95%的有效像素归一化后落在 $(0, 1]$ 范围内,同时对剩余5%的极值进行隐式截断,兼顾了信号有效性与对离群值的鲁棒性。在此基础上对全图进行归一化,得到尺度统一的伪深度标签 $\tilde{D}$:

$$\tilde{D}(p) = \frac{D_{\text{raw}}(p)}{p_{95} + \varepsilon} \quad (6)$$

式中: ε 表示防止除零问题的小常数。归一化后,有效像素的深度值被约束在 $(0, 1]$ 范围内,无效像素的值保持为0。为使深度监督损失仅作用于有效像素,定义逐像素有效性掩码 M :

$$M(p) = 1[\tilde{D}(p) > 0] \quad (7)$$

掩码 $M(p) = 1$ 表示该像素参与监督, $M(p) = 0$ 表示该像素被排除在损失计算之外。经归一化处理后,不同帧、不同场景的伪深度标签具有一致的值域分布,去除了帧与帧之间深度量级不一致对训练稳定性的影响,为之后损失函数的设计提供了可靠的监督信号基础。

本文采用协同训练策略,在确保原有的自监督训练框架基础大体不变的情况下,将伪标签深度图作为监督信号辅助约束,两类监督信号在训练过程中相互补充。其中,自监督信号用来维持相邻帧视角的几何一致性,深度伪标签将环境的深度结构信息提供给网络训练,使得在光度一致性假设因水下环境而破坏的情况下,网络训练稳定性得到改善和提高。

图8为本文所提协同训练框架的完整流程,展示了自监督深度估计、伪标签处理模块和两类监督信号的融合方式与分工。

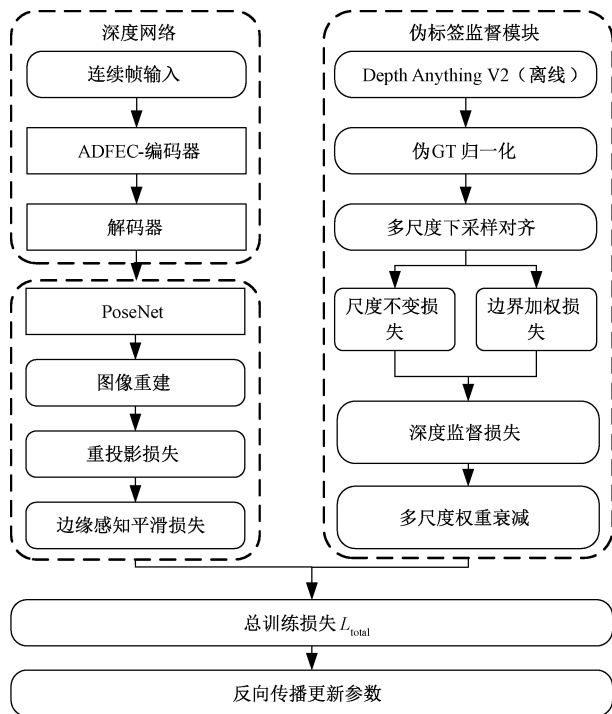


图8 协同训练机制

Fig. 8 Co-training mechanism

通过该协同训练策略,网络能够在保持自监督学习优势的同时,通过引入额外的深度结构信息,提升模型在

复杂水下环境中的深度估计准确性与稳定性。

4 面向水下场景的损失函数优化

自监督单目深度估计的训练目标由多个损失项共同构成。本章首先介绍基线框架沿用的基础损失函数,随后重点阐述本文针对水下场景所设计的伪标签深度监督损失,最后给出完整的综合训练损失函数形式。

4.1 基线损失函数

自监督单目深度估计框架的基线损失函数,由重投影损失与平滑损失两部分组成,总损失在多尺度操作步骤中进行聚合:

$$\mathcal{L}_{\text{base}} = \frac{1}{S} \sum_{s=0}^{S-1} \left(\mathcal{L}_{\text{reproj}}^{(s)} + \frac{\lambda_{\text{disp}}}{2^s} \mathcal{L}_{\text{smooth}}^{(s)} \right) \quad (8)$$

