计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2024,33(7):121−128 [doi: 10.15888/j.cnki.csa.009587]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

自适应多尺度特征融合的单目图像深度估计①

陈国军, 付云鹏, 于丽香, 崔　涛

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通信作者: 陈国军, E-mail: 965823503@qq.com

摘　要: 在基于深度学习的单目图像深度估计方法中, 卷积神经网络在下采样过程中会出现图像深度信息丢失的情况, 导致物体边缘深度估计效果不佳. 提出一种多尺度特征融合的方法, 并采用自适应融合的策略, 根据特征数据动态调整不同尺度特征图的融合比例, 实现对多尺度特征信息的充分利用. 由于空洞空间金字塔池化 (ASPP) 在单目深度估计任务中, 会丢失图像中的像素点信息, 影响小物体的预测结果. 通过在对深层特征图使用 ASPP 时融合浅层特征图的丰富特征信息, 提高深度估计结果. 在 NYU-DepthV2 室内场景数据集的实验结果表明, 本文所提方法在物体边缘处有更准确的预测, 并且对小物体的预测有明显的提升, 均方根误差 (RMSE) 达到 0.389, 准确率 (δ <1.25) 达到 0.897, 验证了方法的有效性.

关键词: 单目图像; 深度估计; 卷积神经网络; 多尺度特征

引用格式: 陈国军,付云鹏,于丽香,崔涛.自适应多尺度特征融合的单目图像深度估计.计算机系统应用,2024,33(7):121–128. http://www.c-s-a.org.cn/1003-3254/9587.html

Monocular Image Depth Estimation with Adaptive Multi-scale Feature Fusion

CHEN Guo-Jun, FU Yun-Peng, YU Li-Xiang, CUI Tao

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: In the monocular image depth estimation method based on deep learning, the depth information of the image is lost during the subsampling process of the convolutional neural networks, which leads to poor depth estimation of object edges. To solve this problem, this study presents a multi-scale feature fusion method, and an adaptive fusion strategy is adopted to dynamically adjust the fusion ratio of feature maps of different scales according to feature data to make full use of multi-scale feature information. In the monocular depth estimation task using atrous spatial pyramid pooling (ASPP), the pixel information loss affects the prediction results of small objects. When using ASPP on deep feature maps, the depth estimation result is improved by fusing rich feature information of shallow feature maps. The experimental results on the NYU-DepthV2 indoor dataset show that the method proposed in this study has a more accurate prediction of object edges and significantly improves the prediction of small objects. The root mean square error (RMSE) reaches 0.389 and the accuracy (δ<1.25) reaches 0.897, which verifies the effectiveness of the method.

Key words: monocular image; depth estimation; convolutional neural network (CNN); multi-scale feature

1　引言

计算机视觉是人工智能的一个重要领域, 它使计算机能够从数字图像, 视频等一些视觉输入中获取有意义的信息. 其中深度信息是计算机视觉任务中基本的任务之一, 可广泛应用于三维重建、自动驾驶、虚拟现实、工业生产等任务.

① 基金项目: 山西省交通建设科技项目 (2019-2-8)
收稿时间: 2023-11-18; 修改时间: 2023-12-20, 2024-01-18, 2024-03-19; 采用时间: 2024-03-28; csa 在线出版时间: 2024-06-05
CNKI 网络首发时间: 2024-06-11

传统的深度信息获取方式主要分为主动深度传感和被动测距传感, 主动深度传感主要是通过机械设备来获取场景的深度信息, 常见的设备比如 LiDAR、RGB-D 相机. 目前由于 LiDAR 设备本身比较昂贵, 且人力成本较高, 难以推广. RGB-D 相机根据近红外光的反射, 通过 TOF (time of flight)、结构光等方法获得场景的深度信息. 但是由于深度测距设备本身的测量距离较短, 且对场景环境有一定要求, 因此无法在室外环境广泛应用. 被动测距传感的方法主要分为两类: 单视图和多视图. 多视图的方法主要使用双目图像进行推理, 根据双目测距原理直接计算出场景的深度信息, 这种方法的优点是成本低, 室内外环境都适用, 缺点是基线长度限制了测量范围. 单目深度估计从单幅图像中估计像素深度信息, 成本低, 应用灵活方便. 然而, 由于缺少诸如运动、立体视觉等可靠的深度线索, 并且对于一张 RGB 图像来说, 其深度信息有无数种可能, 也就是对应无数个可能的深度图, 因此该任务非常具有挑战性.

