视听融合耦合坐标自注意的单目深度估计

马存良, 蒲江川, 许春冬, 易见兵, 嘉明珍*

(江西理工大学信息工程学院 赣州 341000)
(mingzhenj@jxust.edu.cn)

摘 要: 针对单目图片和声音回波信号都含空间信息这一特点, 提出一种视听融合的单目深度估计方法. 首先, 通过池化金字塔模块融合分析回波与材料特征来自适应估计单目图片的离散深度值; 然后, 采用卷积神经网络和 Transformer 相结合的方法对单目图片进行编码, 改进坐标注意力提出坐标自注意力模块对图片特征解码获得离散深度值的概率分布; 最后, 将像素点的深度值建模为离散深度值的期望来构建最终深度图. 实验结果表明, 在仿真数据集 Replica 和 Matterport3D 数据集上, 所提方法的均方根误差分别为 0.204 和 0.875, 相对误差分别为 0.095 和 0.161, 均取得具有竞争力的结果; 在真实数据和含噪声数据中, 该方法能够应用于真实场景的深度估计.

关键词: 单目深度估计; 视听融合; 池化金字塔模块; 自注意力
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2023-00044

Monocular Depth Estimation Based on Audio-Visual Fusion Coupling Coordinate Self-Attention

Ma Cunliang, Pu Jiangchuan, Xu Chundong, Yi Jianbing, and Jia Mingzhen*

(*School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou* 341000)

Abstract: Aiming at the characteristic that both of monocular image and sound echo signal contain spatial information, this paper proposes a monocular depth estimation method based on audio-visual fusion. Firstly, fuse and analyze echo and material features by the pooling pyramid module to estimate adaptively the discrete depth value of the monocular image; secondly, use a method combining convolutional neural network and Transformer to encode monocular images, improve coordinate attention proposes a coordinate self-attention module to decode image features to obtain the probability distribution of discrete depth values; finally, model the depth value of a pixel as the expectation of a discrete depth value to construct the final depth map. Experimental results show that on the simulation data sets Replica and Matterport3D data sets, the root mean square errors of the proposed method are 0.204 and 0.875 respectively, and the relative errors are 0.095 and 0.161 respectively, both achieves competitive results; experimental results on real data and noisy data show that this method can work in monocular depth estimation in real scenes.

Key words: monocular depth estimation; audio-visual fusion; pyramid pooling module; self-attention

收稿日期: 2023-04-22; 修回日期: 2024-02-23. **基金项目**: 国家自然科学基金(11864016); 江西省自然科学基金(20224BAB211012); 江西省教育厅科技项目(GJJ210828). **马存良**(1987—), 男, 博士, 副教授, 硕士生导师, 主要研究方向为深度学习; **蒲江川**(1996—), 男, 硕士研究生, 主要研究方向为单目深度估计; **许春冬**(1976—), 男, 博士, 副教授, 硕士生导师, 主要研究方向为语音与音频信号处理; **易见兵**(1980—), 男, 博士, 副教授, 硕士生导师, 主要研究方向为计算机视觉等; **嘉明珍**(1989—), 女, 博士, 讲师, 硕士生导师, 论文通信作者, 主要研究方向为智能信号处理、深度学习.

单目深度估计是从单张图片预测场景深度信息的技术. 机器学习的发展使得在场景信息先验基础上像素级的单目深度估计成为可能, 促进了即时定位与建图[1]、自动驾驶[2]、虚拟现实[3]的发展. 与激光雷达和多目深度估计相比, 单目深度估计具有低成本的优势. 但由于没有多视图几何的约束, 单目深度估计的映射结果不唯一, 使其具有挑战性[4-5].

在提出基于深度学习的单目深度估计方法前, 已有许多经典的估计方法. 简单场景假设下, 可以将图片转换为三维场景[6]. Saxena 等[7]首次将机器学习应用于单目深度估计, 使用马尔可夫随机场方法构建深度图; 根据异距物体成像模糊效果不同这一相机成像原理, Tang 等[8]提出通过对图片分析获得深度线索信息, 进而得到深度图的方法.

早期的方法欠缺复杂场景建模能力; 近期基于深度学习的方法, 如 ResNet(residual network)[9], DenseNet(dense convolutional network)[10] 和 ViT (vision Transformer)[11]等提升获取图片深层次语义信息的能力, 促进了单目深度估计的发展. Eigen 等[4]应用卷积神经网络(convolutional neural network, CNN), 通过全局和局部信息融合获得了当时的 SOTA(state of the art)结果; Laina 等[12]通过全卷积残差网络构建编码-解码结构进行图片的深度场景预测. CNN 能够很好地建模图片的局部关系, 但是缺乏提取全局信息的能力. 大卷积核[13]、空洞卷积[14]以及特征金字塔[15]等在一定程度上缓解了缺乏长程依赖的问题, 但由于 CNN 固有的空间局部偏置特性, 最低分辨率特征的感受野仍是有限的[16]. 上采样过程中, CNN 中卷积及池化操作的不可逆性将导致解码过程无法完全恢复图片的细节信息[17], 出现物体边缘深度估计不佳的问题. 与视觉信号相比, 声音信号的获取和处理成本更低, 并且声音回波中同样包含丰富的空间信息.

单目深度估计具有的不适定性导致端到端的深度回归变成一项困难的任务[5]. 与深度回归不同, Cao 等[18]将深度估计任务建模为像素级的分类, 然而深度值的离散化造成近邻像素间深度不连续, 导致视觉质量较差. 为了解决这些问题, 一些工作尝试将二者结合起来, 将单目深度估计视为回归-分类任务[5,19-20].

本文提出视觉听觉融合的单目图片深度估计网络 VisionEchoNet, 将单目深度预测视为一个回归-分类任务, 减少了离群值的影响, 使网络更易收敛; 针对实际场景中回波受反射体形状和材料共同影响(不同材料反射和吸收特性有异), 提出材料和回波二者融合的自适应离散深度值估计方法, 通过池化金字塔模块(pyramid polling module, PPM)对回波和材料特征融合分析实现归一化深度区间的估计, 进而计算离散深度值; 针对 CNN 在全局信息提取能力上的不足, 采用 ResNet 与 Transformer 相结合的方法估计像素点离散深度值的概率分布; 针对物体边缘深度估计不佳的问题, 在像素点的深度概率分布估计的解码模块, 采用改进的坐标注意力方法将坐标注意与 Transformer 结合, 提出坐标自注意模块. 在 Replica 和 Matterport3D 数据集上的实验结果表明, 本文方法是有效的.

1 相关工作

Dosovitskiy 等[11]提出一种基于自注意力机制的深度学习模型 ViT, 其利用大卷积核将图片分成一系列的图块, 并将图块送入 Transformer 模型. 由于 Transformer 具有全局感受野, 因此 ViT 模型能够学习到不同图块之间的关系, 很好地获得全局的信息. 受到 ViT 优秀表现的启发, 最近越来越多的工作研究 Transformer 的编码器[21-22]在计算机视觉任务中的应用, 然而由于 Transformer 缺少局部的归纳偏置, 导致最小分辨率的特征缺乏详细的位置信息[23].

为了提高神经网络特征的表达能力, Hu 等[24]提出通道注意力模块, 包含压缩和激励(squeeze excitation, SE)2 部分: 首先将特征图压缩成一维向量, 然后利用全连接层调整每个通道的学习权重. Woo 等[25]在通道注意力 SE 的基础上增加空间注意力机制, 以便关注到空间位置上的信息. Hou 等[26]引入坐标注意力(coordinate attention, CA)机制, 在获取到一个坐标方向上的远程依赖关系的同时, 能够保存沿另一个坐标方向上的位置信息.