式中: S 表示多尺度层数; s 表示当前尺度索引; λ_{disp} 表示平滑损失的权重系数。

重投影损失函数是依据预测深度和相机位姿变换,将相邻帧图像重投影至目标帧后,重投影图像与目标帧之间的光度误差。图像序列的光度误差采用结构相似性 (structural similarity index measure, SSIM) 与 L_1 范数的加权组合:

$$\mathcal{L}_{\text{pe}}(\mathbf{I}_a, \mathbf{I}_b) = \frac{a}{2} \cdot (1 - \text{SSIM}(\mathbf{I}_a, \mathbf{I}_b)) + b \cdot \|\mathbf{I}_a - \mathbf{I}_b\|_1 \quad (9)$$

式中: $\mathbf{I}_a, \mathbf{I}_b$ 表示相邻帧的图像; SSIM 项能够感知图像的结构信息; L_1 项则提供逐像素的绝对误差约束,二者互补; a 设置为 0.85; b 设置为 0.15。

为了滤除因相机静止或静止目标区域的无效监督信号,基线框架使用自动掩码机制,将重投影误差与原始帧进行比较,逐像素判断光度误差差异,只保留重投影后误差更小的像素参与损失计算:

$$\mathcal{L}_{\text{reproj}}^{(s)} = \frac{1}{|P|} \sum_{p \in P} \min(M_1, M_2) \quad (10)$$

$$M_1 = \min_{t \in \mathcal{T}} \mathcal{L}_{\text{pe}}(\mathbf{I}_0^{(s,p)}, \hat{\mathbf{I}}_{t \rightarrow 0}^{(s,p)}) \quad (11)$$

$$M_2 = \min_{t \in \mathcal{T}} \mathcal{L}_{\text{pe}}(\mathbf{I}_0^{(s,p)}, \mathbf{I}_t^{(s,p)}) \quad (12)$$

式中: P 为有效像素集合; $\mathcal{T}$ 为参考帧的集合; $\hat{\mathbf{I}}_{t \rightarrow 0}^{(s,p)}$ 表示将第 t 帧图像重投影至目标帧后在像素 p 处的重采样值; $\mathbf{I}_t^{(s,p)}$ 为未经变换的源参考帧的像素值。取两者之间的最小值,这样有效去除因场景静止而产生的伪监督。

边缘感知平滑损失 $\mathcal{L}_{\text{smooth}}^{(s)}$ 用于约束视差图在梯度上的局部平滑性,利用图像梯度进行边缘感知加权,避免对深度不连续边界施加过强的平滑约束:

$$\mathcal{L}_{\text{smooth}}^{(s)} = |\partial_x \bar{\mathbf{d}}^{(s)}| \cdot e^{-|\partial_x \mathbf{I}^{(s)}|} + |\partial_y \bar{\mathbf{d}}^{(s)}| \cdot e^{-|\partial_y \mathbf{I}^{(s)}|} \quad (13)$$

式中: $\bar{\mathbf{d}}^{(s)}$ 为均值归一化后的视差图; ∂_x, ∂_y 分别表示水平和垂直方向的梯度算子; $\mathbf{I}^{(s)}$ 为对应尺度的输入图像。

图像梯度较大处 (即边缘位置) 的指数项趋近于 0,从而自动降低该区域的平滑约束强度。

4.2 深度监督损失函数

在有效的掩码约束下,深度监督损失通过计算预测深度与伪标签之间的误差来引导网络学习相对深度结构。水下场景存在颜色衰减、散射以及低对比度等特性,使得基于光度一致性的自监督信号在远距离区域容易失效,仅依赖重投影损失难以建立准确的全局深度结构。

自监督学习框架本身没有尺度校准,网络预测的深度与伪标签之间存在整体的尺度偏差。若直接采用绝对误差度量,训练损失会对尺度偏移非常敏感,不利于学习深度的相对结构关系。为解决上述问题,本文采用尺度不变损失 (scale-invariant loss, SI Loss) 在有效像素上度量预测深度与伪标签的误差。在对数域中计算深度误差,使损失函数更加关注深度之间的相对关系,使得误差约束在不同深度尺度上更加均衡。其表达式如式 (14)、(15) 所示。

$$e_p = \log d_{\text{pred}}^{(p)} - \log d_{\text{gt}}^{(p)} \quad (14)$$

$$\mathcal{L}_{\text{SI}} = \frac{1}{N} \sum_p (M_p \cdot e_p)^2 - \frac{1}{2} \left(\frac{1}{N} \sum_p M_p \cdot e_p \right)^2 \quad (15)$$

