文章编号:1007－6735(2025)06－0735－12　DOI: 10.13255/j.cnki.jusst.20241104004

基于两阶段的轻量级深度估计方法

丁程威[1]，　魏国亮[2]
（1. 上海理工大学 光电信息与计算机工程学院，上海 200093；2. 上海理工大学 管理学院，上海 200093）

摘要：针对传统单目深度估计中相对估计方法丢失尺度信息、度量估计方法边缘精度不足，以及现有深度网络参数量大、计算成本高的问题，提出了一种两阶段融合框架 LacDepth。该框架旨在通过 2 个阶段来融合度量估计和相对估计方法。第一阶段，深度残差金字塔模块采用多尺度拉普拉斯残差补偿机制，通过高频特征增强策略有效提升边缘轮廓的几何保真度；第二阶段，轻量化吸引子驱动分类器构建三级级联深度区间预测网络，基于条件对数二项分布建立像素级概率密度函数，通过可微分加权实现相对深度值的亚区间微调。实验结果显示，LacDepth 在 KITTI 数据集上的综合表现最佳，平均相对误差为 0.059，参数量为 9.8×10^6，在精度与效率的平衡性方面展现出显著优势。

关键词：深度估计；多尺度特征融合；特征增强；轻量级网络；分类策略
中图分类号：TP 183　　**文献标志码：**A

A two-stage lightweight approach for depth estimation

DING Chengwei[1]，　WEI Guoliang[2]
(1. *School of Optical−Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*; 2. *Business School, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: To address the issues in traditional monocular depth estimation —where relative estimation methods lose scale information, metric estimation methods suffer from insufficient edge precision, and existing depth networks have large parameter counts and high computational costs—a two-stage fusion framework called LacDepth was proposed. This framework aimed to fuse metric and relative estimation methods through two stages. In the first stage, the deep residual pyramid module adopted a multi-scale Laplacian residual compensation mechanism and effectively improved the geometric fidelity of edge contours via a high-frequency feature enhancement strategy. In the second

收稿日期: 2024−11−04
基金项目: 国家自然科学基金资助项目(62273239)；上海市“科技创新行动计划”国内科技合作项目(20015801100)
第一作者: 丁程威(1999—)，男，硕士研究生．研究方向：机器视觉．E-mail：dingcw9590@163.com
通信作者: 魏国亮(1973—)，男，教授．研究方向：机器视觉、图像处理．E-mail：guoliang.wei@usst.edu.cn

引文格式: 丁程威，魏国亮．基于两阶段的轻量级深度估计方法[J]．上海理工大学学报，2025，47(6)：735-746.
Citation: DING Chengwei, WEI Guoliang．A two-stage lightweight approach for depth estimation[J]．Journal of University of Shanghai for Science and Technology, 2025, 47(6)：735-746.

stage, the lightweight attractor-driven classifier constructed a three-level cascaded depth interval prediction network, established a pixel-level probability density function based on the conditional log-binomial distribution, and realized sub-interval fine-tuning of relative depth values through differentiable weighting. Experimental results show that LacDepth achieves the best comprehensive performance on the KITTI dataset, with an average relative error of 0.059 and a parameter count of 9.8×10^6, demonstrating significant advantages in balancing precision and efficiency.

Keywords: *depth estimation*; *multi-scale feature fusion*; *feature enhancement*; *lightweight network*; *classification strategy*

在计算机视觉领域，单图像深度估计(single-image depth estimation, SIDE)一直是一个备受关注的问题。这项技术对于多个下游任务至关重要，例如3D重建、机器人导航、物体识别以及增强现实等。通过融合深度信息，目标识别系统可以更准确地分割和识别目标对象[1]，从而更好地理解目标的三维结构和姿态。这不仅提升了系统对环境的理解能力，还为多种应用场景提供了技术支持，如自动驾驶、虚拟现实和智能机器人等领域。传统的深度估计方法主要分为两大类：度量深度估计(metric depth estimation，MDE)和相对深度估计(relative depth estimation，RDE)。度量深度估计致力于预测以实际物理单位(如m)表示的深度，而相对深度估计则侧重于预测深度的相对变化，不涉及具体的度量尺度。

尽管度量深度估计在实际应用中具有明显优势，但其模型往往过度拟合特定数据集，难以泛化到其他数据集。相对而言，相对深度估计由于去除了尺度因素，能够更好地泛化到不同的环境和数据集中，但其预测的深度缺乏度量意义，限制了其在需要具体深度信息时的使用。此外，一些基于度量深度估计的工作[2-4]专注于如何提高深度估计精度，导致这些工作中的模型设计得很复杂，模型参数量巨大，训练这些网络模型往往需要较好的实验配置。然而，对于计算资源有限的设备来说，深度估计方法必须同时具备计算效率高和估计精度准确这2个特点。之前的一些工作已经尝试采用轻量级网络来提高计算效率，但是牺牲了深度估计精度。

为了同时兼顾深度估计模型推理速度快、推理精度高这2个优点，本文提出了一种创新的两阶段框架——LacDepth，通过轻量级的两阶段框架将相对深度和度量深度结合，在保持度量尺度的同时，提高模型的性能和参数的大小。在该方法中，第一个阶段，设计了深度残差金字塔模块(pyramid of depth residuals，PDR)，对编码器输出的多尺度特征进行双层空洞残差融合，在解码器高层特征中融入编码器底层特征信息，并引入拉普拉斯残差作为高频特征来增强轮廓信息，即解决解码器高层特征位置信息缺失问题，提高了不同尺度特征信息的语义信息和位置信息，又解决了参数大小问题。在第二个阶段，为了获取丰富的物体边界信息，本文设计了一个轻量级的分类器，该分类器根据吸引子(attractor)来微调深度分类中心，从而得到度量估计结果。这种方法能够更加灵活地调整深度估计结果，特别是在物体边界处，显著提升了深度图的质量。这种基于吸引子的分类策略不仅提高了模型的鲁棒性，还降低了模型的复杂度，使其更适合在计算资源受限的环境中部署。经大量实验证明，本文方法在KITTI数据集[5]的指标Abs Rel上达到了0.059，展示出其先进性。另外，模型规模仅9.8×10^6，具备明显的轻量化优势。

1 相关工作

1.1 单目深度估计(SIDE)

有监督的单图像深度估计方法主要分为两大类：度量深度[2-4]和相对深度[6-8]。度量深度模型通常在特定的数据集上进行训练，因此，容易出现过拟合现象，并且在面对未曾见过的环境或不同深度范围时，其泛化能力相对较弱。相比之下，相对深度模型由于能够在更加多样化的数据集上训练，并且采用相对深度标注(包括尺度不变损失函数)，因而具有更好的泛化性能。然而，相对深度模型在处理需要精确度量深度的下游任务时存

