https://crad.ict.ac.cn
63(2): 338−351, 2026

局部特征引导的室内自监督单目深度估计方法的改进

艾浩军[1,2] 张 锋[1,2] 吕鹏飞[1,2] 唐雪华[3] 王中元[4]

[1](武汉大学国家网络安全学院 武汉 430072)

[2](空天信息安全与可信计算教育部重点实验室(武汉大学) 武汉 430072)

[3](武汉大学遥感信息工程学院 武汉 430079)

[4](武汉大学计算机学院 武汉 430079)

(aihj@whu.edu.cn)

Improving Self-Supervised Monocular Indoor Depth Estimation Method with Local Feature Guidance

Ai Haojun[1,2], Zhang Feng[1,2], Lü Pengfei[1,2], Tang Xuehua[3], and Wang Zhongyuan[4]

[1] (*School of Cyber Science and Engineering*, *Wuhan University*, *Wuhan* 430072)

[2] (*Key Laboratory of Aerospace Information Security and Trusted Computing* (*Wuhan University*), *Ministry of Education*, *Wuhan* 430072)

[3] (*School of Remote Sensing and Information Engineering*, *Wuhan University*, *Wuhan* 430079)

[4] (*School of Computer Science*, *Wuhan University*, *Wuhan* 430079)

Abstract In recent years, self-supervised monocular depth estimation methods have achieved impressive improvements. However, their performance degrades significantly when generating structured depth maps in complex indoor scenarios. To bridge this gap, focusing on the training process, we propose LoFtDepth, a novel method that combines self-supervised monocular depth estimation with local feature guided knowledge distillation. Firstly, an off-the-shelf depth estimation network is used to generate structured relative depth maps as depth priors. Local features are then extracted from these priors as boundary points, guiding the local depth refinement. This reduces the interference of depth-irrelevant features and transfers the boundary knowledge of depth priors to the self-supervised depth estimation network. Additionally, we introduce an inverse auto-mask weighted surface normal loss. This encourages normal directions of depth maps predicted by self-supervised network to align with those of depth priors in untextured regions. As a result, the depth estimation accuracy is enhanced. Finally, according to the coherence of camera motion, we impose a pose consistency constraint on residual pose estimation. This constraint enables effective adaptation to indoor scenes where camera poses change frequently, thereby mitigating training errors and boosting model performance. Extensive experiments on major indoor datasets demonstrate that LoFtDepth outperforms previous methods. It reduces the absolute relative error to 0.121, and successfully generates accurate and well-structured depth maps.

Key words monocular depth estimation; self-supervised learning; local feature; knowledge distillation; surface normal constraint

摘 要 近年来,自监督单目深度估计方法取得了显著的性能提升,但在复杂的室内场景生成结构化深度图时性能明显下降,为此,提出局部特征引导知识蒸馏的自监督单目深度估计方法 LoFtDepth 改进训练过

收稿日期: 2024-12-19; **修回日期**: 2025-06-30

基金项目: 国家自然科学基金项目(72374161); 湖北省国际科技合作计划项目(2025EHA043); 国家重点研发计划项目(2016YFB0502204)

This work was supported by the National Natural Science Foundation of China (72374161), the Hubei Provincial International Science and Technology Collaboration Program (2025EHA043), and the National Key Research and Development Program of China (2016YFB0502204).

通信作者: 唐雪华(tangxuehua@whu.edu.cn)

程。首先，使用预训练的深度估计网络预测结构化的相对深度图作为深度先验，从中提取局部特征作为边界点引导局部深度估计细化，减少深度无关特征的干扰，将深度先验中的边界知识传递到自监督深度估计网络中。同时，引入逆自动掩模加权的表面法线损失，通过对齐自监督网络预测的深度图和深度先验在无纹理区域的法线方向来提升深度估计精度。最后，根据相机运动的连续性，对相机位姿残差估计施加位姿一致性约束以适应室内场景相机位姿的频繁变化来减小训练误差和提升模型性能。主要的室内公开数据集上的实验结果表明，LoFtDepth 性能提升显著，将相对误差降至 0.121，且生成的深度图具有更高的全局准确度和良好的结构特征。

关键词 单目深度估计；自监督学习；局部特征；知识蒸馏；表面法线约束

中图法分类号 TP391

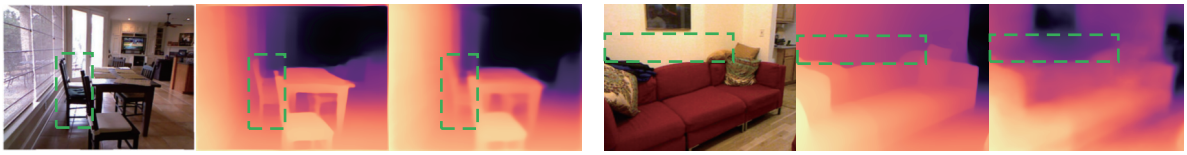
DOI: 10.7544/issn1000-1239.202440951 **CSTR:** 32373.14.issn1000-1239.202440951

深度信息是连接 2 维图像和 3 维世界的桥梁。准确且结构化的深度图对自动驾驶[1]、增强现实[2]和 3 维重建[3]等下游应用至关重要。随着硬件技术的进步，LiDAR 和 Time-of-Flight 等深度传感器能够采集深度信息，但其昂贵的造价及采集过程中巨大的时间、人力开销限制了应用范围。因此，研究者们致力于直接从图像估计深度信息。近年来，基于深度学习的单目深度估计方法受到越来越多的关注[4-10]。

基于深度学习的单目深度估计方法根据训练过程中是否使用真实深度标签，可以分为有监督和自监督(无监督)方法。依赖大规模带深度标签的图像数据，有监督方法在多数场景的误差低于 8%[6-8,11-13]，但数据的采集成本高昂。为此 Zhou 等人[9]提出了自监督方法，使用单目图像序列训练，推理时只需 1 幅图像即可进行深度估计。在训练过程中，选取单目图像序列中的一帧作为目标图像，并将其相邻帧作为参考图像，使用相机位姿估计网络预测 2 幅图像之间的相对相机位姿。再运用相机内参，将参考图像重投影到目标图像平面得到重投影图像。重投影图像和目标图像的外观一致性作为主要损失函数。基于该算法框架，自监督单目深度估计在驾驶场景上深度估计误差低至约 10%[5,14-17]。然而，当这些方法直接应用于室内场景时性能会显著下降，不能生成准确且结构化深度图[4,18]。具体地，这类以图像序列训练自监督深度估计模型的方法应用到室内场景时面临 3 个挑战：1)复杂的物体边界。室内场景结构繁杂，缺乏约束使模型无法准确预测边界深度。2)大面积的无纹理区域。区域内的像素外观一致，无论深度预测是否准确，损失函数值都较低，误导模型优化。3)多样的相机位姿变化。相机位姿估计网络无法正确预测多样的相机位姿，干扰重投影过程，进而影响深度精度。本文针对自监督单目深度估计模型训练过程中的这 3 个挑战，改进和优化从图像序列学习深度知识的过程，提升模型的室内场景泛化能力。

针对室内场景边界复杂的问题，现有工作[4,10,19]主要通过优化边界及其邻域的深度解决该问题。其中，Wu 等人[4]提出了一种基于知识蒸馏的方法，通过匹配教师网络和学生网络生成的深度图的物体边界进行约束。以深度梯度的全局分位数作为阈值，并直接将超过该阈值的位置认定为物体边界。该操作会将梯度较大但变化平滑的非边界区域误判为边界，导致模型在训练过程中学习错误的边界特征，以至于最终预测深度图的边界模糊，如图 1(a)最右侧图像所示。

为解决大面积无纹理区域的深度估计的问题，现有方法基于同一平面颜色分布的一致性[20]或曼哈顿法线的一致性[18]来识别共面的区域，利用共面约束提高无纹理区域的深度估计精度。这些假设在很多室内场景并不成立，需要一个新的无纹理区域深



(a) 更清晰的物体边界 (b) 更准确的无纹理区域预测结果

Fig. 1 The advantages of our proposed method

图 1 本文方法的优点

度估计方法。

在相机位姿预测方面，研究者们通过减弱[21]或预先估计[22]相机旋转运动来优化位姿估计；Ji 等人[23-24]提出位姿残差估计改善相机位姿估计。这些方法利用图像对约束改进位姿估计，显著提升性能。可以利用图像序列的连续性作为相机运动的一致性约束，更精确估计相机位姿。