式中: N 表示有效像素的总数; $\mathcal{L}_{\text{SI}}$ 的第 1 项用于约束预测深度与真实深度在对数域中的逐像素误差,第 2 项通过减去误差均值的平方,削弱整体尺度偏移对损失的影响,即当网络预测的深度图整体乘以某一常数时,该损失不会发生变化。由于自监督学习框架本身没有尺度校准,而尺度不变项能够让模型专注于学习相对深度结构。由此,SI Loss 将网络的优化目标引导至学习深度的相对结构,而非拟合绝对数值。

4.3 基于伪标签梯度的边界加权损失

传统的平滑损失依赖图像梯度来保护深度边界,其隐含假设是深度不连续处通常伴随图像颜色或亮度的突变。然而,在水下环境中,由于光线的吸收与散射作用,颜色相近的不同物体在图像上几乎没有明显的边缘信号,导致基于图像梯度的加权机制失效;平滑损失会错误地跨越深度不连续边界强行平滑,从而模糊物体轮廓的深度估计。

针对上述问题,本文提出利用伪深度标签自身的梯度来定位深度边界,在边界区域加重预测误差的惩罚。Depth Anything V2 生成的伪深度图在物体边界处具有清晰的深度跳变,即使对应的图像颜色差异很小,这一深度跳变信号仍然显著。因此,直接利用伪 GT (ground truth) 的梯度作为边界检测依据,可以绕开图像梯度假设的局限。

计算伪深度标签的水平与垂直方向一阶差分,得到逐像素的边界强度图:

$$G_x(p) = |\tilde{D}(p_{x+1,y}) - \tilde{D}(p_{x,y})| \quad (16)$$

$$G_y(p) = |\tilde{D}(p_{x,y+1}) - \tilde{D}(p_{x,y})| \quad (17)$$

$$G(p) = G_x(p) + G_y(p) \quad (18)$$

式中: $G_x(p)$ 表示像素 p 处水平方向的一阶差分; $G_y(p)$ 表示垂直方向的一阶差分; $G(p)$ 为综合水平与垂直方向的逐像素边界强度。

在此基础上,构建逐像素边界加权系数:

$$w_{\text{edge}}(p) = 1.0 + 5.0 \cdot G(p) \quad (19)$$

非边界区域的权重恒为 1.0,边界强度越大,加权系数越高,最大可将边界处的惩罚放大至原来的 6 倍以上。边界加权损失定义为:

$$\mathcal{L}_{\text{boundary}} = \frac{\sum_p w_{\text{edge}}(p) \cdot M_p \cdot (\log d_{\text{pred}}^{(p)} - \log \tilde{D}^{(p)})^2}{\sum_p M_p + \varepsilon} \quad (20)$$

式中:分母的有效像素数 M_p 用于归一化; ε 表示防止除零问题的小常数。与 SI Loss 中仅在有效像素上计算不同, $\mathcal{L}_{\text{boundary}}$ 在全图上计算对数残差后再乘有效掩码,使边界加权系数能够准确获取空间位置信息,并充分利用伪标签梯度的空间分布信息。

综合上述两项,深度监督损失的完整形式为:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{SI}} + 0.5 \cdot \mathcal{L}_{\text{boundary}} \quad (21)$$

将系数设置为 0.5,用于调节两项损失之间的权重比例,使尺度不变约束作为主导项,边界加权损失作为辅助项,在整体深度结构准确的前提下,强化对水下物体轮廓的深度边界保持能力。