声音回波中包含空间信息. Christensen 等[27]受到蝙蝠、海豚等利用回波进行空间感知的启发, 提出通过声音回波信息进行室内场景三维感知的深度学习方法, 然而基于回波的三维感知无法对纹理区域进行精确的预测. 在此基础上, Gao 等[28]提出利用视听结合的方式完成深度预测, 实验结果表明, 回波特征与视觉特征的融合能够提升深度估计效果. Antonacci 等[29]认为, 空间三维结构和材料共同影响回波波形; 在此基础上, Parida 等[30]提出 BITD(beyond image to depth), 通过在回波特征中融入材料特征, 并利用注意力机制赋予回波深

度图与视觉深度图不同权重，可使模型获得 SOTA 结果．然而，这样的预测结果仍然存在一定问题，而对像素级的深度估计这样的密集任务中，回波在局部细节表达能力差，尤其在一些室内复杂场景下预测出的深度图不准确．

近年来，许多工作将深度预测看作像素级的回归任务[31-35]，然而这样的方法存在训练成本高、容易陷入局部最优解等问题．与此不同，Cao 等[18]将深度预测建模为像素级的分类任务，将连续的深度范围离散成多个深度区间，通过深度网络模型将每个像素进行深度区间分类，该方法减少了训练的成本；但是，像素级分类导致生成的深度图不平滑和预测准确率下降等问题．为了解决上述问题，有些工作通过回归-分类结合的方式构建深度预测网络[5,19-20]，即将深度图转化为离散深度值和离散深度值的概率分布图的线性结合(将深度图的计算转化为深度值的期望求解问题)．与离散深度值的均匀采样和对数采样[19,36]相比，Bhat 等[5]认为自适应的深度值采样有更好的表现，即在每个变化的场景中离散深度值都应该自适应地被预测出来，该策略取得了优秀的表现．将离散深度值自适应地估计的方法也是更加契合深度值随场景而变化的特点[35]．然而以上分类方法都是利用单目图片进行预测，但视觉模态在这种无纹理的区域中，如墙壁、空旷的房间等，由于缺少场景的结构信息，存在对深度预测模糊的问题．

2　本文方法

与近些年常用的回归-分类结合的方法不同，本文提出的 VisionEchoNet 通过回波信号和材料网络 MaterialNet 自适应地估计场景的离散深度值．这种离散深度值的估计符合物理上的直观：对声音回波的一阶近似公式分析发现，物体 i 的深度可以直观地由 $vt_i/2$ 计算，其中，v 表示声波在空气中传播速度(近似为常量)，t_i 表示第 i 个物体反射到达麦克风的时间；同时，声音的反射受目标材料属性的影响．

图 1 所示为 VisionEchoNet 的框架，包括回波网络 EchoNet，MaterialNet，视觉网络 VisionNet，PPM 网络 PPMNet 等．最终，深度图由自适应离散深度值和像素点上离散深度值的概率分布线性结合生成．

图 1 中，EchoNet 获取回波信号的时频信息 $\boldsymbol{E}\in\mathbb{R}^{2\times F\times T}$，通过堆叠的卷积模块从场景提取回波特征；MaterialNet 为预训练好的图片的材料分割模型[30]，本文采用 ResNet18 作为基础网络，通过获取图片信息提取材料特征；PPMNet 为自适应归一化离散深度区间预测网络，通过深度区间计算离散深度值；VisionNet 通过编解码结构提取图片 $\boldsymbol{I}\in\mathbb{R}^{3\times H\times W}$ 的多尺度特征，针对 CNN 在单目深度估计方面全局特征提取欠缺的问题，在编码阶段采用 ResNet50 融合 TransformerEncoder 模块的方

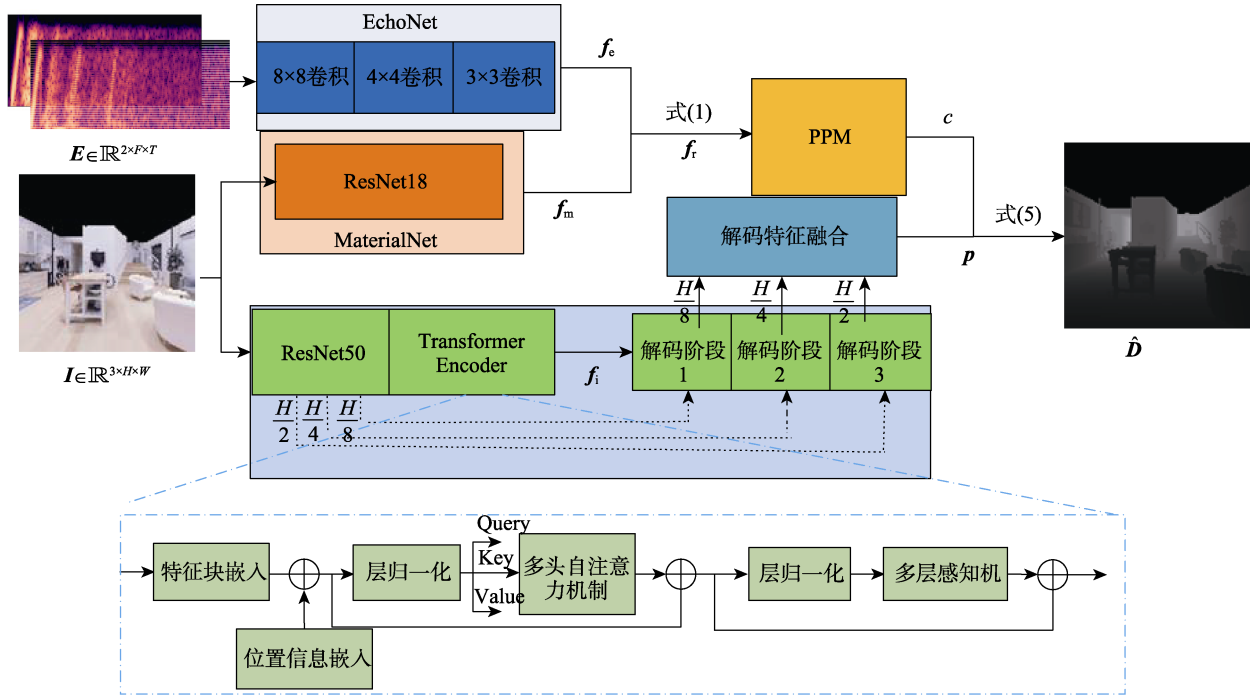


图 1　VisionEchoNet 的结构

案，通过 ResNet50 提取图片的局部信息，将其输入 TransformerEncoder 模块获得融合全局和局部信息的编码特征. 在解码阶段，采用本文改进的坐标自注意力机制(coordinate self-attention, CSA)改善信息损失的问题.

2.1 回波信号预处理

由于回波具有感知场景深度的能力，因此本文利用回波信息预测自适应离散深度值. 与 VisualEchoes[28,37]一致，回波信号 $\boldsymbol{y}(t)$ 通过物理仿真得到，在一阶近似下，有

$$\boldsymbol{y}(t)=\sum_{i=1}^{k}a_i\boldsymbol{x}(t-t_i).$$

其中，$\boldsymbol{x}(t)$ 表示扬声器音频信号；a_i 表示信号到达麦克风的振幅影响因子(受场景内材料和形状等信息影响[37])；t_i 表示返回到麦克风的音频时延，k 表示不同深度材料的数量.