综合上述事实，为提升单目自监督学习深度估计在室内场景的预测精度，提出了结合局部特征引导知识蒸馏的自监督单目深度估计新方法 LoFtDepth。因为深度图上的局部特征具有深度突变的特点，可以将其视为结构特征的边界点。所以，在知识蒸馏中，提取教师深度图的局部特征作为边界点进行局部深度估计细化，鼓励学生网络学习教师网络边界点及邻域的局部深度分布，使学生网络可以预测结构化的深度图；为了提升无纹理区域的深度精度，使用逆自动掩模加权表面法线损失，使表面法线损失针对无纹理区域进行约束。在自监督分支中，受 Li 等人[16]的启发，在位姿残差估计的基础上引入位姿一致损失，使连续多帧之间的相机位姿满足相机运动的连续性，改善相机位姿估计，提升深度估计精度。经过以上改进的自监督学习网络 LoFtDepth，能够生成具备清晰的物体边界且在无纹理区域的深度准确的高精度的深度图，如图 1 所示。本文的主要贡献有 3 点：

1)提出局部特征引导知识蒸馏的自监督单目深度估计方法 LoFtDepth。以深度图的局部特征作为边界点进行知识蒸馏，排除深度无关特征的影响，学生模型能够精确学习教师模型在边界点的局部深度分布，提升学生模型对于物体边界的感知，使其能够预测结构化的深度图。

2)针对无纹理区域深度估计不准确的问题，设计逆自动掩模加权的法线损失，保持学生模型和教师模型在无纹理区域的法线方向一致性，提升学生模型对无纹理区域的深度估计精度。

3)针对相机位姿预测错误的问题，提出位姿一致的相机位姿残差估计，以前后相邻 2 幅参考图像的位姿作为伪真值监督目标图像与参考图像的位姿，提高相机位姿估计精度，有效降低训练误差，提升模型预测深度的精度。

1 相关工作

基于深度学习的单目深度估计旨在学习从单幅 RGB 图像到深度图的映射。自 Eigen 等人[6]将深度学习技术应用于单目深度估计以来，研究者们已经提出诸多重要的有监督方法[6-8,11-12]以提升估计精度。其中，Ranftl 等人[12]提出了 DPT(dense prediction Transformer)，使用 vision Transformer 作为骨干网络，能够预测结构化的深度图。但因为在混合数据集上使用相对深度进行训练的缘故，DPT 只能够提供相对深度。随着视觉大模型 VFM(vision foundation model)的发展，为提升模型的鲁棒性，Yang 等人[25-26]和 Ke 等人[27]提出了 Depth Anything 系列[25-26]及 Marigold[27]，分别使用 DINOv2[28]和 Stable Diffusion v2[29]作为编码器。充分使用编码器预训练时的丰富先验知识，并进行一定的优化，取得了稳健的泛化性能。本文以 DPT 作为教师网络进行边界深度的知识蒸馏，提升学生网络的性能。

在自监督深度估计领域，Godard 等人[30]和 Zhou 等人[9]分别使用立体图像和单目连续帧学习单目深度估计，都采用外观一致性作为监督信号。特别地，后者引入额外的网络来估计相对相机位姿，用于重投影参考图像。在此基础之上，后续研究通过解决遮挡[5]、尺度不一致[15]等问题或联合语义分割[14]取得了显著进展[16,31-32]。然而，这些方法在训练时缺乏对室内场景中复杂边界、大面积无纹理区域以及复杂相机运动等挑战的优化，致使其在室内场景的性能大幅下降。

针对驾驶场景的自监督方法在应对室内场景内复杂的边界存在不足，导致生成的深度图物体边界模糊。为此，Wu 等人[4]融合知识蒸馏，通过边界匹配约束学生网络。但他们使用全局分位数定义边界，将部分深度梯度较大的区域误判为边界。为解决这一问题，我们在教师深度图上提取局部特征作为边界点，在局部特征的引导下增强自监督深度估计网络对于边界知识的认知。

针对室内场景中大面积无纹理区域的问题，Wang 等人[20]通过提取超像素并根据颜色来确定超像素是否共面，利用共面约束改善深度估计。基于曼哈顿世界假设，Li 等人[18]利用曼哈顿法线来对齐由预测深度生成的表面法线，并结合颜色、对齐后的法线和深度来确定共面像素，从而实施共面约束。Zhao 等人[33]则通过更换一个更好的骨干网络来缓解这个问题。

室内场景中复杂的相机运动影响自监督深度估计的训练过程，Ji 等人[23-24]提出了残差姿态估计，通过迭代预测相机位姿来适应复杂相机位姿变化。

Bian 等人[21]分析了相机位姿对深度估计的影响，表明旋转运动可视为噪声。由此提出自动纠正网络，减弱图像序列的旋转运动。Wu 等人[22]则预先估计相机旋转矩阵，然后将目标图像和参考图像进行预处理后再输入给自监督单目深度估计网络进行训练。Zhao 等人[33]使用 COLMAP 方法获得粗略的相机位姿，然后使用一个对齐网络来细化相机位姿。

2 自监督单目深度估计一般原理

本文以 MonoDepth2[5]作为基线模型，为便于叙述，简化其主要流程如图 2 所示。在推理阶段，仅需将单幅图像输入给深度估计网络，即可得到预测的深度图。

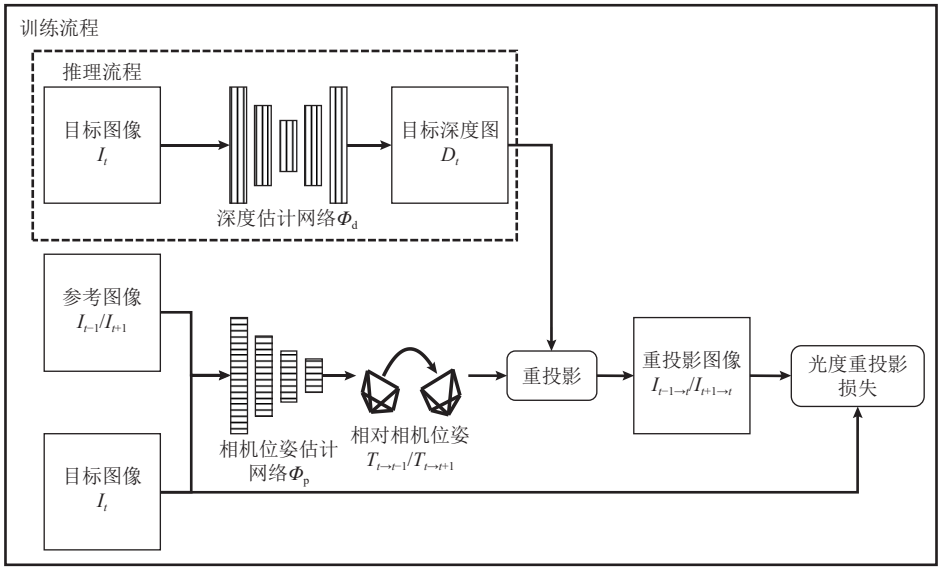


Fig. 2 The framework of self-supervised monocular depth estimation baseline model

图 2 自监督单目深度估计基线模型的框架

在训练时，将单目图像序列的一帧作为目标图像I_t并将其相邻帧作为参考图像$I_{t'}, t' \in \{t-1, t+1\}$。使用深度估计网络$\Phi_d$预测目标深度图$D_t$，相机位姿估计网络$\Phi_p$预测相对相机位姿$T_{t\to t'}$，再结合相机内参$K$将参考图像$I_{t'}$重投影到目标图像$I_t$的图像平面上得到重投影图像$I_{t'\to t}$用于计算损失函数。重投影具体为：先将$I_t$投影到$I_{t'}$，然后根据投影得到的坐标对应关系，在$I_{t'}$上进行双线性采样得到$I_{t'\to t}$，如式(1)所示。

$$I_{t'\to t} = I_{t'}\langle proj(D_t, T_{t\to t'}, K, I_t)\rangle , \tag{1}$$

其中$\langle\cdot\rangle$表示双线性采样算子，函数$proj(\cdot)$表示将I_t投影到$I_{t'}$，投影过程如式(2)所示。

$$I_{t\to t'} \sim KT_{t\to t'}D_tK^{-1}I_t 。 \tag{2}$$

接着，以I_t和$I_{t'\to t}$的外观一致性计算光度重投影损失(photometric reprojection loss)，其构造如式(3)所示。

$$\mathcal{L}_p = \mathcal{M}_a \min_{I_{t'}} pe(I_t, I_{t'\to t}) , \tag{3}$$

其中$\min_{I_{t'}}$表示在每个像素处，取多个参考图像的最小光度误差，而非采用所有参考图像的平均光度误差。$\mathcal{M}_a$为自动掩模，由光度误差pe生成，用于去除训练过程中与相机保持相对静止的像素，其定义如式(4)所示，通过比较重投影图像$I_{t'\to t}$与目标图像I_t之间的光度误差$pe(I_t, I_{t'\to t})$以及参考图像$I_{t'}$与目标图像I_t之间的光度误差$pe(I_t, I_{t'})$，光度误差$pe(I_t, I_{t'})$小于等于光度误差$pe(I_t, I_{t'\to t})$的像素被认为是与相机保持相对静止的像素，需要在自动掩模$\mathcal{M}_a$中被去除。pe为光度误差(photometric error)，由 L1 损失和结构相似性$SSIM$(structural similarity)[34]组合而成。其计算方式如式(5)所示。

$$\mathcal{M}_a = [\min_{I_{t'}} pe(I_t, I_{t'\to t}) < \min_{I_{t'}} pe(I_t, I_{t'})] , \tag{4}$$