深度监督损失在多个分辨率尺度上均施加监督,但为了避免低分辨率伪标签经多次下采样后引入过多噪声,各尺度的监督权重随分辨率降低而指数衰减。设衰减权重为 w_s ,其定义如式(22)所示。

$$w_s = \frac{1}{2^s}, \quad s \in \{0, 1, 2, 3\} \quad (22)$$

对应数值为:尺度 $s=0$ 表示权重 1.0, $s=1$ 表示权重 0.5, $s=2$ 表示权重 0.25, $s=3$ 表示权重 0.125。尺度 $s=0$ 为原始分辨率,具有最丰富的空间细节,经双线性插值上采样后与稀疏伪标签的对齐精度最高,因此赋予最高权重。低分辨率特征图经多次下采样后高频信息已被抑制,对应的监督仅起到方向性约束作用,所以权重逐级减半。最终,综合训练损失为:

$$\mathcal{L}_{\text{total}} = \frac{1}{S} \sum_{s=0}^{S-1} (\mathcal{L}_{\text{reproj}}^{(s)} + \frac{\lambda_{\text{disp}}}{2^s} \cdot \mathcal{L}_{\text{smooth}}^{(s)} + \lambda_{\text{sup}} \cdot w_s \cdot \mathcal{L}_{\text{sup}}^{(s)}) \quad (23)$$

式中: λ_{sup} 表示深度监督损失的全局权重超参数,用于平衡自监督信号与伪标签监督信号之间的相对强度; S 表示多尺度层数。当 $\lambda_{\text{sup}} = 0$ 时,综合损失退化为纯自监督

的基线形式,与原始基线框架完全等价,便于消融实验中逐项验证各损失项的贡献。

本章提出的深度监督损失通过以下 3 个方面对水下场景的自监督深度估计进行针对性优化:1) 利用 Depth Anything V2 生成的伪标签引入绝对深度的结构性先验,弥补自监督信号在远距离区域的失效。2) SI Loss 消除了网络预测与伪标签之间的全局尺度偏差,使监督信号聚焦于相对深度关系。3) 基于伪标签梯度的边界加权损失,有效绕开了图像梯度假设,在水下颜色差异微弱的物体边界处,主动加强深度不连续性的有效值,改善水下场景中深度估计的边缘清晰度。

5 实验结果及分析

为验证所提方法在水下视觉场景中的有效性,本章在 FLSea 数据集上对改进后的 Lite-Mono 模型进行实验评估,并与原始 Lite-Mono 模型进行对比分析,定量对比实验随机选取了 1 715 帧图像作为测试数据集。

网络训练在配备 NVIDIA RTX 4060 GPU 的工作站上进行,训练框架为 PyTorch,训练轮次为 30 epochs。为了验证模型的泛化能力,在定性分析中还加入 USOD10K 数据集^[31]和本研究采集的水下灯阵列数据集进行可视化对比。

5.1 实验系统与数据采集

为验证所提方法在真实水下复杂环境中的深度估计性能,本研究搭建了水下实验系统,如图 9 所示,并采集了面向深度估计任务的图像序列。实验系统由可控光源阵列、水下图像采集工业相机等构成,整体布置于室内水池环境中。水池尺寸约为 5 m×5 m (长×宽),最深处约为 10 m,水体中存在一定浓度的悬浮颗粒物,水面反射与局部高光区域共同构成了具有代表性的复杂光照条件,具有实际水下场景中的典型光学退化特性。

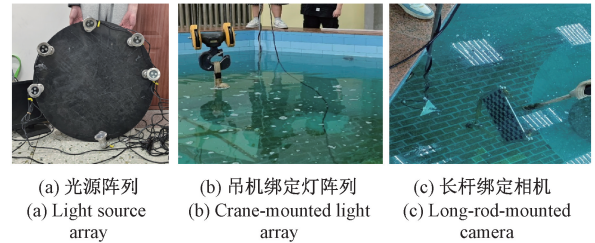


图 9 实验系统

Fig. 9 Experimental system

如图 9 所示,光源阵列由多个点光源按固定空间几何结构排列而成,固定于水池一侧。光源阵列作为具有明确空间结构的光学参照物,用于在不同距离与光照条

件下验证深度估计模型对结构边缘、局部高光区域等情况的深度估计能力。水下相机通过刚性连接固定于可沿水池上方轨道移动的平台, 相机光轴指向光源阵列方向, 确保目标位于视场中。