本文将回波信号通过短时傅里叶变换(short-time Fourier transform, STFT)，得到回波频谱图 $\boldsymbol{E}=\mathrm{STFT}\left(\boldsymbol{y}(t)\right)$，将其作为 EchoNet 的输入. 其中，$\boldsymbol{E}\in\mathbb{R}^{2\times F\times T}$；$T$ 表示时间步数；F 表示频点数量.

2.2 通过 EchoNet 提取回波特征

将频谱图 $\boldsymbol{E}$ 输入 EchoNet 可以得到回波特征 $\boldsymbol{f}_{\mathrm{e}}=\mathrm{EchoNet}\left(\boldsymbol{E}\right)$. 其中，受 BITD[30]的 EchoNet 启发，EchoNet 由 3 层卷积网络堆叠构成，卷积核的个数分别为 32, 64 和 8，卷积核尺度分别为 8×8, 4×4 和 3×3，步长分别为 4, 2 和 1. 每层卷积运算后连接批归一化处理(batch normalization, BN)和 ReLU 激活函数. 最后使用 1×1 的卷积操作将回波特征向量通道数调整为 512，EchoNet 的输出为 $\boldsymbol{f}_{\mathrm{e}}\in\mathbb{R}^{512\times H/32\times W/32}$.

2.3 通过 MaterialNet 提取材料特征

MaterialNet 的输入为单目图片 $\boldsymbol{I}$，其输出场景的材料信息特征为 $\boldsymbol{f}_{\mathrm{m}}=\mathrm{MaterialNet}\left(\boldsymbol{I}\right)$. 其中，$\boldsymbol{f}_{\mathrm{m}}\in\mathbb{R}^{512\times H/32\times W/32}$.

采用材料分类数据集 MINC[38]对 MaterialNet 进行分割预训练，得到的权重作为 MaterialNet 的初始权重. 本文中，MaterialNet 结构和网络权重与 BITD 一致.

2.4 融合材料和回波特征预测自适应离散深度值

应用双线性变换将回波特征和材料特征融合，得到回波材料耦合特征为

$$\boldsymbol{f}_{\mathrm{r}}=B\left(\boldsymbol{f}_{\mathrm{e}},\boldsymbol{f}_{\mathrm{m}}\left|\boldsymbol{W}_{\mathrm{b}},\boldsymbol{b}_{\mathrm{b}}\right.\right)\tag{1}$$

其中，$\boldsymbol{f}_{\mathrm{r}}\in\mathbb{R}^{512\times H/32\times W/32}$；$\boldsymbol{W}_{\mathrm{b}}$ 和 $\boldsymbol{b}_{\mathrm{b}}$ 分别表示可学习的权重矩阵和偏置；B 表示双线性变换.

材料特征和回波特征的线性融合可能会导致对复杂场景的理解不足. 通过 PPMNet 进一步处理回波材料耦合特征 $\boldsymbol{f}_{\mathrm{r}}$，得到归一化深度区间 $\boldsymbol{b}=\mathrm{PPMNet}\left(\boldsymbol{f}_{\mathrm{r}}\right)$. 其中，$\boldsymbol{b}\in\mathbb{R}^{256\times1}$，并满足约束 $\sum_k\boldsymbol{b}_k=1$，该向量将产生可探测深度范围内的 256 个自适应离散深度值 $\boldsymbol{c}\in\mathbb{R}^{256\times1}$. 第 k 个深度区间的中心深度值计算公式为

$$\boldsymbol{c}_k=\boldsymbol{d}_{\min}+\left(\boldsymbol{d}_{\max}-\boldsymbol{d}_{\min}\right)\left(\frac{\boldsymbol{b}_k}{2}+\sum_{j=1}^{k-1}\boldsymbol{b}_j\right).$$

其中，$\boldsymbol{d}_{\min}$ 和 $\boldsymbol{d}_{\max}$ 分别表示可估计深度的最小值和最大值.

2.5 通过 VisionNet 提取多尺度视觉特征

本文构建了编-解码结构的 VisionNet，用于提取单目深度图片特征. 该网络输入为单目图片 $\boldsymbol{I}$，输出为多尺度的解码特征，公式为

$$\left[\boldsymbol{f}_{\mathrm{d1}},\boldsymbol{f}_{\mathrm{d2}},\boldsymbol{f}_{\mathrm{d3}}\right]=\mathrm{VisionNet}\left(\boldsymbol{I}\right).$$

其中，$\boldsymbol{f}_{\mathrm{d1}}\in\mathbb{R}^{512\times H/8\times W/8}$；$\boldsymbol{f}_{\mathrm{d2}}\in\mathbb{R}^{256\times H/4\times W/4}$；$\boldsymbol{f}_{\mathrm{d3}}\in\mathbb{R}^{128\times H/2\times W/2}$.

融合 ResNet50 和 Transformer 编码器构建编码模块. 鉴于在计算机视觉领域优秀的特征提取效果，ViT 成为许多编码器的骨干网络；然而，直接对切割的图片块进行整体的分析，可能会造成图片块之间边缘和细节等信息的丢失. SwinTransformer[39]的层次化构建方法与移动窗口自注意力，使得 Transformer 不仅具有内在的长程依赖性，而且对于局部信息也有一定的弥补，该结构成为流行的骨干网络. 然而在实验中发现，SwinTransformer 在 Replica 数据集上表现结果并不是很好，本文将在消融实验部分进一步探究特征提取网络. 本文采用 CNN 和 Transformer 结合的方法，可以兼顾 CNN 的局部特征提取和 Transformer 的全局特征提取能力. 为了降低计算量，对 ResNet50 进行改进. ResNet50 主要由 1×1, 3×3 和 1×1 的卷积堆叠而成，共堆叠 4 层，分别堆叠 3, 4, 6 和 3 次，输出特征维度分别为 256, 512, 1 024 和 2 048. 本文中，将最后 2 层进行合并，即堆叠次数改为 3, 4 和 9，输出特征维度分别为 256, 512 和 1 024.

为了保留原始的空间细节信息，在解码阶段首先对 Transformer 模块的输出特征进行线性插值上采样；然后使用跳跃连接将编码器的低层特征与解码器高层特征在维度上进行拼接结合；最后进行 2 次 3×3 的卷积操作，更好地捕捉融合后特征图的局部特征，并且用多个卷积操作在尽可能减少参数量的情况下增大感受野，每次卷积操作后

都连接 BN 和 ReLU 激活函数. 由于上采样过程不可避免地存在信息的丢失和模糊化，为了增强特征的表达能力，本文提出改进的 CSA 模块，在通道注意力上嵌入水平和竖直方向上的位置信息，并在 2 个方向上并行进行自注意力计算 2 个方向的每个不同像素之间的影响.

2.6　改进的 CSA 模块

针对现有 CA 模块存在的局部特征损失的问题，本文对其进行改进，提出 CSA 模块，其结构如图 2 所示. 在空间的 2 个方向中，坐标注意力应用全局平均池化获得水平方向和竖直方向特征，能够获取整个场景中的全局特征即场景中的背景信息. 然而对于场景中缺乏特征信息的小目标时，仅以全局信息代表整个空间场景则可能造成信息的损失，如目标物体的细节、纹理等；而全局最大池化更加侧重提取局部的特征信息，更易捕获具有显著特征的小目标. 因此本文提出 CSA 模块：在空间的 2 个不同方向，分别使用全局平均池化和全局最大池化增强信息的提取. 假设输入特征为 $\boldsymbol{x}\in\mathbb{R}^{C\times W\times H}$，输出特征 $\boldsymbol{y}\in\mathbb{R}^{C\times H\times W}$.