其中[·]表示艾弗森括号(Iverson bracket)，括号内的条件成立时，$\mathcal{M}_a$=1，否则$\mathcal{M}_a$=0。

$$pe(I_t, I_{t'\to t}) = \frac{\alpha}{2}(1 - SSIM(I_t, I_{t'\to t})) + (1-\alpha)\|I_t - I_{t'\to t}\|_1 , \tag{5}$$

其中α为权重控制项，设置α=0.85。$SSIM$由 2 幅图像的均值、方差和协方差计算得到，如式(6)所示。

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} 。 \tag{6}$$

除了光度重投影损失外，自监督训练还使用了

边缘感知平滑损失$\mathcal{L}_s$, 该损失函数定义如式(7)所示。

$$\mathcal{L}_s = \left|\partial_x (D_t^{-1})^*\right| \mathrm{e}^{-\left|\partial_x I_t\right|} + \left|\partial_y (D_t^{-1})^*\right| \mathrm{e}^{-\left|\partial_y I_t\right|}, \tag{7}$$

其中$(D_t^{-1})^* = D_t^{-1} / \overline{D_t^{-1}}$为平均归一化的逆深度[35]。

3 LoFtDepth 方法

3.1 方法总览

本文提出了局部特征引导知识蒸馏的自监督单目深度估计方法 LoFtDepth, 整体框架如图 3 所示。

LoFtDepth 由蒸馏分支和自监督分支组成, 使用现有方法常用的统计损失和光度重投影损失[4-5,17]。为让自监督深度估计网络能够生成结构化且准确的深度图, 施加额外的约束。

蒸馏分支包含参数固定的教师网络和自监督学习的学生网络。首先在教师深度图上提取局部特征, 将其作为边界点。深度图上的局部特征具有明显深度变化, 将局部特征视为深度图的边界点, 可以排除深度变化平滑的深度无关特征对知识蒸馏的影响。为了提升学生网络对边界深度的认知, 以局部特征为中心, 分别在教师深度图和学生深度图上提取深度块。使用深度块计算局部细化损失, 指导学生网络学习教师网络在边界点及邻域的局部深度分布。根据教师(学生)深度图计算教师(学生)法线图, 使用教师法线图和学生法线图计算法线损失, 通过对齐法线方向进行共面约束。在室内场景中, 自动掩模去除的像素大多位于无纹理区域。根据这一特点, 为了让共面约束针对于无纹理区域, 使用逆自动掩模对法线损失进行加权。

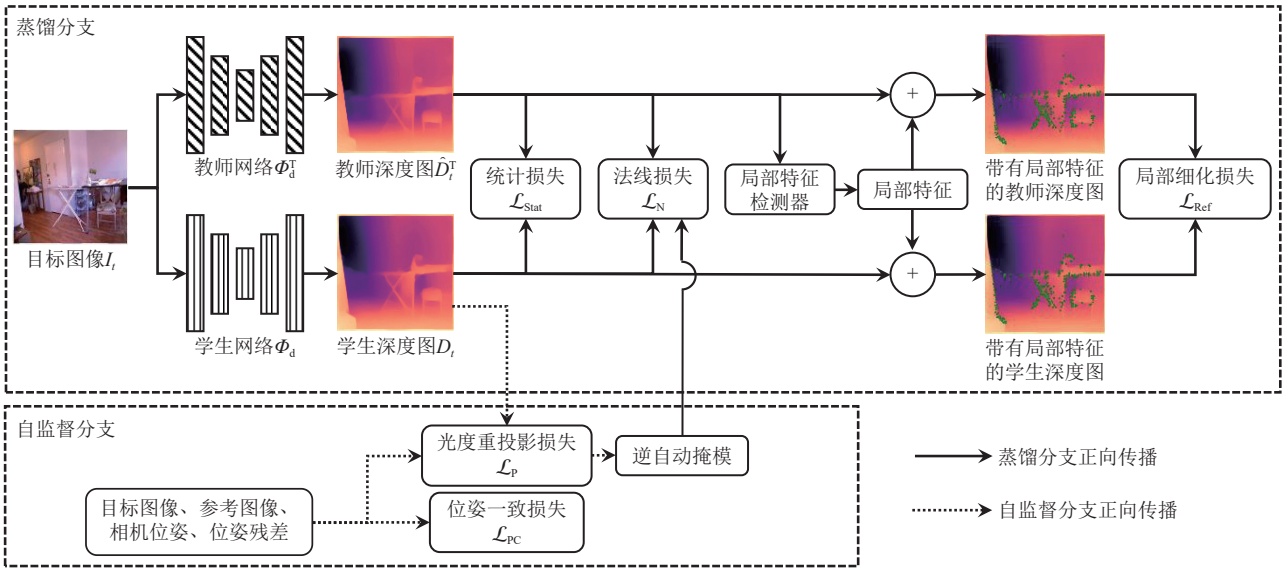


Fig. 3 Overveiw of LoFtDepth method

图 3 LoFtDepth 方法总览

自监督分支则以第 2 节介绍的自监督单目深度估计方法为主要框架。其中, 深度估计网络为蒸馏分支中的学生网络。根据此框架, 引入位姿残差估计[23], 并根据图像序列中相机运动的连续性, 使用位姿残差估计得到的位姿残差与相机位姿网络预测的相机位姿计算相机位姿一致损失, 以适应室内场景中多变的相机位姿。

3.2 局部特征引导的深度估计细化

蒸馏分支以 DPT[12] 作为教师网络Φ_d^T, 并以自监督深度估计网络作为学生网络Φ_d。学生网络Φ_d为编码器-解码器架构, 以卷积神经网络作为编码器, 并使用上采样和卷积层作为解码器进行解码得到深度图。将目标图像I_t同时输入给教师网络和学生网络得到对应的教师深度图D_t^T及学生深度图D_t。由于教师网络生成的是相对深度图, 使用最小二乘法预测尺度和偏移因子对齐教师深度图和学生深度图, 得到对齐的教师深度图$\hat{D}_t^T = scale \times D_t^T + shift$(后续无特别说明, 教师深度图指的是$\hat{D}_t^T$)。

局部特征引导的深度估计细化流程如图 4 所示, 利用在深度图上提取到的局部特征作为边界点。围绕局部特征, 最小化教师深度图和学生深度图在边界点邻域内的局部深度差异。

具体地, 首先在教师深度图$\hat{D}_t^T$上提取局部特征, 得到局部特征集合$\mathcal{S}_t^*$。

$$\mathcal{S}_t^* = f_{LF}(\hat{D}_t^T), \tag{8}$$

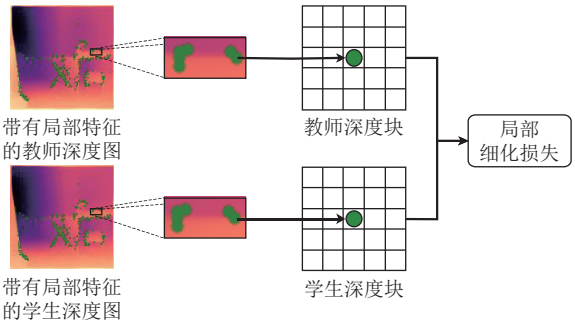


Fig. 4 Local feature guided depth estimation refinement

图 4 局部特征引导的深度估计细化

其中 f_{LF} 表示局部特征检测器，LoFtDepth 取经典的局部特征检测方法 SIFT[36] 作为默认的局部特征检测器。在实验部分，额外将 DSO[37] 和基于深度神经网络的方法 Key.Net[38] 作为局部特征检测器，以检验不同局部特征检测器对于 LoFtDepth 性能的影响。