在局限性，即相对深度模型主要是对未知尺度和偏移的深度进行回归。

近期的研究致力于结合度量信息来进行深度估计。Yin 等[9] 提出了一种两阶段框架，通过将单目深度估计与 3D 点云编码器相结合，从而能够从单一图像中恢复出 3D 场景的形状。这些编码器经过训练后能够预测缺失的深度偏移和焦距。另外，Jun 等[10] 将度量深度分解为归一化深度和尺度特征，并设计了一个多解码器网络架构，其中的度量深度解码器能够利用来自梯度和归一化深度解码器的相对深度特征。

此外，鲁棒视觉挑战赛（robust vision challenge）也在探索通用深度预测技术，涵盖了室内和室外场景。一种流行的方法是将目标深度区间离散化，并将连续深度回归问题转化为分类任务。

1.2　度量深度估计

许多传统的基于学习的单目深度估计方法采用了带有卷积层的编码器–解码器架构，近年来还引入了 Transformers 模块。深度估计通常被视为逐像素回归任务。然而，一个不断发展的研究方向尝试将深度估计重新表述为组合的分类回归问题，以此解释图像上深度值的分布。AdaBins 扩展了标准的编码器–解码器主干网，使用基于 Transformer 的模块，该模块将预测的深度范围离散到深度区间中[11]。其中，深度区间宽度自适应地确定每个图像，最终的深度估计被计算为深度区间中心的线性组合。

LocalBins 基于 AdaBins 的概念，但在局部邻域内考虑给定像素的深度分布，而不是整个图像，并且以多尺度方式跨解码器层计算深度区间嵌入向量[12]。PixelBins 通过卷积层替代变压器块简化了 AdaBins 的架构，从而降低了计算复杂性[13]。BinsFormer 结合一个辅助的场景分类查询来指导深度区间的生成，并且同样利用了多尺度策略来细化自适应生成的深度区间[14]。PixelFormer 将深度估计视为通过跳过注意力机制精炼的像素查询，用于在不利用解码特征的情况下来预测深度区间中心[13]。

1.3　轻量级深度估计方法

为了能实时地估计深度信息，Wofk 等[15] 设计了一个轻量级的网络。使用深度可分离卷积构建整个网络，并进行网络剪枝，以减少计算量。文献 [16-17] 基于知识蒸馏方法，从参数量大的模型中蒸馏出一个参数量较小的模型。以往的轻量级模型都是牺牲深度估计精度以提升推理速度，因此，如何权衡估计精度和提升推理速度是一个重要的研究方向。Sheng 等[17] 设计了一个轻量的金字塔场景 Transformers 模块（pyramid scene transformer，PST）来捕捉多尺度的区域间交互。通过感知每 2 个特征之间深度特征的差异，再预测一个合理的场景结构，在较低参数量下获得较高性能和推理速度。

2　整体模型框架

2.1　融合相对与绝对估计的轻量级深度预测方法

在现有文献和实际应用中，轻量级网络通常指参数量、计算量和存储需求都较低、便于部署在移动或嵌入式设备上的网络架构上。例如，MobileNet 系列网络的参数量通常在 3×10^6～10×10^6 之间，被广泛认为是轻量级网络的典型代表，满足嵌入式实时任务的要求[18]。同时，在实时深度估计等任务中，基于网络的实时性和低功耗前提，提升估计或分类精度。因此，参数量为 9.8×10^6 的网络，符合实际应用场景对轻量级网络的要求。

本文所提轻量级深度估计模型输入 RGB 图像 $\boldsymbol{R}^{3\times H\times W}$（$H$ 表示长度，W 表示宽度），输出预测深度图像 $\boldsymbol{D}\in\boldsymbol{R}^{1\times H\times W}$。该深度估计网络由一个编码–解码架构及一个深度分类模块构成，如图 1 所示。图中，C 表示通道数。编码器基于预训练的 ResNet101 模型[19] 初始化，以提高收敛性。将 RGB 图像输入到 ResNet101[19] 预训练模型中，输入的彩色图像通过深度叠加的卷积块被缩为潜在特征。输入的 RGB 图像通过一系列卷积块被转化为潜在特征图，在 ResNet101 的第 2、4、8、16、32 层 Bottleneck 模块[19] 提取到的特征图为：f_1、f_2、f_3、f_4、f_5。其中，最高特征 f_5 具有丰富的语义信息，底层 f_1 具有丰富的颜色和位置信息。这些特征紧凑地包含了嵌入空间中颜色和深度值之间的关系，这是从不同的场景几何中学习到的。为了方便高层和底层特征信息的融合，首先，分别对 f_5、f_4、f_3、f_2、f_1 使用 1×1 卷积核将特征的通道增大一倍，分别记为 F_5、F_4、F_3、F_2、F_1。

为了克服传统卷积运算带来的感受野不足的限制，本节在编码器的卷积块中引入了两种不同

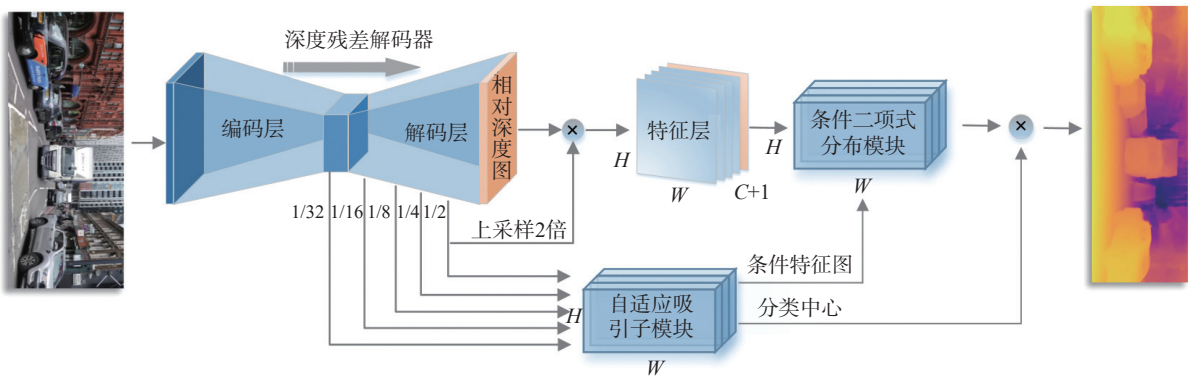


图 1 轻量级深度估计模型结构图

Fig.1 Model structure of lightweight depth estimation

膨胀率的ASPP卷积[20]，分别为1/8和1/16，从而能够在保持空间分辨率的同时捕捉更广泛的上下文信息。解码器部分如图2所示，具体来说，最底层具有丰富语义信息的F_5首先通过线性插值进行2倍上采样，与次高层特征F_4进行通道拼接后（用C表示），经空洞卷积模块扩展感受野，形成融合特征Z_4。该特征随后输入上采样金字塔，通过多级残差连接逐步恢复至原始分辨率。上采样金字塔如图3所示。在相对深度估计模块中，5个多尺度特征分别进行处理：底层特征经空洞卷积实现跨尺度特征融合，高层特征通过常规卷积提炼抽象信息，最终在上采样金字塔中迭代生成多尺度残差特征，并输出对应尺度的深度图预测。相对深度图像$\boldsymbol{R}\in\boldsymbol{R}^{1\times H\times W}$。