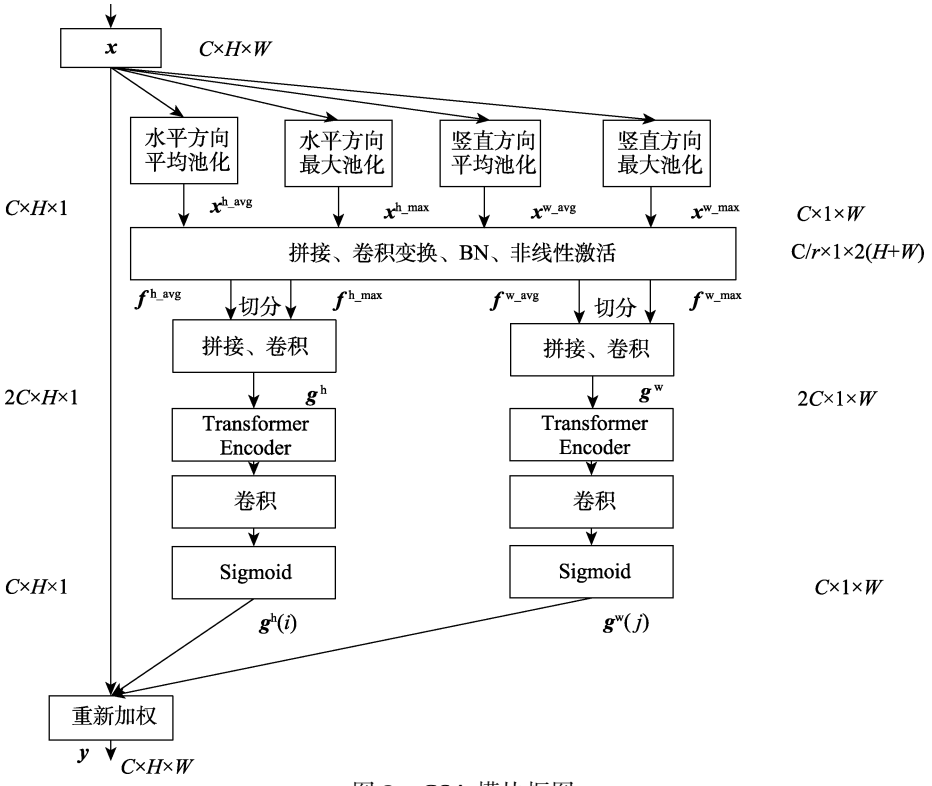


图 2　CSA 模块框图

CSA 模块的特征提取计算过程如下:

用尺寸为(H ,1)和(1, W)的池化核沿着水平方向和竖直方向对特征图分别进行平均池化($\boldsymbol{x}^{\text{h_avg}}$ 和 $\boldsymbol{x}^{\text{w_avg}}$)、最大池化($\boldsymbol{x}^{\text{h_max}}$ 和 $\boldsymbol{x}^{\text{w_max}}$), 对获得的 4 个特征向量进行拼接、激活并进行特征融合，得到特征 $\boldsymbol{f}$，公式为

$$\begin{cases}\boldsymbol{x}^{\text{h_avg}}=\dfrac{1}{W}\displaystyle\sum_{i=0}^{W}\boldsymbol{x}(h,i)\\ \boldsymbol{x}^{\text{h_max}}=\max\left(\boldsymbol{x}(h,i)\right)\\ \boldsymbol{x}^{\text{w_avg}}=\dfrac{1}{H}\displaystyle\sum_{j=0}^{H}\boldsymbol{x}(j,w)\\ \boldsymbol{x}^{\text{w_max}}=\max\left(\boldsymbol{x}(j,w)\right)\end{cases}\tag{2}$$

然后沿着空间维度，融合特征 $\boldsymbol{f}\in\mathbb{R}^{C/r\times 2(H+W)}$ 切分成 4 个张量 $\boldsymbol{f}^{\text{h_avg}}$，$\boldsymbol{f}^{\text{h_max}}\in\mathbb{R}^{C\times H\times 1}$ 和 $\boldsymbol{f}^{\text{w_avg}}$，$\boldsymbol{f}^{\text{w_max}}\in\mathbb{R}^{C\times 1\times W}$；将相同方向的特征在通道维度上拼接，并利用自注意力机制对在 2 个不同方向的特征进行融合分析，计算不同位置的相关性. 计算公式为

$$\begin{cases}\boldsymbol{g}^{\text{h}}=C\left(\left[\boldsymbol{f}^{\text{h_avg}},\boldsymbol{f}^{\text{h_max}}\right]\right)\\ \boldsymbol{g}^{\text{w}}=C\left(\left[\boldsymbol{f}^{\text{w_avg}},\boldsymbol{f}^{\text{w_max}}\right]\right)\end{cases}\tag{3}$$

$$\begin{cases}\boldsymbol{g}^{\text{h}}(i)=\sigma\left(C\left(T\left(\boldsymbol{g}^{\text{h}}\right)\right)\right)\\ \boldsymbol{g}^{\text{w}}(j)=\sigma\left(C\left(T\left(\boldsymbol{g}^{\text{w}}\right)\right)\right)\end{cases}\tag{4}$$

其中，C 和 T 分别表示卷积变换函数和 Trans-

formerEncoder 函数；σ 表示 Sigmoid 激活函数；$[\cdot,\cdot]$表示特征拼接.

最后通过输入和注意力权重计算得到空间 CSA 的输出张量 $\boldsymbol{y}$. 对于坐标(i,j)的输出，计算公式为 $\boldsymbol{y}(i,j)=\boldsymbol{x}(i,j)\times\boldsymbol{g}^{\mathrm{h}}(i)\times\boldsymbol{g}^{\mathrm{w}}(j)$. 其中，× 表示矩阵的 Hadamard 积，应用编程语言的矩阵计算广播机制.

综上所述，本文提出的改进的 CSA 模块用于特征提取的步骤如下：

输入. 特征张量 $\boldsymbol{x}$.

输出. 增强特征张量 $\boldsymbol{y}$.

Step1. 按照式(2)，根据 $\boldsymbol{x}$ 计算横向和纵向平均和最大池化特征$[\boldsymbol{x}^{\mathrm{h_avg}},\boldsymbol{x}^{\mathrm{h_max}},\boldsymbol{x}^{\mathrm{w_avg}},\boldsymbol{x}^{\mathrm{w_max}}]$.

Step2. 对 Step1 输出进行特征融合交互，再进行特征切分，得到$[\boldsymbol{f}^{\mathrm{h_avg}},\boldsymbol{f}^{\mathrm{h_max}},\boldsymbol{f}^{\mathrm{w_avg}},\boldsymbol{f}^{\mathrm{w_max}}]$.

Step3. 按照式(3)(4)，根据$[\boldsymbol{f}^{\mathrm{h_avg}},\boldsymbol{f}^{\mathrm{h_max}}]$和$[\boldsymbol{f}^{\mathrm{w_avg}},\boldsymbol{f}^{\mathrm{w_max}}]$计算横向和纵向注意力权重$[\boldsymbol{g}^{\mathrm{h}},\boldsymbol{g}^{\mathrm{w}}]$.

Step4. 对输入张量 $\boldsymbol{x}$ 进行横向和纵向注意，得到输出张量 $\boldsymbol{y}$.