接着将每个局部特征点 $p_i \in \mathcal{S}_t^*$ 作为边界点，以该局部特征点 p_i 为中心，分别从教师深度图 $\hat{D}_t^{\mathrm{T}}$ 和学生深度图 D_t 中提取大小为 $m \times m$ 的深度块 $\hat{D}_{t,p_i}^{\mathrm{T}}$ 和 D_{t,p_i} 计算 $\hat{D}_{t,p_i}^{\mathrm{T}}$ 和 D_{t,p_i} 在对数空间上的尺度不变损失（scale-invariant loss）作为局部细化损失 $\mathcal{L}_{\mathrm{Ref}}$。局部细化损失 $\mathcal{L}_{\mathrm{Ref}}$ 计算方式如式（9）所示。

$$\mathcal{L}_{\mathrm{Ref}} = \frac{1}{|\mathcal{S}_t^*|} \sum_{p_i \in \mathcal{S}_t^*} Diff_{p_i}, \tag{9}$$

$$Diff_{p_i} = \left(\frac{1}{m^2} \sum_{j=1}^{m} \sum_{k=1}^{m} diff_{p_i,j,k}^2 - \left(\frac{1}{m^2} \sum_{j=1}^{m} \sum_{k=1}^{m} diff_{p_i,j,k} \right)^2 \right), \tag{10}$$

其中 $diff_{p_i,j,k} = |\ln(D_{t,p_i,j,k}) - \ln(\hat{D}_{t,p_i,j,k}^{\mathrm{T}})|$ 表示深度块 $\hat{D}_{t,p_i}^{\mathrm{T}}$ 和 D_{t,p_i} 在第 j 行第 k 列位置上的深度误差。$|\mathcal{S}_t^*|$ 表示 $\mathcal{S}_t^*$ 中局部特征的个数。深度块的大小 $m = 5$。通过最小化局部细化损失 $\mathcal{L}_{\mathrm{Ref}}$，学生网络 Φ_{d} 能够学习教师网络 $\Phi_{\mathrm{d}}^{\mathrm{T}}$ 在边界区域的深度知识，提升深度图的结构化程度。

在局部特征引导的深度估计细化中，选择在深度图上提取局部特征，而不是在 RGB 图像上提取局部特征。这是因为在 RGB 图像中提取的局部特征往往包含深度无关特征，如阴影边缘或物体表面的纹理图案。它们在 RGB 图像中变化显著，但在深度图中变化平滑，如图 5 中的红色矩形框所示。这些深度无关特征在知识蒸馏时不能够提供显著的帮助。相反地，在深度图上提取的局部特征是具有明显深度变化的深度相关特征。这些特征能够在知识蒸馏中对学生网络进行更有效的约束。为验证上述想法，在实验部分，分别在 RGB 图像和深度图上提取局部特征进行训练，评估不同局部特征检测域对深度估计性能的影响。



Fig. 5 Differences in local feature detection on RGB images and depth maps

图 5 在 RGB 图像和深度图上进行局部特征检测的差异

3.3 逆自动掩模加权的法线损失

逆自动掩模加权的法线损失 $\mathcal{L}_{\mathrm{N}}$ 计算过程如图 6 所示。该损失鼓励学生模型和教师模型在无纹理区域的法线方向相似。

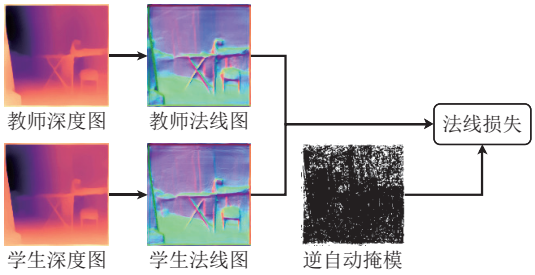


Fig. 6 Inverse auto mask weighted normal loss

图 6 逆自动掩模加权的法线损失

首先分别从教师深度图 $\hat{D}_t^{\mathrm{T}}$ 和学生深度图 D_t 生成各自的表面法线图 $\hat{N}_t^{\mathrm{T}}$ 和 N_t，计算二者之间的余弦相似度作为基础的法线损失。接着使用逆自动掩模 $\neg\mathcal{M}_{\mathrm{a}}$ 加权 $\mathcal{L}_{\mathrm{N}}$。最终法线损失 $\mathcal{L}_{\mathrm{N}}$ 的定义如式（11）所示。

$$\mathcal{L}_{\mathrm{N}} = (\neg\mathcal{M}_{\mathrm{a}})(1 - CosSim(N_t, \hat{N}_t^{\mathrm{T}})), \tag{11}$$

其中 $CosSim(N_t, \hat{N}_t^{\mathrm{T}})$ 表示 N_t 和 $\hat{N}_t^{\mathrm{T}}$ 的余弦相似度。

在室内场景中，自动掩模 $\mathcal{M}_{\mathrm{a}}$ 除去的像素大多位于无纹理区域，如图 7 所示。可能原因是：由式（4）



Fig. 7 Characteristics of auto mask in indoor scenes

图 7 室内场景中自动掩模的特点

可以得知, 当光度误差 $pe(I_t, I_{t'})$小于等于光度误差 $pe(I_t, I_{t'\to t})$时, 像素会被自动掩模 $\mathcal{M}_a$去除。而在室内场景中, 大面积无纹理区域中的像素外观高度一致, 使得这 2 组光度误差近似, 从而导致该区域内的像素被自动掩模 $\mathcal{M}_a$排除。根据这一特点, 可以通过逆自动掩模 $\neg\mathcal{M}_a$确定无纹理区域。所以, 使用逆自动掩模 $\neg\mathcal{M}_a$加权法线损失, 确保法线损失主要作用于无纹理区域。

3.4 位姿一致损失

在自监督分支中引入位姿残差估计[23-24]。位姿残差估计可以认为是重复参考图像的重投影过程。根据位姿残差估计网络 Φ_p^{Res}预测的相机位姿残差 $T_{t\to t'}^{Res} = T_{t\to(t'\to t)}$, 将重投影图像 $I_{t'\to t}$再次重投影到 I_t的平面上得到最终的重投影图像 $I_{t'\to t}^*$。取目标图像 I_t和最终的重投影图像 $I_{t'\to t}^*$计算光度重投影损失, 如式(12)所示。

$$\mathcal{L}_p = \mathcal{M}_a \min_{I_{t'}} pe(I_t, I_{t'\to t}^*)。 \tag{12}$$

考虑到图像序列的相机运动具有连续性, 提出位姿一致损失 $\mathcal{L}_{PC}$(pose consistency loss)约束相机位姿估计网络 Φ_p, 其计算流程如图 8 所示。

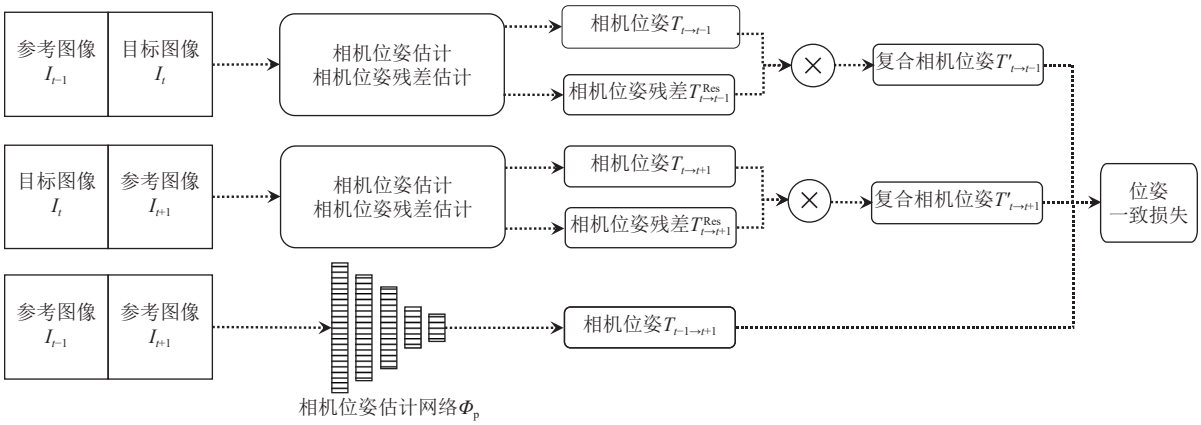


Fig. 8 Pose consistency loss

图 8 位姿一致损失

具体的计算步骤为: 1)利用自监督分支得到的相机位姿 $T_{t\to t'}$和相机位姿残差 $T_{t\to t'}^{Res}$相乘计算得到目标图像和参考图像之间的复合相机位姿 $T'_{t\to t'}$; 2)使用相机位姿估计网络 Φ_p估计 2 幅参考图像 I_{t-1}和 I_{t+1}之间的相机位姿 $T_{t-1\to t+1}$; 3)根据相机运动的连续性, 以 $T_{t-1\to t+1}$作为 $T'_{t\to t+1}$和 $T'_{t-1\to t}$乘积的伪真值计算位姿一致损失 $\mathcal{L}_{PC}$, 如式(13)所示。

$$\mathcal{L}_{PC} = \|\boldsymbol{R}'_{t\to t+1} \otimes \boldsymbol{R}'_{t-1\to t} - \boldsymbol{R}_{t-1\to t+1}\|_1 + 0.01\times\mathcal{L}_{reg}, \tag{13}$$