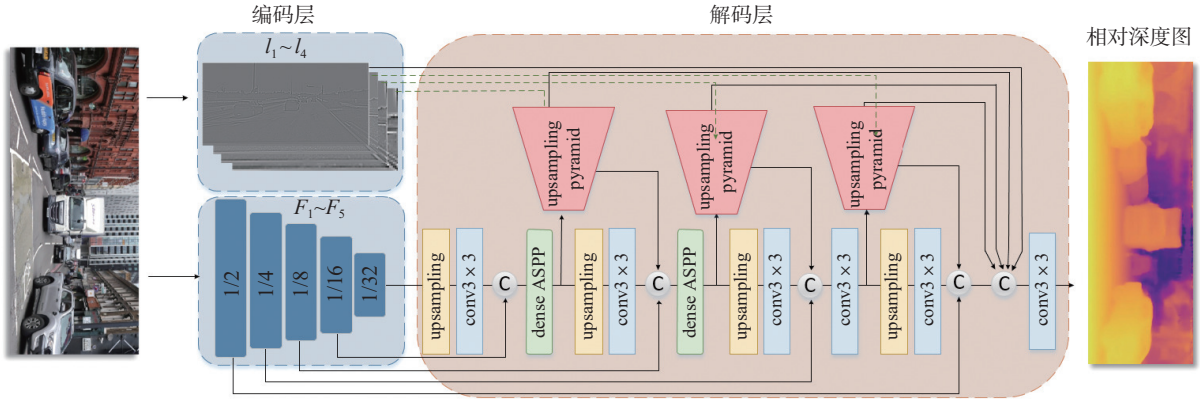


图 2 相对深度模块

Fig.2 Relative depth module

为解决深度回归任务中常见的边缘模糊问题，因此，本文设计了一个分类引导的深度离散化机制，模块细节如图4所示。该模块接受多尺度特征图作为输入，预测像素深度中心区间，并根据不同层级的吸引子数量在深度间隔上动态调整深度区间，实现多尺度细化。获得的深度分类中心及其对应的条件特征图进一步用于后续处理。通过将相对深度图与第4层的解码器输出进行通道拼接，形成增强特征F'，并与条件特征图通过条件概率分布模块计算像素级深度区间可能性，每个像素的深度条件概率p，再与分类中心进行加权求和，最终得到绝对深度图像$\boldsymbol{D}\in\boldsymbol{R}^{1\times H\times W}$。

$$d(i)=\sum_{k=1}^{Z}p_i(k)c_i(k) \tag{1}$$

式中：$d(i)$表示像素i对应的深度值；Z表示分类区间的总数；$p_i(k)$为该像素深度值属于第k个区间的条件概率；$c_i(k)$为离散化深度区间中心。

2.2 第一阶段：相对估计模块

2.2.1 编码器

在样本数据较少的情况下，为了提高深度学习模型的准确率，通常会考虑增加模型的深度、宽度及输入分辨率。然而，传统的卷积神经网络（convolutional neural network，CNN）虽然允许调整这些维度，但往往需要手动调节参数，并且可能

伴随着模型性能和效率的下降。相比之下，ResNet101 通过加深网络结构来提取图像中的底层纹理特征和高层语义信息。此外，借助迁移学习技术，该模型能够显著加速训练过程，并增强其在小规模数据集上的泛化能力。预训练模型提供的高质量初始权重不仅有助于模型快速收敛，还有效减少了过拟合的风险，从而确保在资源受限条件下的高效学习。

2.2.2　深度残差解码器

低分辨率的高层特征虽然富含语义信息，但缺乏精确的位置信息；而高分辨率的底层特征则相反。语义信息决定了图像的内容，而位置信息则提供了图像中的细节信息，例如物体的位置和纹理。在深度估计任务中，通常采用连续的卷积运算，这往往会导致语义信息和位置信息的丢失。此外，解码器在上采样阶段也会导致特征逐渐退化，从而使预测的深度图变得模糊，缺乏细节。

为了解决深度估计的精度、场景边界信息的精度和推理速度等问题，本文设计了一种深度残差解码器，如图 2 所示。为了克服深度估计中细节信息丢失的问题，引入了拉普拉斯残差信息。这一方法通过计算图像的多分辨率表示来捕捉并保留图像中的高频细节，有效补充了在下采样和上采样过程中丢失的高频细节，提升了深度图的边缘清晰度和表面纹理信息，使预测的深度图更加真实和细腻。计算拉普拉斯残差 l_j，如下所示：

$$l_j = I_j - \mathrm{UP}(I_{j+1}),\quad j = 1,2,3,4 \tag{2}$$

式中：j为拉普拉斯金字塔的水平指数；I_j为原始输入图像根据金字塔尺度进行下采样；UP(·)表示为上采样。采样过程采用双线性插值。

相对估计模块具体实现流程，首先，将待融合的两个特征F'、F_{i-1}进行通道拼接，并通过空洞卷积来获得更强的全局视野，以及语义信息和位置信息，通过上采样金字塔来恢复该分辨率下的深度图。如图 3 所示，在上采样金字塔中，执行 1×1卷积以减少通道，同时保留大部分上下文信息。然后，将局部特征x_k通过上采样得到深度残差金字塔中不同尺度下的深度图$\boldsymbol{D}'_k$以及下一尺度的特征x_{k+1}。将上一尺度的特征x_k尺度放大 2 倍，并用3×3卷积对融合后的特征图进行进一步的细化和降噪，与特征F'_{i+1}以及x_{k+1}进行通道拼接，将拼接后的特征送入一组堆叠的卷积块中，进行多尺度下的特征融合。最后，生成深度图$\boldsymbol{D}$。这一过程可描述为

$$\mathrm{conv}_{3\times3}(\mathrm{concat}\{x_k^{'}, x_{k+1}, F_{k+1}\}) \tag{3}$$

最后，将生成的$\boldsymbol{D}'_k$使用密集连接，将深度图拼接到一起，通过卷积，得到最终的深度图$\tilde{\boldsymbol{D}}_k$。

$$\tilde{\boldsymbol{D}}_k = \mathrm{conv}[\mathrm{concat}(F'_i, F'_{i-1}, \ldots)] \tag{4}$$