如图 10 所示, 实验采集的图像中出现明显的局部过曝与反射光斑。在远距离或大角度条件下, 目标区域照度下降、对比度减弱。上述变化过程覆盖了水下深度估计任务中常见的光照不均、局部高亮、纹理退化等典型场景。



图 10 复杂水下光照环境下的采集图像
Fig. 10 Acquired images in complex underwater illumination environments

每帧图像均保留原始分辨率与色彩信息。该数据集与 FLSea、USOD10K 公开数据集形成互补: 公开数据集主要提供静态水下场景的深度真值, 而本实验数据强调动态光照变化与连续距离变化下的深度估计的鲁棒性, 为模型在真实物理环境中的泛化能力提供验证依据。

5.2 定量分析

实验使用单目深度估计常用的评价指标进行定量分析, 其中包括绝对相对误差 (absolute relative error, Abs Rel)、平方相对误差 (square relative error, Sq Rel)、均方根误差 (root mean square error, RMSE, 单位为 m)、对数均方根误差 RMSE log 和精度指标 $\delta < 1.25$ 、 $\delta < 1.25^2$ 和 $\delta < 1.25^3$ 。其中, 误差类指标数越小表示深度预测越准确, 而精度类指标数越大表示预测深度值和真实深度具有越高的一致性。精度指标 $\delta < 1.25^k (k = 1, 2, 3)$ 定义为: 在所有有效像素中, 满足像素占比:

$$\max\left(\frac{d_{\text{pred}}}{d_{\text{gt}}}, \frac{d_{\text{gt}}}{d_{\text{pred}}}\right) < 1.25^k \tag{24}$$

其中, d_{pred} 为预测深度, d_{gt} 为真实深度。 $k = 1, 2, 3$ 分别对应精度指标 a1、a2、a3, 阈值依次放宽, 该指标值越大表示深度预测精度越高。

表 1 给出了改进后的 Lite-Mono 模型和基础 Lite-Mono 模型在水下图像数据集 FLSea 上的定量评估结果。

表 1 消融实验定量分析

Table 1 Quantitative analysis of ablation experiments

模型	Abs Rel	Sq Rel	RMSE	RMSE log	a1	a2	a3
Lite-Mono	0.259	0.366	1.151	0.354	0.572	0.816	0.918
仅 ADFEC	0.194	0.232	0.938	0.263	0.684	0.889	0.959
ADFECD+DA2	0.183	0.237	0.872	0.263	0.732	0.906	0.962

由表 1 中数据可以看出, 引入 ADFEC 模块后 (仅 ADFEC), 模型各项误差指标均有显著下降: Abs Rel 由 0.259 降至 0.194, 降幅约 25.1%; RMSE 由 1.151 降至 0.938, 降幅约 18.5%; RMSE log 由 0.354 降至 0.263, 降幅约 25.7%。精度指标 a1、a2、a3 分别从 0.572、0.816、0.918 提升至 0.684、0.889、0.959, 说明 ADFEC 模块对提升水下场景深度估计精度具有明显效果。

在此基础上进一步引入伪标签辅助训练策略 (ADFECD+DA2), Abs Rel 继续降低至 0.183, RMSE 降至 0.872, 精度指标 a1、a2、a3 分别提升至 0.732、0.906、0.962, 各项主要指标均优于仅使用 ADFEC 的结果。Sq Rel 指标在完整改进后略高于仅嵌入 ADFEC 模块的模型, 这是由于 Sq Rel 对大误差敏感, 伪标签的几何约束在部分样本上的预测偏差经过了平方放大后显现, 但其余六项指标的一致改善表明整体性能提升是稳定有效的。总体指标表明, 改进后的 Lite-Mono 模型在 FLSea 数据集上的各项关键指标均显著优于基线 Lite-Mono, 验证了各改进模块的有效性。