其中 $\boldsymbol{R}'_{t\to t+1}$和 $\boldsymbol{R}'_{t-1\to t}$分别表示复合相机位姿 $T'_{t\to t+1}$和 $T'_{t-1\to t} = (T'_{t\to t-1})^{-1}$的旋转矩阵, $\boldsymbol{R}_{t-1\to t+1}$表示相机位姿 $T_{t-1\to t+1}$的旋转矩阵。$\mathcal{L}_{reg} = \sum(1 - \|\boldsymbol{I} - \boldsymbol{R}\|_1), \boldsymbol{R} \in \{\boldsymbol{R}'_{t\to t+1}, \boldsymbol{R}'_{t-1\to t}, \boldsymbol{R}_{t+1\to t-1}\}$是正则项, 用于防止相机位姿估计网络预测不变的相机位姿。

在深度估计训练过程中, 相机的旋转运动可视为噪声, 而平移运动则能提供有效的监督信号[21]。为了减少不准确旋转运动预测对深度估计的负面影响, 同时不影响平移运动对训练的积极作用, 因此仅针对相机位姿的旋转矩阵进行约束。

3.5 总损失函数

在蒸馏分支中, 额外使用了统计损失 $\mathcal{L}_{Stat}$[4] 进行知识蒸馏, 该损失构成如式(14)所示。

$$\mathcal{L}_{Stat} = 1 - SSIM(D_t, \hat{D}_t^T)。 \tag{14}$$

LoFtDepth 的总损失函数由光度重投影损失 $\mathcal{L}_p$、平滑损失 $\mathcal{L}_s$、统计损失 $\mathcal{L}_{Stat}$、局部细化损失 $\mathcal{L}_{Ref}$、逆自动掩模加权的法线损失 $\mathcal{L}_N$以及位姿一致损失 $\mathcal{L}_{PC}$构成, 具体如式(15)所示。

$$\mathcal{L} = \mathcal{L}_p + \omega_1\mathcal{L}_s + \omega_2\mathcal{L}_{Stat} + \omega_3\mathcal{L}_{Ref} + \omega_4\mathcal{L}_N + \omega_5\mathcal{L}_{PC}, \tag{15}$$

其中 $\omega_i(i = 1,2,3,4,5)$表示损失函数权重, 与先前自监督单目深度估计方法相同[4], 将 ω_1和 ω_2分别设置为 0.001 和 0.1。根据多次实验结果, 将 ω_3, ω_4, ω_5依次设置为 1, 0.01, 1。

4 实 验

4.1 实验数据及评估指标

本文选择在仿真数据 SimSIN[4] 和虚拟公寓 VA(virtual apartment)[4], 以及真实数据 NYUv2[39] 和 ScanNet[40] 上进行实验。SimSIN 由多达 1 000 种不同仿真环境中的 50 万对双目立体图像组成。与 DistDepth[4] 相同, SimSIN 中的所有图像都被用于训练。VA 包

含 3 524 张图像，其场景为包含了许多具有复杂结构的物体的公寓，用于评估模型性能。NYUv2 是目前单目深度估计的主要室内数据集。该数据集的 RGB 图像和对应的深度图使用 Microsoft Kinect 进行采集，共包含 26 295 幅训练图像和 654 幅测试图像。ScanNet 提供了 1 513 个室内 RGB-D 视频。使用 Wang 等人[20]发布的 ScanNet 的测试数据，共包含来自不同场景的 533 张图像。

评估指标包括平均绝对误差（mean absolute error, *MAE*）、相对误差（absolute relative error, *AbsRel*）、均方根误差（root mean square error, *RMSE*）和对数均方根误差*RMSElog*，以及δ计算有效像素占所有像素百分比的深度准确度。具体指标的计算方式为：

$$MAE = \frac{1}{|N|}\sum_{i\in N}|d_i - d_i^*|, \tag{16}$$

$$AbsRel = \frac{1}{|N|}\sum_{i\in N}\frac{|d_i - d_i^*|}{d_i^*}, \tag{17}$$

$$RMSE = \sqrt{\frac{1}{|N|}\sum_{i\in N}\|d_i - d_i^*\|^2}, \tag{18}$$

$$RMSElog = \sqrt{\frac{1}{|N|}\sum_{i\in N}\|\ln(d_i) - \ln(d_i^*)\|^2}, \tag{19}$$

$$\delta_k = \max\left(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}\right) < 1.25^k, \tag{20}$$

其中d_i表示i像素上的预测深度，d_i^*表示i像素上的真实深度，N表示用于计算评估指标的像素集合。

4.2 实验设置

使用 1 个 Nvidia RTX 3090 GPU 进行实验。输入图像的大小设置为 256×256，批量大小（batch size）为 16。利用 AdamW[41]作为优化器。在网络结构方面，以 ResNet[42]作为自监督单目深度估计网络的编码器，同时考虑精度和实用性，主要以 ResNet-152 作为编码器。为了实验的公平性，额外提供以 ResNet-18 作为编码器的实验结果。以 DPT-Hybrid[12]作为教师网络。将 SIFT[36]作为默认的局部特征检测器f_{LF}，提取的局部特征数量$|\mathcal{S}_i^*|$设置为 300。在 NYUv2[39]上训练时，总学习轮数为 30，学习率（learning rate）设置为 1E−4，在第 20 轮的时候将学习率下降为 1E−5，总训练时长约为 15.35 h。在 SimSIN[4]上训练时，总学习轮数为 10，学习率设置为 2E−4，在第 8 轮的时候将学习率降为 2E−5，总训练时长约为 4 天。以 ResNet-152 作为编码器的学生网络，测试时的推理速度为 33 fps。

4.3 基于仿真数据的对比实验

首先在 SimSIN 上训练模型，然后在 VA 数据集上评估该模型的性能，并将其与先前的自监督单目深度估计方法进行比较。

表 1 展示了在 VA 数据集上的定量结果。通过定量结果得知，与先前的自监督方法相比，LoFtDepth 在 VA 数据集上取得了更好的性能。

Table 1 Quantitative Results on VA Dataset

表 1 VA 数据集上的定量结果

方法	*MAE*↓	*AbsRel*↓	*RMSE*↓	*RMSElog*↓
MonoDepth2[5]	0.295	0.203	0.432	0.251
DepthHints[17]	0.291	0.197	0.427	0.248
DistDepth[4]	0.276	0.188	0.394	0.227
LoFtDepth（本文）	**0.254**	**0.176**	**0.374**	**0.213**

注：“↓”表示指标越低，模型性能越好。黑体数值表示最优结果。

图 9 展示了在 VA 数据集上的定性结果。当物体靠近地面或远处背景的时候（矩形框标记），DistDepth 产生的深度图物体边界模糊，而 LoFtDepth 预测的深度图则保持清晰的物体边界。这种区别突出表现了利用局部特征作为边界点排除被错误定义的边界对于结构知识蒸馏的重要性。

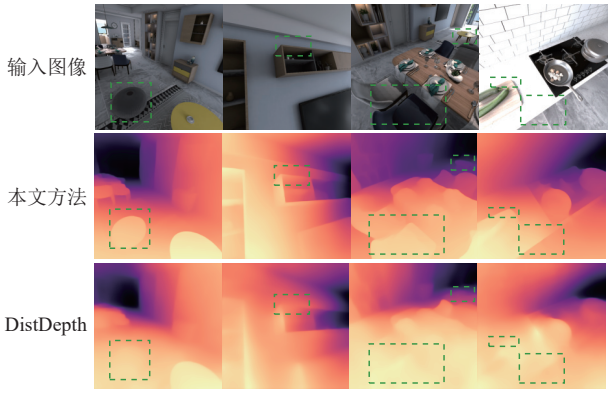


Fig. 9 The qualitative results on VA dataset

图 9 VA 数据集上的定性结果

4.4 基于真实数据的对比实验

在 NYUv2 数据集上以 2 种方式进行训练：1）在 NYUv2 上直接训练；2）先在 SimSIN 上进行预训练，然后在 NYUv2 上进行微调，并将它们与近年来最先进的室内自监督方法进行比较。同时，测试了仅在 SimSIN 上训练的模型的效果，以探索 LoFtDepth 从仿真数据到真实数据的泛化能力。