2.7 像素级离散深度值的概率分布

将 VisionNet 输出的 3 个尺度特征 $\boldsymbol{f}_{\mathrm{d1}}$，$\boldsymbol{f}_{\mathrm{d2}}$ 插值至图片分辨率 1/2 后与 $\boldsymbol{f}_{\mathrm{d3}}$ 在通道维度上拼接，然后使用 1×1 的卷积操作将通道数调整至 256，最后使用 Softmax 函数将多尺度融合特征转换为像素级离散深度值的概率分布 $\boldsymbol{p}\in\mathbb{R}^{256\times H/2\times W/2}$.

2.8 深度图的构建

由于将深度离散化为 k 个深度值并为每个像素分配一个深度值会导致最终预测的深度图不平滑[5]，因此本文的深度预测最终结果是自适应离散深度值与离散深度值的概率分布的线性结合，即

$$\hat{\boldsymbol{D}}(i,j)=\sum_k \boldsymbol{c}_k\times\boldsymbol{p}_k(i,j) \tag{5}$$

其中，$\hat{\boldsymbol{D}}$ 表示线性结合后的深度图；$\boldsymbol{c}_k$ 表示第 k 个深度值；$\boldsymbol{p}_k(i,j)$ 表示对于像素点 (i,j) 中的第 k 个深度值的概率. 与 Bhat 等[5]的方法相同，本文的上采样至原始分辨率的 1/2，最后将深度图双线性插值至原始分辨率.

从概率论的角度来看，二者相乘及求和可以理解为随机变量的期望求解. 本文将每个像素点的深度值的期望值看作最终该像素点的深度. 在计算过程中，每个像素点的深度看作一个随机变量，$\boldsymbol{c}_k$ 和 $\boldsymbol{p}_k(i,j)$ 共同确定了像素 (i,j) 的深度值的分布. 满足概率分布约束 $\sum_k \boldsymbol{p}_k(i,j)=1$.

2.9 损失函数

为了缓解不同尺度给深度预测误差带来的不平衡的问题，Eigen 等[4]提出尺度不变损失函数. 给定某个像素的预测深度值 $\hat{\boldsymbol{d}}_i$ 和真实深度值 $\boldsymbol{d}_i$，深度的真实值与预测值之间的对数距离 $\boldsymbol{g}_i=\ln\hat{\boldsymbol{d}}_i-\ln\boldsymbol{d}_i$. 假设像素的数量为 n，则损失函数[20,40-41]为

$$L(\boldsymbol{D},\hat{\boldsymbol{D}})=\alpha\sqrt{\frac{1}{n}\sum_i \boldsymbol{g}_i^2-\frac{\lambda}{n}\left(\sum_i \boldsymbol{g}_i\right)^2}\ .$$

其中，与训练时的学习率参数类似，α 的取值能够影响模型训练收敛的速度. 与文献[20,40-41]相同，本文中，α=10，λ=0.85.

3 实验与结果分析

3.1 数据集及评价指标

为了验证本文方法的可行性，在 Replica[42]和 Matterport3D[43]这 2 个数据集上进行实验. Replica 数据集包含公寓、办公室、卧室、酒店等 18 个不同的场景；在这些场景中对 1 740 个位置拍照，每个位置取 0°, 90°, 180°和 270°这 4 个不同方向，最终得到 6 960 张图片；将其中 15 个场景中的 5 496 张图片作为训练集，3 个场景中的 1 464 张图片作为测试集. Matterport3D 数据集包含 77 个场景中的 16 844 个位置拍摄 67 376 张图片信息，将其中 59 个场景中的 40 176 张图片作为训练集，10 个场景中的 13 592 张图片作为验证集，8 个场景中的 13 602 张图片作为测试集. 2 个数据集中，输入模型的 RGB 单目图片的分辨率相同，均为 128×128 像素. 为了增加图片的多样性，本文对图片的对比度、色度和亮度进行数据增强，以提高模型的泛化能力.

回波通过房间脉冲响应(room impulse response, RIR)[28]与处于人耳听觉范围内 20 Hz~20 kHz 的 3 ms 发射信号卷积进行仿真. 与 BITD 一致，对于 Replica 数据集，本文使用采样率 44.1 kHz 的扫描信号；在 Matterport3D 数据集，则使用 16 kHz 的扫描信号. EchoNet 的输入是回波经过 STFT 的频谱图，STFT 的窗口大小为 512，窗口长度为 64(通过零填充补到 512)，窗函数采用 Hann 窗，帧移为 16.

为了验证本文方法在真实数据的有效性，本文还在 BatVision 数据集[44]上进行实验. 2023 年，Brunetto 等[44]对回波预测深度信息进行研究并公开了 BatVision 数据集. BatVision 数据集是通过一辆搭载双目 3D 相机、扬声器、2 个麦克风的移动小车在伯克利大学的教室、办公室、长走廊等的场

景收集得到的；位于车头的扬声器发射 20 Hz~20 kHz 的 3 ms 信号，位于两侧的麦克风接收到时长为 72.5 ms 的回波信号，位于车头的双目 3D 相机获得场景的深度图和 RGB 图片(以左目图片作为模型的视觉输入). BatVision 数据集共有 52 220 张图片和对应回波信号，其中，39 654 张图片作为训练集，7 618 张图片作为验证集，5 038 张图片作为测试集. 为与 BatVision 数据集进行对比，本文将深度图中可估计的最大深度限定在 12 m，RGB 分辨率设为 128×128 像素，其他设置均相同.

为了与 BITD 进行对比，本文也使用均方根误差(root mean square error, RMSE)、相对误差(relative error, REL)、对数误差(logarithmic error, LE)及阈值误差(A)为评价指标，计算公式分别为

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i\in n}\left|\boldsymbol{d}_i-\hat{\boldsymbol{d}}_i\right|^2},$$

$$\mathrm{REL}=\frac{1}{n}\sum_{i\in n}\frac{\left|\boldsymbol{d}_i-\hat{\boldsymbol{d}}_i\right|}{\boldsymbol{d}_i},$$

$$\mathrm{LE}=\frac{1}{n}\sum_{i\in n}\ln\left(\boldsymbol{d}_i\right)-\ln\left(\hat{\boldsymbol{d}}_i\right),$$

$$A=\max\left(\frac{\boldsymbol{d}_i}{\hat{\boldsymbol{d}}_i},\frac{\hat{\boldsymbol{d}}_i}{\boldsymbol{d}_i}\right)\leqslant T.$$

其中，n 表示像素个数；$\boldsymbol{d}_i,\boldsymbol{d}_i^*$ 表示预测深度和真实深度；T 表示阈值集合，大小设置为{1.25, 1.25^2,1.25^3}[28,30]，评价指标 A 的在 3 个阈值下的准确率为 $A_{1.25}$, $A_{1.25^2}$, $A_{1.25^3}$ (下文同).

3.2　实验配置

本文中，实验平台如下：Intel Xeon(R) W-2133 CPU，64 GB 内存，GeForce RTX 2080Ti 图片处理器，通过 PyTorch 深度学习框架训练模型，Ubuntu 22.04 操作系统. 使用 Adam 优化器，学习率为10^{-4}，动量为 0.9，权重衰减为5×10^{-5}，批处理大小为 32，训练 150 轮.