随着深度学习的兴起与发展, 卷积神经网络在单目视觉中的应用也快速发展. 深度学习在解决目标识别、图像分割等问题中性能优越, 这些方法也大量迁移应用于单目深度估计. 基于 RGB 图像与深度图像存在某种映射这一基本假设, Eigen 等[1]将卷积神经网络应用于单目深度估计任务, 并预测了较准确的深度图. 随后 Eigen 和他的团队提出可以用于多任务的通用多尺度网络框架. Laina 等[2]提出了全卷积残差网络 (fully convolutional residual network, FCRN), 提高了输出分辨率, 并提出了更高效的上采样模块, 同时加入 berHu 损失函数, 进一步提高了网络性能. Hu 等[3]提出一种多尺度特征融合的网络结构, 包含编码器、解码器、多尺度特征融合模块和细化模块 4 个部分, 提升了网络预测结果的精度以及在物体边缘处的预测. Chen 等[4]提出一种基于结构感知的残差金字塔网络结构, 并提出了残差细化模块 (residual refinement module, RRM) , 用于在解码器优化不同尺度的深度图. 此外, 还提出自适应密集特征融合模块 (adaptive dense feature fusion, ADFF) 来融合所有尺度的特征, 用于残差预测, 进一步优化解码器的残差深度图.但是以固定比例融合所有尺度的特征并不能更好地利用不同尺度特征图的相同物体的特征信息. Song 等[5]、Zhang 等[6]都将拉普拉斯金字塔结构用于架构设计中，对不同层次的编码端特征进行残差恢复, 并逐步融合预测结果. Miangoleh 等[7]提出一种通过场景内容自适应合并的方法, 该方法通过合并多尺度特征和不同场景物体的深度估计来生成一个高分辨率的深度图. Ranftl 等[8]引入 vision Transformer 代替卷积神经网络作为网络主干, 为深度估计提供了更细粒度和更全局性的预测. 由于深度信息有连续性的特点, 所以也有结合条件随机场 (conditional random field, CRF) 的深度估计方法, 比如 Cao 等[9]、Yuan 等[10]通过优化全连接 CRF 并与多尺度网络结合来提高性能. Wu 等[11]提出了一种多层次上下文和多模态融合网络 (multilevel context and multimodal fusion network, MCMFNet), 用于融合多尺度多层次上下文特征图, 并从深度信息中学习目标边缘, 可以获取具有清晰边界的检测结果. 但是, Hu 等[12]已经证明, 不同的通道特征在模型中并不发挥同样重要的作用, 因此, 加强重要通道的特征, 弱化不重要通道的特征有助于提升模型功能性. Yang 等[13]提出了深度自适应融合模块 (deep adaptive fusion module, DAFM), 对不同尺度的深度图进行估计, 并对这些深度图进行加权求和, 通过减少不同尺度深度图之间的信息差距, 提高融合效果. Xu 等[14]提出一种结合坐标注意力和特征融合的预测框架, 包含注意力、多尺度和特征融合模块, 将有用的底层和高层上下文特征集成在一起. Xu 等[15]提出一种多层次特征融合模块, 通过水平和垂直结构连接特征图. 并引入一个可学习的权重, 通过该权重实现自适应融合, 以保留最佳特征.