表 2 展示了在 NYUv2 数据集上的定量结果。从定量结果中可以看出，LoFtDepth 的性能优于先前针对室内环境的自监督方法。此外，表 2 的最后 2 行表明 LoFtDepth 具有更好的从仿真数据泛化到真实数据的性能。

Table 2 Quantitative Results on NYUv2 Dataset

表 2 在 NYUv2 数据集上的定量结果

方法	骨干网络	方法来源	*AbsRel*↓	*RMSE*↓	δ_1/%↑	δ_2/%↑	δ_3/%↑
MonoDepth2[5]	ResNet-18	ICCV 2019	0.172	0.617	73.9	94.4	98.7
P²Net+PP[20]	ResNet-18	ICCV 2021	0.147	0.553	80.4	95.2	98.7
StructDepth[18]	ResNet-18	CVPR 2021	0.142	0.540	81.3	95.4	98.8
ARN[21]	ResNet-18	TPAMI 2021	0.138	0.532	82.0	95.6	98.9
MonoIndoor[23]	ResNet-18	ICCV 2021	0.134	0.526	82.3	95.8	98.9
MonoIndoor++[24]	ResNet-18	TCSVT 2022	0.132	0.517	83.4	95.8	99.0
HI-Net[22]	ResNet-18	RAL 2022	0.130	0.522	82.8	95.9	98.9
GasMono[33]	ResNet-18	ICCV 2023	0.141	0.533	81.2	95.7	98.9
F²Depth+PP[43]		EAAI 2024	0.153	0.569	78.7	95.0	98.7
GAM-Depth[10]	ResNet-18	ICRA 2024	0.131	0.507	83.6	96.0	99.0
LoFtDepth-18	ResNet-18	本文	0.130	0.498	83.5	96.6	99.2
LoFtDepth-152	ResNet-152	本文	**0.121**	**0.474**	**85.9**	**97.1**	**99.2**
GasMono[33]	MPViT-base	ICCV 2023	0.113	0.459	87.1	97.3	99.2
DistDepth†[4]	ResNet-152	CVPR 2022	0.130	0.517	83.2	96.3	99.0
LoFtDepth-152†	ResNet-152	本文	**0.119**	**0.470**	**86.0**	**97.1**	**99.2**
DistDepth‡[4]	ResNet-152	CVPR 2022	0.165	0.569	77.7	93.4	97.9
LoFtDepth-152‡	ResNet-152	本文	**0.152**	**0.545**	**80.0**	**94.4**	**98.3**

注：“↓”表示指标越低，模型性能越好；“↑”表示指标越高，模型性能越好；“†”表示方法先在 SimSIN 上预训练，然后在 NYUv2 上微调；“‡”表示方法仅在 SimSIN 上训练；黑体数值表示最优结果。

图 10 展示了 NYUv2 数据集上的定性结果。与先前的自监督方法相比，LoFtDepth 生成的深度图具有明显的物体边界。另外，LoFtDepth 生成的深度图具有与真实深度图更相似的深度分布。这表明 LoFtDepth 具有生成准确且高度结构化的深度图的能力。

为了进一步评估 LoFtDepth 的泛化能力，在 ScanNet 上进行了零样本实验。表 3 展示了 ScanNet 零样本实验的结果，所有的方法都只在 NYUv2 上进行训练。结果表明，LoFtDepth 的零样本泛化能力优于先前的自监督方法。

为了评估 LoFtDepth 对室外场景的适应性，在室外数据集 KITTI 上进行了实验，图像的大小为 192×

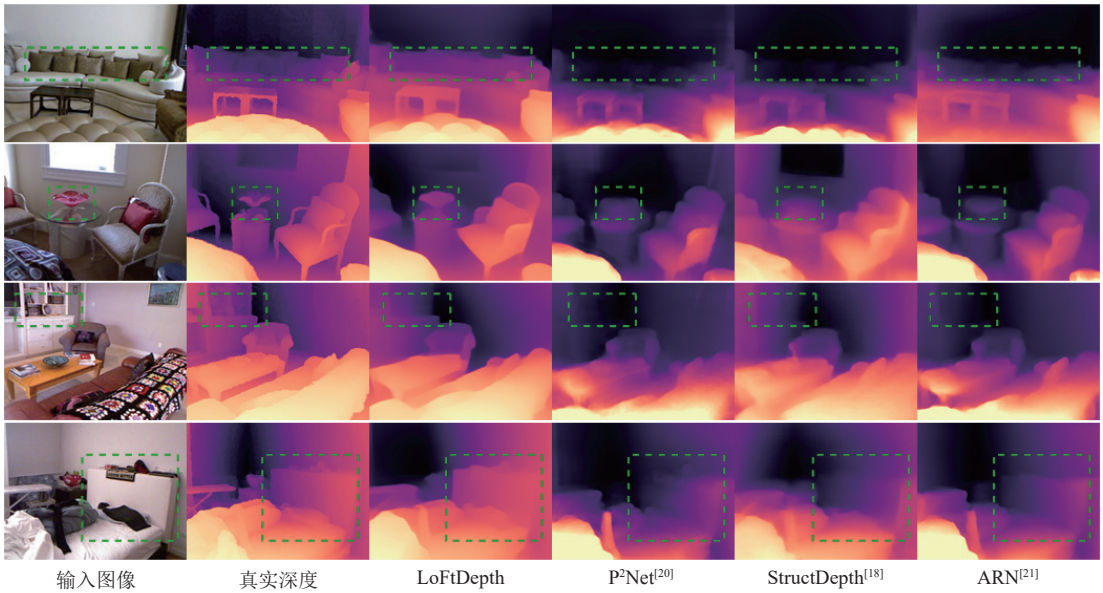


Fig. 10 The qualitative results on NYUv2 dataset

图 10 NYUv2 数据集上的定性结果

Table 3 Zero-Shot Quantitative Results on ScanNet Dataset

表 3 ScanNet 数据集上的零样本定量结果

方法	*AbsRel*↓	*RMSE*↓	δ_1/%↑	δ_2/%↑	δ_3/%↑
MonoDepth2[5]	0.191	0.451	69.3	92.6	98.3
P^{2}Net[20]	0.175	0.420	74.0	93.2	98.2
SC-Depth[15]	0.169	0.392	74.9	93.8	98.3
StructDepth[18]	0.165	0.400	75.4	93.9	98.5
ARN[21]	0.156	0.361	78.1	94.7	98.7
GAM-Depth[10]	0.159	0.400	77.0	93.9	95.8
LoFtDepth-18（本文）	0.148	0.352	79.9	95.7	99.1
LoFtDepth-152（本文）	**0.137**	**0.332**	**82.2**	**96.4**	**99.2**

注：“↓”表示指标越低，模型性能越好；“↑”表示指标越高，模型性能越好；黑体数值表示最优结果。

640。表 4 展示了 KITTI 数据集上的定量实验结果，说明 LoFtDepth 在室外场景下也有适度的提升。

4.5 消融实验

为了更好地了解 LoFtDepth 的每个部分对提升深度估计性能的贡献，在 NYUv2 数据集上进行消融实验。

1）知识蒸馏中损失函数的不同组合。首先探索结构蒸馏中不同的损失函数组合对模型性能的影响。具体比较了如下方法：①基线模型；②只有统计损失的基线模型；③有统计损失和法线损失的基线模型；④包含结构蒸馏中所有的损失但在 RGB 图上提取局部特征的基线模型；⑤包含结构蒸馏中所有的损失且在深度图上提取局部特征的基线模型。图 11 展示了不同损失函数组合的表面法线图。经过改进后的方法在物体边界处呈现出较大的表面法线变化，这一现象表明，在知识蒸馏中提出的逆自动掩模加权

Table 4 Quantitative Results on KITTI Dataset

表 4 KITTI 数据集上的定量结果

方法	*AbsRel*↓	*RMSE*↓	δ_1/%↑	δ_2/%↑	δ_3/%↑
MonoDepth2[5]	0.115	4.863	87.7	95.9	98.1
PackNet[31]	0.111	4.788	86.4	95.4	98.0
SC-Depth[15]	0.119	4.950	86.3	95.7	98.1
DepthHints[17]	**0.106**	4.695	87.5	95.8	98.0
SC-Depth v3[19]	0.118	4.709	86.4	96.0	**98.4**
LoFtDepth-18（本文）	0.109	**4.671**	**88.1**	**96.1**	98.2

注：“↓”表示指标越低，模型性能越好；“↑”表示指标越高，模型性能越好；黑体数值表示最优结果。

的表面法线损失以及局部特征引导的深度估计细化具备有效性。表 5 展示了方法①~⑤的实验结果，该结果体现了法线损失和局部细化损失有效地提升了深度估计模型的性能。方法④和方法⑤的实验结果说明在深度图上提取局部特征更有效于深度域的结构蒸馏。