3.3　数据评估和分析

为了验证本文方法的有效性，在 Replica 和 Matterport3D 数据集上，与 ECHO2DEPTH, RGB2-DEPTH, VisualEchoes[28]和 BITD[30]这 4 种方法进行定性实验，其中, ECHO2DEPTH 和 RGB2DEPTH 分别为仅使用回波和单目图片通过 VisualEchoes[28]. 实验结果如表 1 和表 2 所示，对比实验所用训练与测试数据均一致，训练前的预处理、EchoNet 和 MaterialNet 的骨干网络与 BITD[30]的设置一致. 可以看出，本文方法在 4 个评价指标上均取得有竞争力的结果，在 Replica 数据集上，RMSE 由 0.249 降至 0.204，提升约 18%，REL 由 0.118 降至 0.095，提升约 19%；在 Matterport3D 数据集上，RMSE 由 0.950 降至 0.875，提升约 8%，REL 由 0.175 降至 0.161，提升约 8%，展示了本文方法的优越性.

表 1　5 种方法在 Replica 数据集上的比较

方法	RMSE↓	REL↓	LE↓	$A_{1.25}$↑	$A_{1.25^2}$↑	$A_{1.25^3}$↑
ECHO2DEPTH	0.713	0.347	0.134	0.580	0.772	0.868
RGB2DEPTH	0.374	0.202	0.076	0.749	0.883	0.945
VisualEchoes[28]	0.346	0.172	0.068	0.798	0.905	0.950
BITD[30]	0.249	0.118	0.046	0.869	0.943	0.970
本文	**0.204**	**0.095**	**0.038**	**0.907**	**0.964**	**0.980**

注. 粗体表示最优值，↓表示指标越小越好，↑表示越大越好.

表 2　5 种方法在 Matterport3D 数据集上的比较

方法	RMSE↓	REL↓	LE↓	$A_{1.25}$↑	$A_{1.25^2}$↑	$A_{1.25^3}$↑
ECHO2DEPTH	1.778	0.507	0.192	0.464	0.642	0.759
RGB2DEPTH	1.090	0.260	0.111	0.592	0.802	0.910
VisualEchoes[28]	0.998	0.193	0.083	0.711	0.878	0.945
BITD[30]	0.950	0.175	0.079	0.733	0.886	0.948
本文	**0.875**	**0.161**	**0.074**	**0.744**	**0.905**	**0.962**

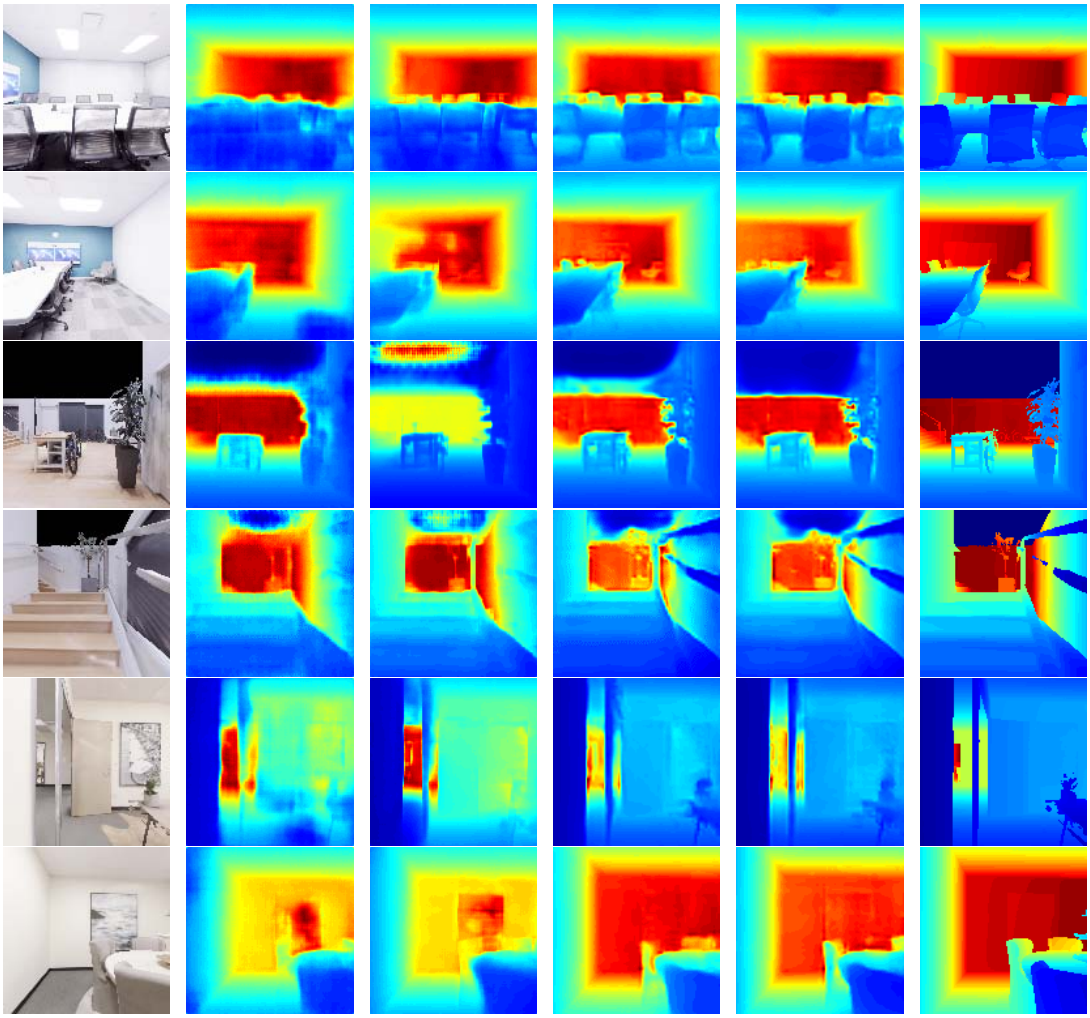
注. 粗体表示最优值，↓表示指标越小越好，↑表示越大越好.

本文还在 Replica 数据集上进行了定性实验，结果如图 3 所示，其中, NoEchoes 代表不添加回波信息. 可以看出，第 5 行和第 6 行中，本文方法对于无纹理的区域(墙壁)和无任何物体的空间都有效，而 BITD 方法[30]难以还原较远处的距离信息；第 1 行和第 5 行中，室内场景本身包含许多小物体(桌椅、盆栽)和背景(墙壁)，场景信息较为复杂，细小物体容易被忽视，本文提出的 CSA 模块能够同时聚焦于某些特征较为细小的物体和背景信息，因此能够对特征明显的细节进行还原.

3.4　消融实验

为了验证本文方法以及不同模块的有效性，在 Replica 数据集上进行消融实验.

本文的基线(Baseline)为 VisionNet 和 MaterialNet 的回归-分类方法，并通过在 Baseline 上加入回波和不同的模块来验证每个模块是否有效，结果如表 3 所示. 其中，Echo 表示回波、DF 表示解码特征融合(decoder fusion, DF). 实验结果表明，加入上述模块对网络都能起到递增的效果.



a. 输入图片 b. VisualEchoes[28] c. BITD[30] d. NoEchoes e. 本文 f. 真实深度图

图 3 3 种方法在 Replica 数据集上的定性比较

表 3 Replica 数据集上加入不同模块的消融实验

方法	RMSE↓	REL↓	LE↓	$A_{1.25}$ ↑	$A_{1.25^2}$ ↑	$A_{1.25^3}$ ↑
Baseline	0.266	0.140	0.056	0.827	0.928	0.967
+PPM	0.232	0.110	0.045	0.869	0.946	0.975
+PPM+Echo	0.220	0.101	0.041	0.902	0.959	0.975
+PPM+Echo+CSA	0.212	0.096	0.040	0.903	0.961	0.977
+PPM+Echo+CSA+DF	**0.204**	**0.095**	**0.038**	**0.907**	**0.964**	**0.980**

注. 粗体表示最优值，↓表示指标越小越好，↑表示越大越好.