Lee 等[16]将空洞空间卷积池化 (atrous spatial pyramid pooling, ASPP) 模块引入单目深度估计任务, 并提出局部平面引导模块 (local planar guidance, LPG) , 将解码器的特征图与最终输出深度图联系起来, 取得了较好的预测结果. Wu 等[17]提出 DenseASPP 网络模型, 通过融合两种跳层连接来利用不同尺度下的视觉特征, 学习其中的有用特征, 并将残差模块集中到 ASPP 中, 将残差块中的特征用于增加后向梯度. 廖志伟等[18]提出一种分层压缩激励 ASPP 的结构, 引入了分层的压缩激励结构块 (hierarchical compress excitation, H-CE), 将空间注意纳入网络结构中, 允许网络进行特征重新校准. 但是单目深度估计任务是密集预测任务, 空洞卷积会丢失较多像素点信息, 对小物体的预测有一定的影响. 在这些方法中, 随着卷积神经网络深度的加深, 经过不断的下采样, 特征图中的深度信息会大量丢

失, 对深度的预测结果产生较大的影响, 导致边缘模糊等问题的出现. 虽然通过跳层连接以及对特征图的上采样可以保留和恢复一部分深度信息, 但是效果不理想.

本文提出一种自适应多尺度特征融合, 对不同特征图进行统一尺度, 根据训练学习得到的权重进行融合, 并在已经融合的特征图基础上进行下一次融合, 实现自底向上的堆叠融合, 更好保留重要的深度信息, 充分利用相同场景中的物体在不同尺度下的特征信息, 解决物体边缘模糊的问题, 进一步提高深度预测的准确性; 其次在使用 ASPP 模块对深层特征图进行空洞卷积时, 引入将浅层特征图池化后的特征图, 丰富场景信息, 改善在单目深度估计中使用 ASPP 模块导致的小物体预测不准确的问题.

2 网络结构设计

本文的网络结构如图 1 所示, 采用编码-解码网络结构, 并在解码端增加多尺度自适应融合模块 (adaptive multi-scale feature fusion, AMFF) 和融合浅层特征信息的 ASPP 模块.

2.1 编码—解码网络结构

整体网络采用编码器-解码器的网络结构. 对于编码器, 使用在 ImageNet[19]数据集上预训练的 DenseNet-161 网络, 将输入的 640×480 大小的 RGB 图像编码为特征向量并输入到解码器. 在解码器端, 将浅层特征图中的特征信息融合到 ASPP 模块, 构建特征中的上下文信息, 然后输入到上采样层. 此外, 图像在下采样过程中得到的不同尺度的特征图通过多尺度自适应融合模块进行融合并将大小固定到不同尺度, 然后通过跳层连接与深层特征融合, 最终通过双线性插值进行上采样, 输出 320×240 大小的深度图. 融合浅层特征信息的 ASPP 模块、多尺度自适应融合模块 (AMFF)、跳层连接以及上采样层共同构成解码器部分.