式中，concat( )代表通道拼接。

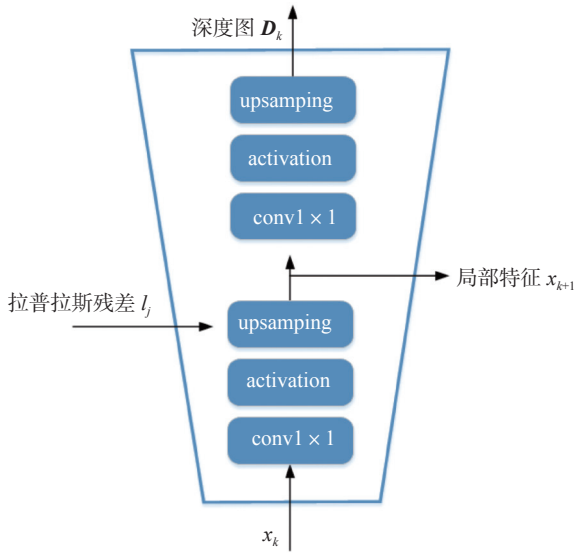


图 3　上采样金字塔模块

Fig.3　Upsampling pyramid module

2.3　第二阶段：量估计模块

2.3.1　自适应吸引子深度中心模块

为了解决传统回归问题在深度估计中导致的高不确定性、过拟合风险，以及对连续数值细微变化敏感等问题，Bhat 等[11] 设计了 LocalBins，将预测的深度范围划分为多个深度区间。此外，深度区间的上下界会根据输入图像自适应估计，从而为每张图像动态确定合适的深度范围。这样的设计虽然能提升深度估计场景边界的准确率，但是，适应全局的深度分布会导致模型冗余，大大增加模型参数量，不利于模型训练和在低算力平台的部署。

为了解决这些问题，本文基于 LocalBins 设计了一个轻量级深度分类模块，如图 4 所示。与 LocalBins 通过多尺度特征对深度区间进行分割以实现多尺度细化不同，本文通过引入“吸引子”来调整深度区间的位置，在深度间隔上向左或向右移动它们，从而实现多尺度细化。吸引子可以自由调整，没有局部约束（只有总宽度保持不变）。

本文设置深度区间总数 Z 为 64，吸引子张量为{16，8，4，1}，具体设置原因由消融实验说明，具体流程如图 3 所示。首先，将具有高维信息的 f_5 通过连续的卷积和激活函数，获得该特征图中深度分类区间向量 $\boldsymbol{B}_{\mathrm{c}}$ 和嵌入特征向量 $\boldsymbol{B}_{\mathrm{e}}$，$\boldsymbol{B}_{\mathrm{c}}$ 作为初始状态的深度区间分类中心位置的张量，$\boldsymbol{B}_{\mathrm{e}}$ 作为先验信息作用于上一层特征图。在第二个解码层，将上一层的 $\boldsymbol{B}_{\mathrm{e}}$ 上采样为当前特征图的尺寸大小，并与当前特征图作残差连接，得到新的特征图 F_k，这有助于在保持梯度传播的同时保留上一层的信息。F_k 通过一个多层感知机(multilayer perceptron, MLP) 后，再通过一层卷积将特征转成具有 4 个通道的特征图，即特征吸引子张量 $\boldsymbol{a}_k$ $(k=1,\cdots,4)$。计算吸引子张量 $\boldsymbol{a}_k$ 与 $\boldsymbol{B}_{\mathrm{c}}$ 的距离 $\Delta \boldsymbol{c}_i$，通过 $\Delta \boldsymbol{c}_i$ 来调整下一层的深度区间中心向量 $\boldsymbol{c}_i'$，即 $\boldsymbol{c}_i' = \boldsymbol{c}_i + \Delta \boldsymbol{c}_i$。$\Delta \boldsymbol{c}_i$ 的计算公式为

$$\Delta \boldsymbol{c}_i = \sum_{k=1}^{4} (\boldsymbol{a}_k - \boldsymbol{c}_i) \mathrm{e}^{-\alpha |\boldsymbol{a}_k - \boldsymbol{c}_i|^{\gamma}} \tag{5}$$

式中：$\boldsymbol{a}_k$ 表示特征吸引子张量；$\boldsymbol{c}_i$ 表示当前层级下深度区间的中心向量；超参数 α 和 γ 决定了吸引子的强度。

随着吸引子的不断减小，最终该模块输出了不断细化后的 $\boldsymbol{B}_{\mathrm{c}}$ 和嵌入特征向量 $\boldsymbol{B}_{\mathrm{e}}$。该嵌入特征向量作为条件特征传递到条件二项分布模块中。

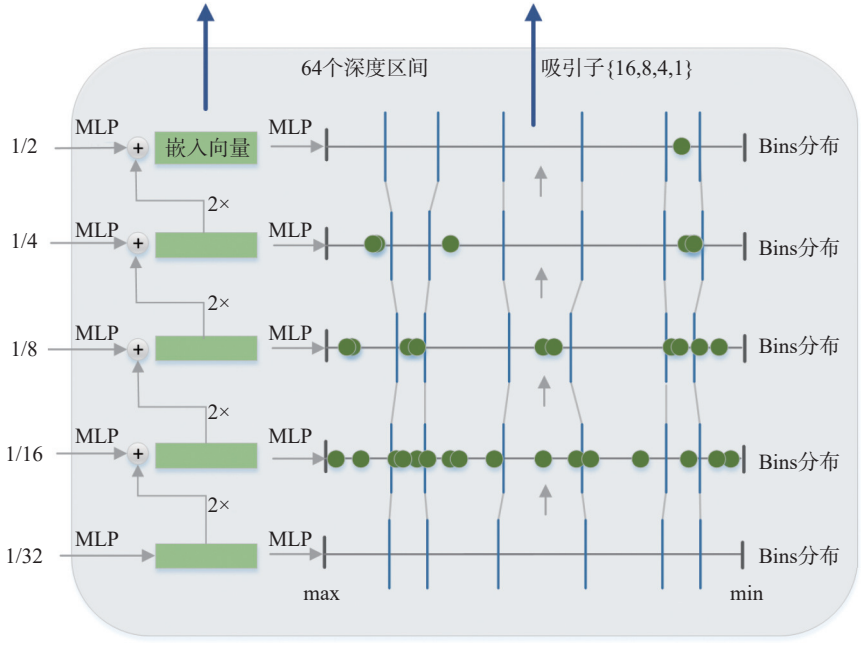


图 4　自适应吸引子深度中心模块

Fig.4　Deep center module of adaptive attractors

2.3.2　条件二项分布模块

为了获得最终的度量深度预测，本文将深度区间中心进行线性组合，并根据式 (1) 对其概率得分进行加权。先前基于自适应深度区间的模型[21] 使用 softmax 函数来预测深度区间中心的概率分布。softmax 被选用主要是受到离散分类问题的启发。尽管 softmax 在处理无序类别时表现良好，但在有序的深度区间，它可能会导致相邻深度区间中心之间的概率差异很大($|p_i - p_{i+1}| \gg 0$)。考虑到深度区间本质上是有序的，因此，可以直观地采用排序感知的概率预测。