本文对 VisionNet 编码器的骨干网络进行消融实验，选择 SwinTransformer，DenseNet121，ResNet50 和 Baseline 进行实验，除编码器网络不同外，其他均保持一致，结果如表 4 所示. 可以看出，在图片分割的任务上，虽然 SwinTransformer 在 ADE20K 数据集[45]上取得了比基于 CNN 的特征提取骨干网络更好的结果[39]，但在 Replica 这样相对较小的数据集中，SwinTransformer 缺乏足够的数据学习到 CNN 这样的局部偏置，因此基于 CNN 的骨干网络的表现比 SwinTransformer 更好. 实验结果表明，本文的混合模型取得了最好的结果.

通过与 SE, CA 和卷积注意力(convolutional block attention module, CBAM)进行实验，验证在解码过程中 CSA 模块的有效性，结果如表 5 所示，其他设置均相同. 可以看出，本文的 CSA 模块加强了重点区域的关注，在一定程度上缓解了信息的丢失.

表 4　Replica 数据集上不同特征编码器的消融实验

编码器骨干网络	RMSE↓	REL↓	LE↓	$A_{1.25}$↑	$A_{1.25^2}$↑	$A_{1.25^3}$↑
SwinTransformer[38]	0.320	0.163	0.066	0.794	0.909	0.959
DenseNet[10]	0.290	0.145	0.061	0.809	0.923	0.965
ResNet50[9]	0.280	0.141	0.059	0.819	0.922	0.961
Baseline	**0.266**	**0.140**	**0.056**	**0.827**	**0.928**	**0.967**

注. 粗体表示最优值，↓表示指标越小越好，↑表示越大越好.

表 5　不同注意力模块的消融实验

注意力模块	RMSE↓	REL↓	LE↓	$A_{1.25}$↑	$A_{1.25^2}$↑	$A_{1.25^3}$↑
CA[26]	0.215	**0.095**	0.042	0.889	0.961	**0.980**
CBAM[25]	0.214	0.099	0.040	0.903	0.962	0.979
SE[24]	0.211	0.097	0.039	**0.907**	**0.965**	0.979
CSA	**0.204**	**0.095**	**0.038**	**0.907**	0.964	**0.980**

注. 粗体表示最优值，↓表示指标越小越好，↑表示越大越好.

为了验证视觉、回波和材料各模态对模型的影响，在视觉模态基础上依次添加材料和回波信息进行比较，结果如表 6 所示，其中，V 表示视觉信息，M 表示材料分类信息，E 表示回波信息. 实验结果证明，添加不同模态的信息的对模型的预测能起到更好的效果，对于单目深度图的预测更加精确.

表 6　Replica 数据集上不同模态信息的消融实验

模态信息	RMSE↓	REL↓	LE↓	$A_{1.25}$↑	$A_{1.25^2}$↑	$A_{1.25^3}$↑
V	0.222	0.111	0.046	0.868	0.948	0.978
V+M	0.215	0.102	0.044	0.874	0.946	0.978
V+M+E	**0.204**	**0.095**	**0.038**	**0.907**	**0.964**	**0.980**

注. 粗体表示最优值，↓表示指标越小越好，↑表示越大越好.

尽管音频对于局部误差比较大，但在大量包含无纹理区域的室内场景下，回波却包含丰富的距离信息. 因此本文采用回波预测离散深度值，采用视觉预测深度值的概率分布，本文称该方案为 E-V 方案；与之对应，还有一套相反的方案，称为 V-E 方案，通过视觉预测离散深度值，通过回波预测深度值的概率分布：这 2 套方案在 Replica 数据集上的实验结果如表 7 所示. 可以看出，E-V 方案优于 V-E 方案，说明通过回波信息预测离散深度值是可行的.

表 7　不同模态组合对深度估计的消融实验

方案	RMSE↓	REL↓	LE↓	$A_{1.25}$↑	$A_{1.25^2}$↑	$A_{1.25^3}$↑
V-E	0.226	0.103	0.041	0.900	0.957	0.975
E-V	**0.204**	**0.095**	**0.038**	**0.907**	**0.964**	**0.980**

注. 粗体表示最优值，↓表示指标越小越好，↑表示越大越好.

由于声音不可避免地受噪声影响，因此本文研究包含噪声的回波对模型的影响. 在信噪比为 10 dB 和 20 dB 的情况下，分别进行 3 组训练-测试组合实验，结果如表 8 所示，其中，在回波中所加噪声为加性高斯噪声. 可以看出，本文方法在包含一定的噪声情况下仍有一定的稳定性.

Replica 和 Matterport3D 数据集是经过仿真平台计算得到的. 为了验证本文方法在真实数据下的表现，在 BatVision 数据集上进行实验，结果如表 9 所示. 可以看出，由于 BatVision 与 Materporrt3D 数据集具有相似的结构，而在任务难度相当的情况下二者具有相似的推理结果，因此本文方法同样适用于真实场景的深度估计；真实场景的数据已经包含真实噪声信息，进一步证明本文方法对噪声的鲁棒性.

表 8　不同信噪比噪声的鲁棒性分析

信噪比/dB	训练集噪声	测试集噪声	RMSE↓	REL↓	LE↓	$A_{1.25}$↑	$A_{1.25^2}$↑	$A_{1.25^3}$↑
10	有	无	0.237	0.124	0.057	0.818	0.924	0.972
	无	有	0.236	0.107	0.046	0.879	0.954	0.978
	有	有	0.220	0.098	0.040	0.904	0.960	0.979
20	有	无	0.228	0.128	0.052	0.845	0.944	0.975
	无	有	0.233	0.106	0.045	0.893	0.960	0.980
	有	有	0.216	0.097	0.040	0.904	0.961	0.980

注. ↓表示指标越小越好，↑表示越大越好.

表 9 2 种方法在 2 个数据集上的比较

方法	数据集	RMSE↓	REL↓	LE↓	$A_{1.25}$↑	$A_{1.25^2}$↑	$A_{1.25^3}$↑
BITD[30]	Matterport3D	0.950	0.175	0.079	0.733	0.886	0.948
	BatVision	0.901	0.234	0.097	0.688	0.888	0.942
本文	Matterport3D	0.875	0.161	0.074	0.744	0.905	0.962
	BatVision	0.877	0.195	0.089	0.699	0.904	0.960

注. ↓表示指标越小越好, ↑表示越大越好.

4 结 语

为了解决视觉模态对于无纹理区域的预测困难以及编码-解码结构的信息丢失问题, 本文提出视听融合的单目深度估计方法. 首先设计一个双分支网络 VisionEchoNet: 一端在 CNN 提取回波特征与材料特征后双线性结合在一起, 利用 PPM 结合多尺度特征信息获得不同大小感受野的信息, 从而自适应地估计离散深度值的集合; 另一端利用 CNN 提取单目图片的局部特征, 送入 Transformer-Encoder 模块进行全局交互, 使全局信息与局部信息相互补偿. 为防止信息丢失, 在解码阶段提出改进的 CSA 模块, 并将不同阶段的解码特征融合在一起生成像素深度值的概率分布. 最后, 通过像素深度的期望值预测深度. 在 Replica 和 Matterport3D 数据集上的实验结果表明, 最终的深度预测结果在 5 个评价指标上均取得具有竞争力的结果, 在 2 个数据集上, RMSE 分别达到了 0.204 和 0.875; REL 分别达到 0.095 和 0.161; 该方法对于深度图预测的边缘轮廓更加清晰、对无纹理区域距离预测准确.