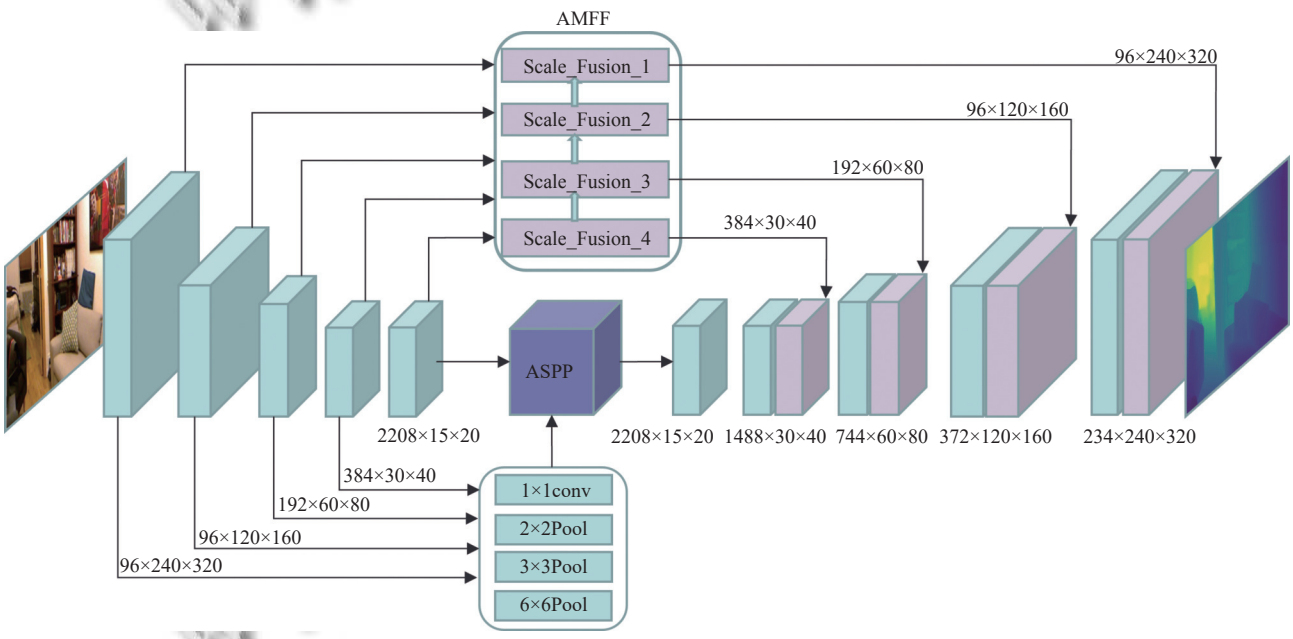


图 1 网络结构

2.2 多尺度自适应融合模块 (AMFF)

多尺度特征融合是一种常用的特征提取方法, 不同尺度的融合特征能够得到更丰富的语义信息, 更多的空间特征细节信息. 受到 Liu 等[20]提出的 ASFF (adaptively spatial feature fusion) 结构启发, 本文引入 AMFF 模块. 但由于 ASFF 结构起初是用于目标检测中, 解决不同尺度的特征图中实例的不一致性, 但是对属于密集预测的单目深度估计任务来说, 与目标检测所需的启发式引导特征选择不同, 因为单目深度估计需要在各级特征下预测相同的深度图, 所以与 ASFF 在解码阶段进行融合不同, 我们对编码阶段的所有特征图进行融合, 减少网络加深造成的特征损失, 并对已融合的特征向后传播, 实现自底向上的堆叠融合, 在这个过程中, 同样采用自适应融合方式, 加强重要通道的特征, 这样融合后的浅层特征更能保留深层特征的信息.

考虑到需要融合的特征图大小与通道数都不一致

的问题, 所以设计融合模块对不同特征图进行尺度调整, 以尺度调整到 96×240×320 为例, 如图 2 所示, 对通道数分别为 96、96、192、384、2 208 的特征图进行尺度统一, 通过步长为 1 的 1×1 卷积实现不同尺度特征图的通道数相同, 得到 5 个通道数都为 96 的特征图, 加上前一个融合模块 Scale_Fusion_2 的输出结果, 对它们进行双线性插值上采样, 得到 6 个 96×240×320 的特征图, 然后通过对这些特征图进行 1×1 卷积得到 λ_0、λ_1、λ_2、λ_3、λ_4、λ_5, 经过 Softmax 函数满足:

$$\alpha^0 = \frac{e^{\lambda_0}}{e^{\lambda_0}+e^{\lambda_1}+e^{\lambda_2}+e^{\lambda_3}+e^{\lambda_4}} \tag{1}$$

得到权重参数α^0、α^1、α^2、α^3、α^4、α^5. 融合模块利用这些权重参数动态调整融合比例, 代替固定融合比例, 实现对编码器阶段不同尺度特征图的自适应融合, 以此更充分地利用不同特征图中的特征信息, 得到的融合特征图可如式 (2) 所示.

$$y_{ij}^{l} = \alpha_{ij}^{0}x_{ij}^{0} + \alpha_{ij}^{1}x_{ij}^{1} + \alpha_{ij}^{2}x_{ij}^{2} + \alpha_{ij}^{3}x_{ij}^{3} + \alpha_{ij}^{4}x_{ij}^{4} + \alpha_{ij}^{5}y_{ij}^{l-1} \tag{2}$$

其中, y_{ij}^{l}表示输出特征图 y^l 在 (i,j) 位置上的特征向量, x_{ij}^{n}是第 n 个输入特征图调整到与第 l 个输入特征图相同的特征图在 (i,j) 位置上的特征向量. α_{ij}^{0}、α_{ij}^{1}、α_{ij}^{2}、α_{ij}^{3}、α_{ij}^{4}、$\alpha_{ij}^{5}\in[0,1]$, 且满足$\alpha_{ij}^{0}+\alpha_{ij}^{1}+\alpha_{ij}^{2}+\alpha_{ij}^{3}+\alpha_{ij}^{4}+\alpha_{ij}^{5}=1$. 最终得到一个大小为 240×320, 通道数为 96 的融合特征图.