受 Beckham 等[22] 的启发，使用条件对数二项分布来解决这个问题。模型首先将相对深度预测与解码器的特征图拼接在一起，形成一个新的特征表示。然后，通过这个新的特征表示预测出二项式分布的众数 q 和温度参数 t。最后，利用这两个参数来计算每个深度区间中心点的概率得分，即

$$p(k|q,t) = \frac{\exp\left(-(d_k - q)^2 / 2t^2\right)}{\sum_{k=1}^{Z} \exp\left(-(d_k - q)^2 / 2t^2\right)} \tag{6}$$

式中，d_k 表示第 k 个深度区间的中心深度值。

通过温度参数 t 调整后的高斯分布概率密度函数，计算给定 q 和 t 时每个深度区间中心点的概率

得分。在实践中，由于深度区间总数 Z 比较大，为避免数值下溢并简化计算，取 $\ln p$，并对阶乘使用斯特林近似，然后应用 $\mathrm{softmax}(\{(\ln p_k)/t\}_{k=1}^{Z})$来获得数值稳定性的归一化概率。这种方法不仅保留了对数的单峰性，还考虑了深度区间的有序性，同时保证了数值的稳定性。斯特林近似用于处理大数阶乘的计算问题，而 softmax 函数则确保输出的概率分布既反映各深度区间的相对可能性，又避免因直接计算指数而导致的数值溢出问题。这种组合方式有效地平衡了模型的准确性和计算效率。

2.4　损失函数

本文应用损失函数L_{t}对网络的可训练参数进行优化，该损失函数由数据损失L_{d}和梯度损失L_{g}两部分组成，公式如下:

$$L_{\mathrm{t}}=\begin{cases} uL_{\mathrm{d}}(y,y'), & \text{迭代次数小于 30} \\ uL_{\mathrm{d}}(y,y')+vL_{\mathrm{g}}(y,y'), & \text{其他} \end{cases} \tag{7}$$

式中：y和y'分别表示预测的深度映射和地面真实值；u和v分别表示L_{d}和L_{g}的平衡因子，分别设置为 10 和 0.1。梯度损失是在 30 个循环之后进行计算的，在训练开始时，由于地面真值的稀疏性，在同时使用数据和梯度损失时，KITTI 数据集中的训练往往不稳定。具体来说，由插值深度图计算的深度梯度与原始深度图略有不同，干扰了数据损失的收敛。为了缓解这一问题，只使用数据损失进行适当的深度图恢复后，再额外计算梯度损失。

数据损失：一般情况下，由于 3D 传感器的限制，近距离采集的深度数据比较密集，而距离较远的深度数据比较稀疏。为了缓解不平衡问题，引入损失函数的平方根作为数据损失L_{d}，该方法计算预测深度值与对数空间中真实值之间的差异，即

$$L_{\mathrm{d}}(y,y^*)=\sqrt{\frac{1}{n}\sum_{i\in i_V}h_i^2-\frac{\lambda}{n^2}\left(\sum_{i\in i_V}h_i^2\right)^2} \tag{8}$$

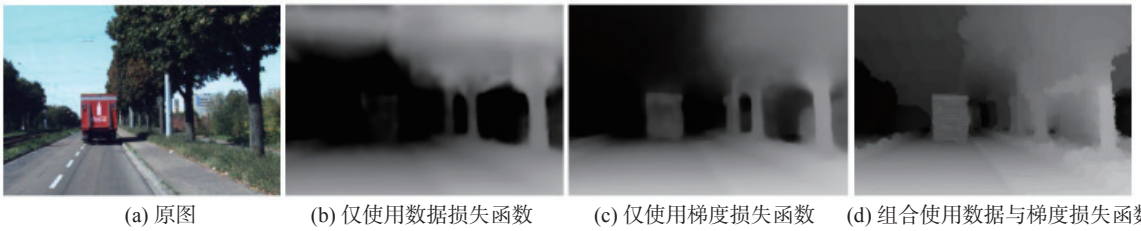
式中：y^*为预测真实值；$h_i=\ln y_i-\ln y_i^*$；V为深度图中的一组有效像素，i_V表示有效像素的总数；n 为样本数量；λ为平衡因子，设置为 0.85。

梯度损失：为了增强局部细节，特别是在深度边界，利用深度图的梯度作为损失函数。由于地面真值的稀疏深度数据，在水平方向和垂直方向上的梯度都难以精确计算，因此，使用了 Matlab 工具包进行点插值，该工具包在先前的方法中被广泛采用。梯度损失的公式如下：

$$L_{\mathrm{g}}(y,y^*)=\frac{1}{n}\sum_{i\in i_V}\left|y_{h,i}-m(y^*)_{h,i}\right|+\left|y_{iV,i}-m(y^*)_{iV,i}\right| \tag{9}$$

式中：$m(\cdot)$为文献 [23] 中使用的插值函数；$y_{h,i}$和$m(y^*)_{h,i}$分别表示估计深度图的第i个梯度值和水平方向上插值的地面真值；$y_{iV,i}$和$m(y^*)_{iV,i}$分别表示垂直方向上的对应梯度分量。

值得注意的是，梯度损失会迫使局部边缘在多层次金字塔精确对齐，从而使深度边界清楚地显示在最终的深度图中。梯度损失的影响如图 5 所示。可以看出，利用梯度损失预测的深度图成功地展现了远处物体的深度边界。



(a) 原图　(b) 仅使用数据损失函数　(c) 仅使用梯度损失函数　(d) 组合使用数据与梯度损失函数

图 5　不同损失函数组合的深度估计结果视觉比较

Fig.5　Visual comparison of depth estimation results according to different combinations of loss functions

3　实　验

3.1　数据集和预处理

本文采用 KITTI 和 NYU Depth V2 数据集进行实验。KITTI 数据集由德国卡尔斯鲁厄理工学院 (Karlsruhe Institute of Technology) 和丰田工业大学 (Toyota Technological Institute) 联合创建，被广泛应用于单目深度估计任务中。该数据集的图像分辨率为 1 242×375。本文对图像缩放并随机裁剪为 704×352 的图像。测试集包含从 29 个场景中选取的 697 张图像，而训练集则由剩余 32 个场景中的 23 488 张图像组成。