2）不同的局部特征检测方法。为评估不同的局部特征检测方法对模型性能的影响，测试了 SIFT[36]、DSO[37] 和基于深度学习的 Key.Net[38] 这 3 种不同的局部特征检测方法。表 5 中的方法⑤~⑦分别展示了使用这 3 种不同局部特征检测方法的实验结果。该实验结果表明：虽然方法⑤~⑦使用了不同的局部特征检测器，得到的结果几乎相同，说明 LoFtDepth 并不局限于特定类型的局部特征检测器。因不同的局部特征检测方法存在差异，将不同局部特征检测方法提取到的局部特征进行对比，同时与 DistDepth[4] 中的边缘检测比较，进一步研究局部特征的置信以及不同局部特征检测方法之间的差异。图 12 展示了边缘检测及不同局部特征检测方法的结果对比。首先

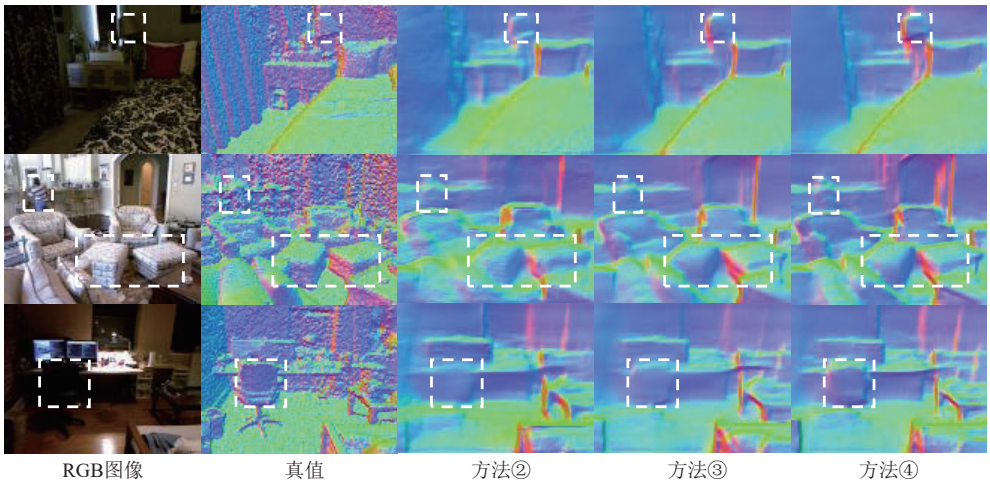


Fig. 11 Surface normal maps for different combinations of loss functions

图 11 不同损失函数组合的表面法线图

Table 5 Quantitative Results of Ablation Studies

表 5 消融实验的定量结果

方法	知识蒸馏				检测器	$\neg\mathcal{M}_a$	位姿残差	$\mathcal{L}_{PC}$	*AbsRel*↓	*RMSE*↓	δ_1/%↑	δ_2/%↑	δ_3/%↑
	$\mathcal{L}_{Stat}$	$\mathcal{L}_N$	$\mathcal{L}_{Ref}$	检测域									
①									0.172	0.617	73.9	94.4	98.7
②	√								0.144	0.532	80.7	96.4	99.1
③	√	√				√			0.135	0.514	82.9	95.7	98.8
④	√	√	√	RGB	SIFT	√			0.130	0.503	84.0	94.4	99.0
⑤	√	√	√	深度图	SIFT	√			**0.127**	**0.494**	**84.7**	**96.6**	**99.1**
⑤	√	√	√	深度图	SIFT	√			**0.127**	0.494	**84.7**	96.6	99.1
⑥	√	√	√	深度图	DSO	√			0.128	0.494	84.3	**96.8**	**99.2**
⑦	√	√	√	深度图	Key.Net	√			0.128	0.494	84.5	96.5	99.1
⑤	√	√	√	深度图	SIFT	√			**0.127**	**0.494**	**84.7**	**96.6**	**99.1**
⑧	√	√	√	深度图	SIFT	×			0.130	0.500	84.0	96.5	99.0
⑨							√		0.134	0.521	83.1	96.0	99.0
⑩							√	√	0.131	0.508	83.8	96.1	99.0
⑪	√	√	√	深度图	SIFT	√	√		0.123	0.480	85.2	96.8	99.2
⑫	√	√	√	深度图	SIFT	√	√	√	**0.121**	**0.474**	**85.9**	**97.1**	**99.2**

注：“↓”表示指标越低，模型性能越好；“↑”表示指标越高，模型性能越好；“√”表示包含了该模块；黑体数值表示最优结果。

进行边缘检测与局部特征检测的对比。在正常情况下，边缘检测方法和局部特征检测的结果在空间上有重合。也就是说，这里提取到的局部特征大多为深度相关的物体边界点。但是当图像中包含远处场景的时候，由于这些区域的深度梯度数值较大，而边缘检测将深度梯度的前 5% 的区域定义为边界，导致边缘检测误将这部分区域判定为边界，而局部特征检测因为关注局部的深度变化，则不会有这个问题。其次比较不同局部特征检测方法。不同局部特征检测器提取的局部特征也不同，特地选取经典方法的 SIFT[36] 和 DSO[37]，以及基于深度学习方法的 Key.Net[38] 进行实验。从图 12 中可以看出，虽然 3 个方法提取到的局部特征都分布在物体边界上，但是与 SIFT[36] 提取到的局部特征相比，DSO[37] 提取到的局部特征更加集中于边界，Key.Net[38] 检测的局部特征则相对较为稀疏，这些差异导致使用不同局部特征检测器训练出来的网络结果有略微区别。

3）逆自动掩模加权。为检测逆自动掩模加权法线损失的有效性，评估了未使用逆自动掩模加权法线损失的模型（方法⑧），并将其与方法⑤进行比较。表 5 展示了二者的实验结果对比，表明使用逆自动掩模加权法线损失可以提高模型的性能。

4）位姿一致损失。为了验证位姿一致损失对于深度估计性能提升的有效性。比较结合位姿残差估

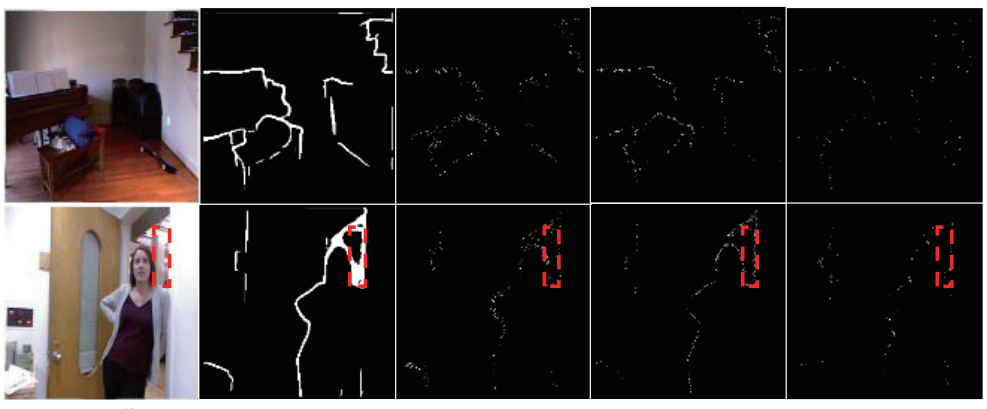


Fig. 12 Comparison of edge detection and different local feature detection methods

图 12 边缘检测及不同局部特征检测方法的对比

计的基线模型(方法⑨)、结合位姿残差估计并添加位姿一致损失的基线模型(方法⑩)、结合结构蒸馏和位姿残差估计的基线模型(方法⑪)、完整的LoFtDepth的定量结果(方法⑫)。表5中方法⑨~⑫的实验结果表明无论是否结合知识蒸馏，位姿一致损失都能够有效地提升模型的性能。

5 结　论

针对进一步提升室内场景自监督单目深度估计精度的问题，提出训练过程的增强改善方法。结合知识蒸馏，以教师深度图的局部特征作引导，使自监督深度估计网络学习教师网络的边界深度分布，提升模型的边界深度精度；使用逆自动掩模加权的法线损失，对齐教师网络和自监督深度估计网络在无纹理区域的法线方向，提升自监督网络在该区域的深度估计准确性；引入位姿残差估计，并根据图像序列相机运动的连续性改善位姿估计，减小训练误差，提升深度估计精度。由此构成的LoFtDepth方法在多个公开的室内数据集上取得了显著的性能提升，将相对误差*AbsRel*降低至0.121，对比DistDepth提升约7个百分点，并能够在复杂的室内场景生成深度准确且边界清晰的深度图。值得一提的是，深度估计中引入局部特征引导以及对无纹理区域的适应性方法并不局限于LoFtDepth本身，在其他深度估计方法或不同场景同样具有应用潜力。