由于相机和麦克风位置不同, 因此单目相机图片与声音回波信号往往产生一定的误匹配. 虽然信息采集装置的绝对位置不同, 但是对于特定的数据集或深度估计装置, 视觉信号采集装置(相机)和回声信号采集装置(麦克风)相对位置固定. 信息采集装置的绝对位置不同造成的误匹配信息包含于训练集中, 可以通过数据驱动的方法学习这些误匹配, 在训练中自动地校准二者信息的误匹配. 数据驱动的方法能应用于特定的数据分布, 对于视听融合的深度估计应用, 训练好的模型仅能够应用于特定传感器布局的场景. 这些都是本文方法的局限性, 但是对于具体的产品而言, 固定传感器布局的需求是容易满足的. 未来, 对声音和相机相对位置鲁棒的深度估计是一个具有挑战性的任务.

参考文献(References):

[1] Durrant-Whyte H, Bailey T. Simultaneous localization and mapping: Part I[J]. IEEE Robotics & Automation Magazine, 2006, 13(2): 99-110

[2] Ding M Y, Huo Y Q, Yi H W, *et al*. Learning depth-guided convolutions for monocular 3D object detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Los Alamitos: IEEE Computer Society Press, 2020: 1000-1001

[3] Gudoniene D, Rutkauskiene D. Virtual and augmented reality in education[J]. Baltic Journal of Modern Computing, 2019, 7(2): 293-300

[4] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network[C] //Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2366-2374

[5] Bhat S F, Alhashim I, Wonka P. AdaBins: depth estimation using adaptive bins[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 4009-4018

[6] Hoiem D, Efros A A, Hebert M. Automatic photo pop-up[J]. ACM Transactions on Graphics, 2005, 24(3): 577-584

[7] Saxena A, Sun M, Ng A Y. Make3D: learning 3D scene structure from a single still image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 824-840

[8] Tang C, Hou C P, Song Z J. Depth recovery and refinement from a single image using defocus cues[J]. Journal of Modern Optics, 2015, 62(6): 441-448

[9] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[10] Huang G, Liu Z, Van Der Maaten L, *et al*. Densely connected convolutional networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 4700-4708

[11] Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale[OL]. [2023-04-22]. https://arxiv.org/abs/2010.11929v1

[12] Laina I, Rupprecht C, Belagiannis V, *et al*. Deeper depth prediction with fully convolutional residual networks[C] //Proceedings of the 4th International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 2016: 239-248

[13] Peng C, Zhang X Y, Yu G, *et al*. Large kernel matters-improve semantic segmentation by global convolutional network[C] //Proceedings of the IEEE Conference on Computer Vision and

Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 4353-4361

[14] Chen L C, Papandreou G, Kokkinos I, *et al*. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848

[15] Zhao H S, Shi J P, Qi X J, *et al*. Pyramid scene parsing network[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 2881-2890

[16] Zheng S X, Lu J C, Zhao H S, *et al*. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 6881-6890

[17] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 3431-3440

[18] Cao Y Z H, Wu Z F, Shen C H. Estimating depth from monocular images as classification using deep fully convolutional residual networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(11): 3174-3182

[19] Li Z Y, Wang X Y, Liu X M, *et al*. BinsFormer: revisiting adaptive bins for monocular depth estimation[OL]. [2023-04-22]. https://arxiv.org/abs/2204.00987

[20] Agarwal A, Arora C. Attention attention everywhere: monocular depth prediction with skip attention[C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2023: 5861-5870

[21] Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 12179-12188

[22] Carion N, Massa F, Synnaeve G, *et al*. End-to-end object detection with transformers[C] // Proceedings of the 16th European Conference on Computer Vision. Berlin: Springer, 2020: 213-229

[23] Chen J N, Lu Y Y, Yu Q H, *et al*. TransUNet: transformers make strong encoders for medical image segmentation[OL]. [2023-04-22]. https://arxiv.org/abs/2102.04306

[24] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7132-7141

[25] Woo S, Park J, Lee J Y, *et al*. CBAM: convolutional block attention module[C] //Proceedings of the 15th European Conference on Computer Vision. Heidelberg: Springer, 2018: 3-19

[26] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 13713-13722

[27] Christensen J H, Hornauer S, Yu S X. BatVision: learning to see 3D spatial layout with two ears[C] //Proceedings of the IEEE International Conference on Robotics and Automation. Los Alamitos: IEEE Computer Society Press, 2020: 1581-1587

[28] Gao R H, Chen C A, Al-Halah Z, *et al*. VisualEchoes: spatial image representation learning through echolocation[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 658-676

[29] Antonacci F, Filos J, Thomas M R P, *et al*. Inference of room geometry from acoustic impulse responses[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(10): 2683-2695

[30] Parida K K, Srivastava S, Sharma G. Beyond image to depth: improving depth prediction using echoes[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 8268-8277

[31] Hu J J, Ozay M, Zhang Y, *et al*. Revisiting single image depth estimation: toward higher resolution maps with accurate object boundaries[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 1043-1051

[32] Hao Z X, Li Y, You S D, *et al*. Detail preserving depth estimation from a single image using attention guided networks[C] // Proceedings of the International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 2018: 304-313

[33] Huynh L, Nguyen-Ha P, Matas J, *et al*. Guiding monocular depth estimation using depth-attention volume[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 581-597

[34] Xu D, Ricci E, Ouyang W L, *et al*. Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 5354-5362

[35] Xu D, Wang W, Tang H, *et al*. Structured attention guided convolutional neural fields for monocular depth estimation[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 3917-3925

[36] Díaz R, Marathe A. Soft labels for ordinal regression[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 4738-4747

[37] Chen C A, Jain U, Schissler C, *et al*. SoundSpaces: audio-visual navigation in 3D environments[C] //Proceedings of the 16th European Conference on Computer Vision. Heidelberg: Springer, 2020: 17-36

[38] Bell S, Upchurch P, Snavely N, *et al*. Material recognition in the wild with the materials in context database[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 2479-2387

[39] Liu Z, Lin Y T, Cao Y, *et al*. Swin transformer: hierarchical vision transformer using shifted windows[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 10012-10022

[40] Wang Qiuchen, Shuai Hui, Liu Qingshan. Monocular accumulated depth estimation with recursive feature fusion[J]. Journal of Computer-Aided Design & Computer Graphics, 2022,

34(10): 1533-1541(in Chinese)
(王秋晨, 帅惠, 刘青山. 递归特征融合的单目深度累积估计[J]. 计算机辅助设计与图形学学报, 2022, 34(10): 1533-1541)

[41] Lee J H, Han M K, Ko D W, *et al*. From big to small: multi-scale local planar guidance for monocular depth estimation[OL]. [2023-04-22]. https://arxiv.org/abs/1907.10326v4

[42] Straub J, Whelan T, Ma L N, *et al*. The replica dataset: a digital replica of indoor spaces[OL]. [2023-04-22]. https://arxiv.org/abs/1906.05797

[43] Chang A, Dai A, Funkhouser T, *et al*. Matterport3D: learning from RGB-D data in indoor environments[C] //Proceedings of the International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 2017: 6387-6395

[44] Brunetto A, Hornauer S, Yu S X, *et al*. The audio-visual BatVision dataset for research on sight and sound[OL]. [2023-04-22]. https://arxiv.org/abs/2303.05257

[45] Zhou B L, Zhao H, Puig X, *et al*. Semantic understanding of scenes through the ADE20K dataset[J]. International Journal of Computer Vision, 2019, 127(3): 302-321