5.3 定性分析

图 11 展示了 3 组消融配置的深度图可视化对比结果, 在 FLSea 数据集同样训练 30 个 epoch 后, 分别在 FLSea (图 11 第 1~3 列)、USOD10K (图 11 第 4~6 列) 和本文采集的光源阵列数据集 (图 11 第 7~9 列) 上进行测试。其中第 1 行为原始水下图像, 第 2~4 行依次为基线 Lite-Mono、仅 ADFEC 模块和 ADFEC + Depth Anything V2 的深度图预测结果, 热力图中颜色越亮, 表示预测距离越近, 颜色越暗, 则表示预测距离越远。

由图 11 中所列的图像可以看出, 基线 Lite-Mono 模型在各场景下输出的深度图整体结构模糊, 近景与远景的深度层次区分能力不足, 目标边界存在模糊, 难以有效反映真实场景的空间结构。网络加入 ADFEC 模块和伪标签辅助训练策略后, 深度图在局部细节上有所改善, 深度图质量得到显著提升。

本文所提模型在 FLSea 数据集中, 前景目标的轮廓得到清晰刻画, 近景的礁石和背景水体的深度层次明显分离。在 USOD10K 数据集中, 沉船等复杂结构的边界响应更加准确。在本实验系统采集的灯阵列数据集中,

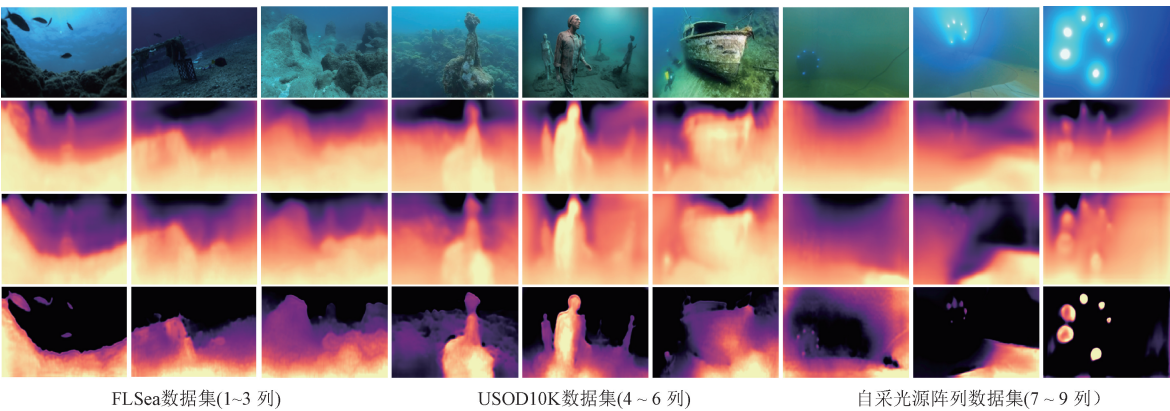


图 11 消融实验定性分析
Fig. 11 Qualitative analysis of ablation experiments

面对不同角度、不同距离,光照复杂、纹理稀疏的场景,模型仍能输出具有合理空间分布的深度估计结果,体现了方法在不同水下场景的泛化能力。综合 3 个数据集的定性结果与 5.2 节的定量分析,所提的改进方法有效性得到充分验证。

6 结 论

针对水下无人航行器的自主对接回收作业中因图像光照不均、纹理退化导致轻量化深度估计网络特征表达能力不足和边缘细节模糊的问题,构建了自适应解耦特征增强模块 ADFEC,将其嵌入到 Lite-Mono 编码器的多尺度特征提取阶段,通过对特征图中空间与语义信息的动态调整,强化网络对关键特征区域的表达能力。同时,设计了基于 Depth Anything V2 伪标签的协同训练策略,为网络提供额外的几何结构约束,并结合面向水下场景的损失函数优化,提升了模型在复杂水下环境中的深度估计精度与稳定性。在 FLSea 数据集上的定量实验结果表明,改进后的模型相较于基线 Lite-Mono 在 Abs Rel、RMSE 等主要指标上均取得了显著提升,定性实验验证了所提方法在 FLSea、USOD10K 和本研究自采灯阵列数据集上具有良好的泛化能力。所提方法的轻量化特性与精度提升,为 UUV 对接回收任务中的实时水下环境感知提供了可行的技术支撑。

参考文献

[1] 宋巍,朱孟飞,张明华,等. 基于深度学习的单目深度估计技术综述[J]. 中国图象图形学报, 2022, 27(2): 292-328.
Song Wei, Zhu Mengfei, Zhang Minghua, et al. A review of monocular depth estimation techniques based on deep learning[J]. Journal of Image and Graphics, 2022, 27(2): 292-328.