NYU Depth V2 数据集由纽约大学的研究人员发布。该数据集的图像分辨率为 640×480，包含来

自不同室内环境的464个场景，提供了1449对对齐的RGB-***D***图像，即彩色图像和对应的深度图像。

为了避免过拟合问题，在训练前采用以下方法对RGB图像和相应的深度图进行预处理：

a. 翻转，单目彩色图像和对应深度图像都以50%的概率水平翻转。

b. HSV增强，随机调整单目彩色图像和对应深度图像的亮度、对比度和饱和度值，比例因子选择在[0.9,1.1]范围内。

3.2 训练细节

本次实验训练网络为25个轮次。编码器模块基于ILSVRC[24]数据集，通过预训练模型初始化特征参数，非编码器部分的初始化是随机的。该模型使用Adam优化器进行训练，功率和动量分别设置为0.9和0.999。编码器的权重衰减因子设为0.0005，而对于所提出的解码器，则不使用权重衰减(即权重衰减因子为0)。初始学习率首先设置为0.0001，并且每5个轮次依次递减10%，直至达到10^{-5}为止。实验过程中，权重衰减统一设为0.0001。实验平台系统为Ubuntu18.04.5，在Pytorch中实现模型框架，所用硬件环境是一张NVIDIA 3090 16G显卡，CUDA版本为11.1。

3.3 实验结果与分析

3.3.1 定量分析

为了进行定量评价，本文使用Eigen等[25]引入的4个指标，这些指标在单目深度估计的性能评价中应用最为广泛，具体定义见文献[25]。

前3个度量标准即是在阈值$\delta\in[1.25^1,1.25^2,1.25^3]$预测的深度图像素精度，参数值越高代表性能越好。后3个指标分别是平均相对误差(REL)、均方误差(RMS)、平均log误差(log10)，参数值越小代表性能越好。

基于这些指标，本文在KITTI[5]数据集上与最新的轻量级和采用分类策略的方法进行了比较，相应的结果如表1、2所示。从表中可以看出，在所有量级网络中，Depth Anything等在绝对相对误差(Abs Rel)、均方根误差(RMSE)上的指标性能最佳，但其最小网络模型参数量高达24.8×10^6。模型依赖于大量额外的数据和一个强大的预训练编码器作为辅助训练，这样导致其模型参数量巨大，无法在嵌入式硬件中实时运行。而本文的网络模型仅是Depth Anything的12%，极大地减少了模型的参数量，并且保证了深度估计精度。在轻量级网络中，与文献[11]中的方法相比，本文方法在6个指标上都是领先的，平衡了推理速度和深度估计的精确度。这充分说明，本文设计的深度残差解码器能有效地筛选出高层特征的语义信息和底层特征的位置信息。另外，轻量级深度中心吸引子模块，通过将连续的深度值离散化为有限的类别，可以减少模型需要处理的数据量，从而提高计算效率，还可以有效提升深度图预测在边缘和纹理丰富区域的准确度。

此外，本文在NYU Depth V2数据集上与目前现有的模型进行比较，相应的结果如表3所示。Depth Anything在Abs Rel(0.056)和RMSE(0.206)2个核心指标上表现最优，这可能是由于其大规模预训练数据带来的强大泛化能力。而本文提出的模型在相对误差指标上展现出显著优势，其中，$\delta<1.25$达到0.953，逼近Depth Anything的0.984，且在更高精度的$\delta<1.25^3$指标上同样达到0.999的优异水平。这种表现差异可能与模型设计定位相关，Depth Anything作为大数据预训练模型，在数值精度上更具优势，而本文模型可能在结构设计上更注重对深度关系的学习，从而在相对准确率指标上表现突出。

在轻量级模型对比中，本文模型的Abs Rel(0.0787)显著优于同为轻量架构的Lite-Mono模型的Abs Rel(0.101)，且在δ指标上全面领先。这表明本文方法在模型效率与精度间取得了更好的平衡。关于室内场景的深度估计挑战，由于室内环境存在大量遮挡、反光表面和复杂的几何结构，深度值的绝对精度确实较难把握。未来可通过引入场景几何约束、融合多模态信息等方式，进一步提升模型对室内空间结构的理解能力。

3.3.2 定性分析

不同的场景光照强度不同，景深范围不同，给模型的推理带来困难。特别是城市街道上复杂的背景和物体遮挡也使得精确估计深度变得更加困难。尤其是在房子和车辆众多的环境中，模型往往难以准确还原场景内物体的细节。这些多样性要求模型具备更强的鲁棒性和适应性。图6是本文方法和文献中的预测结果，最后一列是本文预测结果。从图6的视觉效果来看，本文方法在应对城市街道上复杂的背景和深度信息突变的场景时具有较强的预测能力，并且在各种道路条件下，车子和道路轮廓等细节的保留更为丰富。相

表 1　非轻量级方法在 KITTI 数据集上的定量结果比较

Tab.1　Quantitative comparison of non-lightweight methods on the KITTI dataset

方法	参数量	Abs Rel	Sq Rel	RMSE	RMSELog	δ<1.25	δ<1.25^2	δ<1.25^3
Geonet[26]	31.6×10^6	0.149	1.06	5.567	0.226	0.796	0.935	0.975
Monoformer[27]	23.9×10^6	0.108	0.806	4.594	0.184	0.806	0.963	0.983
LocalBins[11]	36.0×10^6	0.059	0.269	2.405	0.075	0.97	0.96	0.99
Depth Anything[28]	24.8×10^6	**0.047**	**1.928**	**0.125**	**0.071**	**0.981**	**0.998**	**1.000**

表 2　轻量级方法在 KITTI 数据集上的定量结果比较

Tab.2　Quantitative comparison of lightweight methods on the KITTI dataset

方法	参数量	Abs Rel	Sq Rel	RMSE	RMSELog	δ<1.25	δ<1.25^2	δ<1.25^3
R-MSFM6[29]	3.8×10^6	0.126	0.944	4.981	0.204	0.857	0.952	0.978
Lite-Depth[30]	**3.1×10^6**	0.116	0.845	4.841	0.190	0.866	0.957	0.982
Lite-Mono[31]	8.7×10^6	0.101	0.729	4.454	0.098	0.897	0.965	0.983
PrimeDepth[32]	9.0×10^6	0.069	—	—	0.102	0.913	0.968	0.980
本文	9.8×10^6	**0.059**	**0.208**	**2.415**	**0.091**	**0.959**	**0.995**	**0.980**

表 3　不同的方法在 NYU Depth V2 数据集上的定量结果比较

Tab.3　Quantitative comparison of different methods on the NYU Depth V2 dataset

方法	Abs Rel	RMSE	RMSELog	δ<1.25	δ<1.25^2	δ<1.25^3
AdaBins[11]	0.103	0.364	0.044	0.903	0.984	0.997
Depth Anything[29]	**0.056**	**0.206**	**0.024**	**0.984**	**0.998**	**1.000**
Lite-Mono[31]	0.101	0.454	0.078	0.897	0.965	0.983
PrimeDepth[32]	0.058	—	0.030	0.966	0.998	0.999
本文	0.078 7	0.287	0.033	0.953	0.995	0.999