作者贡献声明：艾浩军和张锋共同提出算法思路和实验方案设计；张锋完成实验并撰写论文；吕鹏飞、唐雪华和王中元提出指导意见并修改论文。

参　考　文　献

[1] Kerl C, Sturm J, Cremers D. Dense visual SLAM for RGB-D cameras[C]//Proc of the 26th IEEE/RSJ Int Conf on Intelligent Robots and Systems (IROS). Piscataway, NJ: IEEE, 2013: 2100−2106

[2] Holynski A, Kopf J. Fast depth densification for occlusion-aware augmented reality[J]. ACM Transactions on Graphics, 2018, 37(6): 1−11

[3] Im S, Ha H, Choe G, et al. Accurate 3D reconstruction from small motion clip for rolling shutter cameras[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(4): 775−787

[4] Wu C, Wang Jialiang, Hall M, et al. Toward practical monocular indoor depth estimation[C]//Proc of the 40th IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2022: 3814−3824

[5] Godard C, Aodha M O, Firman M, et al. Digging into self-supervised monocular depth estimation[C]//Proc of the 37th IEEE/CVF Int Conf on Computer Vision. Los Alamitos, CA: IEEE Computer Society, 2019: 3828−3838

[6] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]//Proc of the 28th Conf and Workshop on Neural Information Processing System (NIPS). Cambridge, MA: MIT, 2014: 2366−2374

[7] Laina I, Rupprecht C, Belagiannis V, et al. Deeper depth prediction with fully convolutional residual networks[C]//Proc of the 4th Int Conf on 3D Vision (3DV). Piscataway, NJ: IEEE, 2016: 239−248

[8] Fu Huan, Gong Mingming, Wang Chaohui, et al. Deep ordinal regression network for monocular depth estimation[C]//Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2018: 2002−2011

[9] Zhou Tinghui, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]//Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2017: 1851−1858

[10] Cheng Anqi, Yang Zhiyuan, Zhu Haiyue, et al. GAM-Depth: Self-supervised indoor depth estimation leveraging a gradient-aware mask and semantic constraints[C]//Proc of the 41st IEEE Int Conf on Robotics and Automation (ICRA). Piscataway, NJ: IEEE, 2024: 5367−5374

[11] Bhat S F, Alhashim I, Wonka P. Adabins: Depth estimation using adaptive bins[C]//Proc of the 39th IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2021: 4009−4018

[12] Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2021: 12179−12188

[13] Piccinelli L, Sakaridis C, Yu Fisher. iDisc: Internal discretization for monocular depth estimation[C]//Proc of the 41st IEEE/CVF Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2023: 21477−21487

[14] 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 (in Chinese)
(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)

[15] 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

[16] Li Rui, He Xiantuo, Zhu Yu, et al. Enhancing self-supervised monocular depth estimation via incorporating robust constraints[C]//Proc of the 28th ACM Int Conf on multimedia. New York: ACM, 2020: 3108−3117

[17] Watson J, Firman M, Brostow G J, et al. Self-supervised monocular depth hints[C]//Proc of the 37th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 2162−2171

[18] Li Boying, Huang Yuan, Liu Zeyu, et al. StructDepth: Leveraging the

structural regularities for self-supervised indoor depth estimation[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2021: 12663−12673

[19] Sun Libo, Bian Jiawang, Zhan Huangying, et al. SC-Depth v3: Robust self-supervised monocular depth estimation for dynamic scenes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 46(1): 497−508

[20] Wang Bing, Chen Changhao, Cui Zhaopeng, et al. P^2-Net: Joint description and detection of local features for pixel and point matching[C]//Proc of the 18th IEEE/CVF Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2021: 16004−16013

[21] 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, 2021, 44(12): 9802−9813

[22] Wu Guanghui, Li Kunhong, Wang Longguang, et al. HI-Net: Boosting self-supervised indoor depth estimation via pose optimization[J]. IEEE Robotics and Automation Letters, 2023, 8(1): 224−231

[23] Ji Pan, Li Runze, Bhanu B, et al. MonoIndoor: Towards good practice of self-supervised monocular depth estimation for indoor environments[C]//Proc of the IEEE/CVF Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2021: 12787−12796

[24] Li Runze, Ji Pan, Xu Yi, et al. MonoIndoor++: Towards better practice of self-supervised monocular depth estimation for indoor environments[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(2): 830−846

[25] Yang Lihe, Kang Bingyi, Huang Zilong, et al. Depth anything: Unleashing the power of large-scale unlabeled data[C]//Proc of the 42nd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2024: 10371−10381

[26] Yang Lihe, Kang Bingyi, Huang Zilong, et al. Depth anything V2[J]. arXiv preprint, arXiv: 2406.09414, 2024

[27] Ke Bingxin, Obukhov A, Huang Shengyu, et al. Repurposing diffusion-based image generators for monocular depth estimation[C]//Proc of the 42nd IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2024: 9492−9502

[28] Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning robust visual features without supervision[J]. arXiv preprint, arXiv: 2304.07193, 2024

[29] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proc of the 40th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2022: 10674−10685

[30] Godard C, Aodha M O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]//Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2017: 270−279

[31] Guizilini V, Ambrus R, Pillai S, et al. 3D packing for self-supervised monocular depth estimation[C]//Proc of the 38th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2020: 2485−2494

[32] Zhang Ning, Nex F, Vosselman G, et al. Lite-mono: A lightweight CNN and transformer architecture for self-supervised monocular depth estimation[C]//Proc of the 41st IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2023: 18537−18546

[33] Zhao Chaoqiang, Poggi M, Tosi F, et al. GasMono: Geometry-aided self-supervised monocular depth estimation for indoor scenes[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2023: 16209−16220

[34] Wang Zhou, Bovik A C, Sheikh H R, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600−612

[35] Wang Chaoyang, Buenaposada J M, Zhu Rui, et al. Learning depth from monocular videos using direct methods[C]//Proc of the 36th IEEE Conf on Computer Vision and Pattern Recognition. Los Alamitos, CA: IEEE Computer Society, 2018: 2022−2030

[36] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91−110

[37] Engel J, Koltun V, Cremers D. Direct sparse odometry[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(3): 611−625

[38] Barroso-Laguna A, Riba E, Ponsa D, et al. Key.Net: Keypoint detection by handcrafted and learned CNN filters[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision (ICCV). Los Alamitos, CA: IEEE Computer Society, 2019: 5836−5844

[39] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from RGBD images[C]//Proc of the 12th European Conf on Computer Vision (ECCV). Berlin: Springer, 2012: 746−760

[40] Dai A, Chang A X, Savva M, et al. ScanNet: Richly-annotated 3D reconstructions of indoor scenes[C]//Proc of the 35th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2017: 5828−5839

[41] Loshchilov I, Hutter F. Decoupled weight decay regularization[J]. arXiv preprint, arXiv: 1711.05101, 2017

[42] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proc of the 34th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA: IEEE Computer Society, 2016: 770−778

[43] Guo Xiaotong, Zhao Huijie, Shao Shuwei, et al. F^{2}Depth: Self-supervised indoor monocular depth estimation via optical flow consistency and feature map synthesis[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108391



Ai Haojun, born in 1972. PhD, associate professor. Senior member of CCF. His main research interests include computer vision, artificial intelligence, and deepfake detection.

艾浩军, 1972 年生。博士，副教授。CCF 高级会员。主要研究方向为计算机视觉、人工智能、深度伪造检测。



Zhang Feng, born in 2001. Master candidate. His main research interests include computer vision, deepfake detection, and depth estimation.

张　锋, 2001 年生。硕士研究生。主要研究方向为计算机视觉、深度伪造检测、深度估计。



Lü Pengfei, born in 1995. PhD candidate. His main research interests include computer vision and deepfake detection.

吕鹏飞, 1995 年生。博士研究生。主要研究方向为计算机视觉、深度伪造检测。



Tang Xuehua, born in 1977. PhD, lecturer. Her main research interests include geographic information science, spatial relationship representation reasoning, uncertainty modeling, trajectory data mining, and abnormal behavior pattern discovery.

唐雪华, 1977 年生。博士，讲师。主要研究方向为地理信息科学、空间关系表达推理、不确定性建模、轨迹数据挖掘、异常行为模式发现。



Wang Zhongyuan, born in 1972. PhD, professor. His main research interest includes multimedia technologies.

王中元, 1972 年生。博士，教授。主要研究方向为多媒体技术。