[2] 赵世泉,刘佳兴,王宇超. 基于改进 RTMPose 的水下光阵关键点检测[J]. 仪器仪表学报, 2026, 47(3): 158-169.
Zhao Shiquan, Liu Jiaxing, Wang Yuchao. The keypoint detection of underwater optical guidance arrays based on the enhanced RTMPose method[J]. Chinese Journal of Scientific Instrument, 2026, 47(3): 158-169.
[3] 杜道平,李响,梁野,等. 改进暗通道的水下图像复原算法[J]. 电子测量技术, 2024, 47(5): 142-149.
Du Daoping, Li Xiang, Liang Ye, et al. Improved dark channel algorithm for underwater image restoration[J]. Electronic Measurement Technology, 2024, 47(5): 142-149.
[4] 裴晓芳,杨继海,周进,等. 混合注意力与动态特征引导网络的水下检测[J/OL]. 电子测量技术, 2026-03-05. <https://link.cnki.net/urlid/11.2175.TN.20260304.1806.016>.
Pei Xiaofang, Yang Jihai, Zhou Jin, et al. Underwater detection with hybrid attention mechanism and dynamic feature guidance network[J/OL]. Electronic Measurement Technology, 2026-03-05. <https://link.cnki.net/urlid/11.2175.TN.20260304.1806.016>.
[5] 陈哲,周旭,沈洁,等. 基于梯度先验的水下图像恢复[J]. 仪器仪表学报, 2022, 43(8): 39-46.
Chen Zhe, Zhou Xun, Shen Jie, et al. Underwater image restoration based on gradient prior[J]. Chinese Journal of Scientific Instrument, 2022, 43(8): 39-46.
[6] 麦仁贵,王骥. 基于关键特征传递的 KFT-GAN 水下图像增强模型[J]. 电子测量与仪器学报, 2026, 40(2): 76-85.
Mai Rengui, Wang Ji. Underwater image enhancement model based on generative adversarial network with key