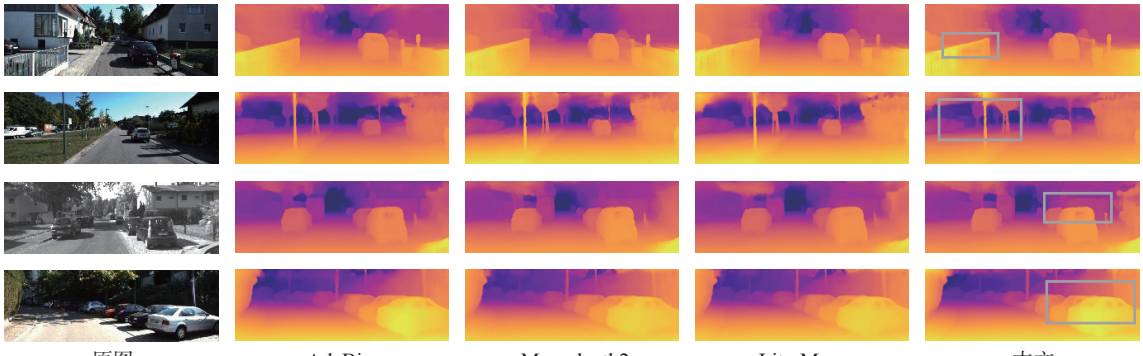


图 6　不同方法在 KITTI 数据集的预测深度图示例

Fig.6　Prediction depth map examples of different methods on the KITTI dataset

较于 Lite-Mono[32] 轻量级模型，本文方法获取到的场景边缘深度信息更为准确，与真实的深度信息视觉效果更为接近。相较于同为带有分类策略的模型，本文方法获得的场景信息更为平滑。综上所述，本文提出的轻量级深度估计方法可以估计出复杂的场景深度信息。

3.3.3　消融实验

为了验证各个模块的改进对深度估计的贡献，本文设计了消融实验。

Backbone 对比实验：如表 4 所示，从实验数据可以看出，替换 backbone 的方法有效提升了网络特征的提取能力。本文通过使用 4 个主流框架

（MobileNetV2[19]、VGG19[29]、InceptionV3[33]、ResNet-101[19]），同时保持其他设置不变，进行性能比较，相应的结果如表 4 所示。可以看出，MobileNetV2 的速度更快，但是 ResNet-101 相较于其他模型误差更小，精确度更高。这得益于 ResNet-101 采用了更深的网络结构（101 层）和创新的残差连接机制。这两种特性共同解决了深层网络训练中的梯度消失问题，提高了模型的学习能力和泛化能力，同时保持了较高的计算效率。实验结果从定量分析的角度证明了替换 backbone 方法的有效性，从而进一步减少了深度估计模型的深度预测误差。

解码器结构对比：在解码器结构的对比研究中，评估了不同金字塔层数对性能的影响。如表 5 中所示，通过增加深度残差金字塔层数能够精确地恢复局部细节以及深度图的全局性。为了在相同的实验条件下进行公平比较，鉴于 6 层和 7 层深度残差金字塔存在 GPU 存储器的限制，所有模型均采用较小的批量大小 (16→8) 进行训练。此方法不仅有效地解决了硬件资源限制的问题，还确保了实验结果的可比性和可靠性。如表 5 所示，层级增多导致性能略有改善。然而，随着金字塔层级增加到 4 以上，性能改善趋于平缓，接近上限。因此，可以认为在第 4 层深度残差金字塔处，已经能够充分恢复深度图的全局信息。

上采样金字塔对比实验：特征融合残差连接和密集连接是两种常用的特征融合方法。残差金字塔方法是将特征相加在一起，而密集金字塔方法是将特征连接在一起，将这 2 种方法与没有上采样金字塔的模型进行比较。表 6 中的结果表明，与无金字塔和特征融合残差连接方法相比，密集连接方法的性能最佳，仅需要增加少量参数。

深度中心吸引子模块：为了研究深度中心吸引子模块的各种类别对整体性能的贡献，首先，删除分类模块，并将卷积块附加到来自基础 ResNext-101 模型的解码器特征中，直接预测度量深度（标准回归）。本文称没有采用分类模块的为 Native；采用分类模块，分类模块的变体为 Metric，基于分裂的模块称为 splitter，深度中心吸引子模块称为 attract。如表 7 所示，设置吸引子数

表 4 KITTI 数据集上不同 backbone 的性能比较

Tab.4 Performance comparison of different backbones on the KITTI dataset

结构	参数量	速度/(帧·s^{-1})	Abs Rel	Sq Rel	RMSE	RMSELog	δ<1.25	δ<1.25^2	δ<1.25^3
MobileNetV2	**15.7**	**38.6**	0.072	0.284	2.76	0.109	0.941	0.99	0.998
VGG19	23.9	34.1	0.066	0.242	2.619	0.098	0.95	0.992	0.998
InceptionV3	17.7	36.7	0.063	0.232	2.563	0.095	0.955	0.993	0.999
ResNet-101	73.5	32.1	**0.059**	**0.212**	**2.446**	**0.091**	**0.964**	**0.994**	**0.999**

表 5 KITTI 数据集上不同深度残差模块层级的性能比较

Tab.5 Performance comparison of residual module levels at different depths on the KITTI dataset

模块结构	Abs Rel	Sq Rel	RMSE	RMSELog	δ<1.25	δ<1.25^2	δ<1.25^3
基准模块	0.080	0.315	3.174	0.123	0.921	0.972	0.992
第3层	0.065	0.239	2.695	0.107	0.950	0.989	0.997
第4层	0.059	0.208	2.415	0.090	0.964	0.994	0.999
第6层	0.058	0.206	2.396	0.089	0.964	0.995	0.999
第7层	0.058	0.205	2.350	0.086	0.964	0.995	0.999

表 6 KITTI 数据集上不同上采样金字塔结构的性能比较

Tab.6 Performance comparison of different upsampling pyramid structures on the KITTI dataset

结构	Abs Rel	RMSE	RMSELog	δ<1.25	δ<1.25^2	δ<1.25^3
No Pyramid	0.062	3.174	0.223	0.891	**0.962**	**0.992**
Residual	0.061	2.434	0.205	0.892	0.952	0.992
Dense	**0.060**	**2.431**	**0.203**	**0.895**	0.955	0.990

量为{16,8,4,1}的策略是最优的，对比分裂策略 REL 下降 3%。总体来说，吸引模块能够更好地将深度中心不断细化。

表 7　KITTI 数据集上不同分类策略的性能比较

Tab.7　Performance comparison of different classification strategies on the KITTI dataset

方法	类型	Config	REL	RMSE
Naive	—	—	0.062	2.65
Metric	splitter	factor=2	0.060	2.49
	attractor	{8,8,8,8}	0.061	2.51
		{16,2,2,16}	0.061	2.48
		{1,4,8,16}	0.061	2.44
		{16,8,4,1}	**0.059**	**2.41**

条件对数二项模块：通过研究本文模型在 KITTI 上使用条件对数二项分布和 softmax 概率的性能，评估了使用条件对数二项分布的效果。与文献 [17] 一致，本文观察到使用对数二项式 (Abs Rel 为 2.41) 代替 softmax(Abs Rel 为 2.44) 导致指数 Abs Rel 约 3% 的性能改善。这体现了单峰分布对于序数问题的重要性。

4　结　论

推理速度和估计精度是评估深度估计算法优劣的重要指标，而推理速度快和估计精度高两者往往不能兼得，本文提出一种面向两阶段的轻量级深度估计方法 LacDepth。LacDepth 通过第一阶段的深度残差金字塔机制有效融合多尺度特征，生成初步的相对估计结果；并在第二阶段利用基于吸引子策略的轻量级分类器，对相对估计结果进行度量估计的微调，显著提高了对物体边界估计的准确性和计算效率。实验结果表明，LacDepth 不仅在 KITTI 数据集上展现了强大的边缘信息预测能力，而且相比现有的轻量级深度估计方法，其平均相对误差更低（5.9%），参数量更少（9.8×10^6），具有优秀的推理速度和精度，体现了其在实际应用中的潜力和优势。

参考文献：

[1] 仇旭阳, 黄影平, 郭志阳, 等. 基于深度学习的障碍物检测与深度估计 [J]. 上海理工大学学报, 2020, 42(6): 558−565.

[2] 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.

[3] 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 [J/OL]. 计算机研究与发展. https://link.cnki.net/urlid/11.1777.tp.20250717.1700.010.

[4] 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 [J/OL]. 北京航空航天大学学报. https://doi.org/10.13700/j.bh.1001-5965.2024.0904.

[5] KONDA K R, MEMISEVIC R. Learning visual odometry with a convolutional network[C]//Proceedings of the 10th International Conference on Computer Vision Theory and Applications. Berlin: SciTePress, 2015: 486–490.

[6] 李坤乾, 王西亚, 宋大雷, 等. 基于语义信息约束的水下图像相对深度估计方法 [P]. 中国, CN113822919A. 2025-12-21.

[7] 曲熠, 陈莹. 基于尺度线索增强的无监督单目深度估计 [J]. 电子学报, 2024, 52(9): 3217−3227.

[8] 罗会兰, 周逸风. 深度学习单目深度估计研究进展 [J]. 中国图象图形学报, 2022, 27(2): 390−403.

[9] YIN W, ZHANG J M, WANG O, et al. Learning to recover 3D scene shape from a single image[C]//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 204–213.

[10] JUN J, LEE J H, LEE C, et al. Depth map decomposition for monocular depth estimation[C]//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022: 18–34.

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

[12] BHAT S F, ALHASHIM I, WONKA P. LocalBins: improving depth estimation by learning local distributions[C]//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022: 480–496.

[13] SARWARI K, LAINE F, TOMLIN C. Progress and proposals: a case study of monocular depth estimation[R]. Berkeley: University of California, Berkeley, 2021.

[14] LI Z Y, WANG X Y, LIU X M, et al. BinsFormer: revisiting adaptive bins for monocular depth estimation[J]. IEEE Transactions on Image Processing, 2024, 33: 3964−3976,

[15] WOFK D, MA F C, YANG T J, et al. FastDepth: fast monocular depth estimation on embedded systems[C]//Proceedings of 2019 International Conference on Robotics and Automation. Montreal: IEEE, 2019: 6101–6108.

[16] ALEOTTI F, ZACCARONI G, BARTOLOMEI L, et al. Real-time single image depth perception in the wild with

handheld devices[J]. Sensors, 2021, 21(1): 15.

[17] SHENG F, XUE F, CHANG Y C, et al. Monocular depth distribution alignment with low computation[C]// Proceedings of 2022 International Conference on Robotics and Automation. Philadelphia: IEEE, 2022: 6548–6555.

[18] SANDLER M, HOWARD A, ZHU M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4510–4520.

[19] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770–778.

[20] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[Z]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017,40(4): 834–848.

[21] AGARWAL A, ARORA C. Attention attention everywhere: monocular depth prediction with skip attention[C]//Proceedings of 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2023: 5850–5859.

[22] BECKHAM C, PAL C J. Unimodal probability distributions for deep ordinal classification[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney: PMLR, 2017: 411–419.

[23] SILBERMAN N, HOIEM D, KOHLI P, et al. Indoor segmentation and support inference from RGBD images[C]//Proceedings of the 12th European Conference on Computer Vision. Florence: Springer, 2012: 746–760.

[24] KAWANO Y, YANAI K. ILSVRC on a smartphone[J]. Information and Media Technologies, 2014, 9(3): 371−375.

[25] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014: 2366–2374.

[26] YIN Z C, SHI J P. GeoNet: unsupervised learning of dense depth, optical flow and camera pose[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1983–1992.

[27] ZHAO C Y, SONG Y X, WANG W H, et al. MonoFormer: one transformer for both diffusion and autoregression[EB/OL]. [2024-09-24]. https://doi.org/10.48550/arXiv.2409.16280.

[28] XU M Y, GUO Z Q, WANG A, et al. DAUA-Plane: depth anything with uncertainty map and attention-based plane reconstruction of surgical scenes[J]. Procedia Computer Science, 2024, 250: 67−73,

[29] ZHOU Z K, FAN X N, SHI P F, et al. R-MSFM: recurrent multi-scale feature modulation for monocular depth estimating[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 12757–12766.

[30] LI Z Y, CHEN Z H, XU J L, et al. LiteDepth: digging into fast and accurate depth estimation on mobile devices[C]//Proceedings of European Conference on Computer Vision. Tel Aviv: Springer, 2022: 507–523.

[31] ZHANG N, NEX F, VOSSELMAN G, et al. Lite-Mono: a lightweight CNN and transformer architecture for self-supervised monocular depth estimation[C]//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 18537–18546.

[32] ZAVADSKI D, KALŠAN D, ROTHER C. PrimeDepth: efficient monocular depth estimation with a stable diffusion preimage[C]//Proceedings of the 17th Asian Conference on Computer Vision. Hanoi: Springer, 2024: 21–40.

[33] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 2818–2826.

（编辑：丁红艺）