- feature transfer[J]. *Journal of Electronic Measurement and Instrumentation*, 2026, 40(2): 76-85.
- [7] 周露珊, 赵磊, 李恒, 等. 基于轻量级密集残差网络的水下图像增强[J]. *电子测量与仪器学报*, 2023, 37(1): 70-77.
- Zhou Lushan, Zhao Lei, Li Heng, et al. Underwater image enhancement based on lightweight dense residual net-work[J]. *Journal of Electronic Measurement and Instru-mentation*, 2023, 37(1): 70-77.
- [8] Zhou T H, Brown M, Snavely N, et al. Unsu-pervised learning of depth and ego-motion from video[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1851-1858.
- [9] 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, 2018: 1983-1992.
- [10] Mahjourian R, Wicke M, Angelova A. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 5667-5675.
- [11] Johnston A, Carneiro G. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 4756-4765.
- [12] Yan J X, Zhao H, Bu P, et al. Channel-wise attention-based network for self-supervised monocular depth estimation[C]//2021 International Conference on 3D Vision, 2021: 464-473.
- [13] Zhou Zh M, Dong Q L. Self-distilled feature aggre-gation for self-supervised monocular depth estimation[C]//European Conference on Computer Vision, 2022: 709-726.
- [14] Lyu X Y, Liu L, Wang M M, et al. HR-depth: High resolution self-supervised monocular depth estima-tion[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2294-2301.
- [15] Zhao C Q, Zhang Y M, Poggi M, et al. Monovit: Self-supervised monocular depth estimation with a vision transformer[C]//2022 International Conference on 3D Vision, 2022: 668-678.
- [16] Godard C, Aodha O M, Brostow G J. Unsupervised monocular depth estimation with left-right consistency [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6602-6611.
- [17] 谢昭, 马海龙, 吴克伟, 等. 基于采样汇集网络的场景深度估计[J]. *自动化学报*, 2020, 46(3): 600-612.
- Xie Zhao, Ma Hailong, Wu Kewei, et al. Sampling aggregate network for scene depth estimation[J]. *Acta Automatica Sinica*, 2020, 46(3): 600-612.
- [18] Amitai S, Klein I, Treibitz T. Self-supervised monocular depth underwater [C]//2023 IEEE Inter-national Conference on Robotics and Automation, 2023: 1098-1104.
- [19] Yu B X, Wu J Y, Islam M J. Udepth: Fast monocular depth estimation for visually-guided underwater robots[J]. *ArXiv preprint arXiv:2209.12358*, 2022.
- [20] Zhang F, You S D, Li Y, et al. Atlantis: Enabling underwater depth estimation with stable diffusion[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 11852-11861.
- [21] Gu W X, Qi L. Dense geometry supervision for under-water depth estimation [C]//Advanced Fiber Laser Conference, 2025: 1354405.
- [22] Petrovai A, Nedevschi S. Exploiting pseudo labels in a self-supervised learning framework for improved monocular depth estimation [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022: 1578-1588.
- [23] Watson J, Firman M, Brostow G, et al. Self-supervised monocular depth hints [C]//2019 IEEE/CVF International Conference on Computer Vision, 2019: 2162-2171.
- [24] Saunders K, Vogiatzis G, Manso L J. Self-supervised monocular depth estimation: Let's talk about the weather [C]//2023 IEEE/CVF International Con-ference on Computer Vision, 2023: 8907-8917.
- [25] Zhang N, Nex F, Vosselman G, et al. Lite-Mono: A lightweight CNN and transformer architecture for self-supervised monocular depth estimation[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 18537-18546.
- [26] Yang L H, Kang B Y, Huang Z L, et al. Depth anything V2[J]. *Advances in Neural Information Processing Systems*, 2024, 37: 21875-21911.
- [27] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite

[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012: 3354-3361.

[28] Wang Y, Peng L, Li L, et al. Decoupling-and-aggregating for image exposure correction [C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 18115-18124.

[29] Basri R, Jacobs D W. Lambertian reflectance and linear subspaces [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(2) : 218-233.

[30] Randall Y. FLSea: Underwater visual-inertial and stereo-vision forward-looking datasets [D]. Haifa: University of Haifa, 2023.

[31] Hong L, Wang X, Zhang G, et al. USOD10K: A new benchmark dataset for underwater salient object detection [J]. IEEE Transactions on Image Processing, 2025, 34: 1602-1615.

作者简介



E-mail:fuhuixuan@ hrbeu. edu. cn

傅荟璇, 2007 年于哈尔滨工程大学获得硕士学位, 2009 年于哈尔滨工程大学获得博士学位, 现为哈尔滨工程大学副教授, 主要研究方向为模式识别与人工智能、深度学习及机器视觉检测、无人系统视觉伺服控制。

Fu Huixuan received her M. Sc. degree from Harbin Engineering University in 2007, received her Ph. D. degree from Harbin Engineering University in 2009. She is currently an associate professor at Harbin Engineering University. Her main research interests include pattern recognition and artificial intelligence, deep learning and machine vision detection, unmanned system visual servo control.



王宇超 (通信作者), 2009 年于哈尔滨工程大学获得硕士学位, 2011 年于哈尔滨工程大学获得博士学位, 现为哈尔滨工程大学副教授, 博士生导师, 主要研究方向为水下无人系统智能感知、海洋运载器自主决策与智能控制、嵌入式系统设计与集成。

E-mail:wangyuchao@ hrbeu. edu. cn

Wang Yuchao (Corresponding author) received his M. Sc. degree from Harbin Engineering University in 2009, received his Ph. D. degree from Harbin Engineering University in 2011. He is currently an associate professor and doctoral supervisor at Harbin Engineering University. His main research interests include intelligent perception for underwater unmanned systems, autonomous decision-making and intelligent control for marine vehicles, embedded system design and integration.