整体网络结构如图 1 所示, 其中 AMFF 模块包括 4 个融合模块, 选取编码阶段中大小分别为 240×320、120×160、60×80、30×40、15×20 的特征图以及邻近下一层融合模块的输出作为输入, 4 个融合模块分别对其进行尺度调整和融合操作, 输出结果分别为 96×240×320、96×120×160、192×60×80、384×30×40 的融合特征图, 然后与解码器端的特征图进行融合上采样.

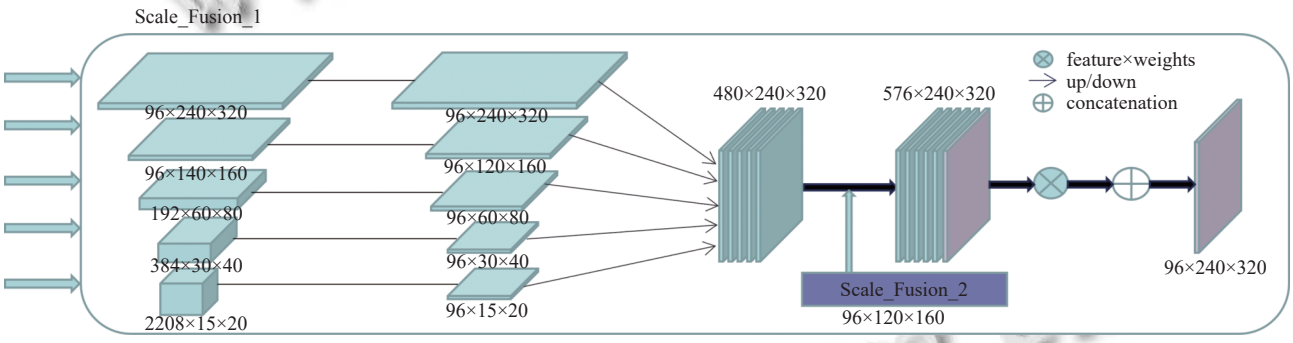


图 2 多尺度自适应融合模块

2.3 融合浅层特征信息的 ASPP 模块

ASPP 模块最初是在 DeepLabV2[21]模型中提出来, 该模块由 4 个不同扩张率的空洞卷积并行构成, 目的是获取更大的感受野, 提取更多的上下文信息. 本文采用空洞率分别为 6、12、18 的空洞卷积, 对于这些大小为 3×3 的卷积核, 通过提高空洞率, 扩大要提取的像素之间的距离, 实现更大的感受野. ASPP 采用多个不同扩张率的空洞卷积对特征图进行提取, 同时对其进行池化以及 1×1 卷积, 最后进行融合. 空洞卷积在增大感受野的同时, 受卷积核结构的影响, 在卷积过程中会丢失一部分像素信息, 虽然多个并行的空洞卷积可以缓解该问题, 但对于像素级的深度预测任务还是有一定的影响, 并且由于深层特征图经过下采样后会丢失大量信息, 并不能较好的保留小物体的特征信息. 所以我们提出融合浅层特征图中图像信息的方法, 通过池化获取浅层特征图中丰富的场景信息, 对深层特征图 F_d 经过空洞卷积后丢失的特征信息进行补全, 将编码段 4 个浅层特征图作为输入, 其中以 96×240×320 的特征图输入为例, 如图 3 所示. 分别使用 6×6、3×3、2×2 池化和 1×1 卷积提取原图 1/2、1/4、1/8、1/16 大小的特征图, 经过提取的特征图通过双线性插值的上采样将分辨率调整到与特征图 F_d 相同的分辨率, 并且通过 1×1 卷积将特征通道数调整到 1/4 大小, 将得到的 4 个特征图进行拼接, 得到与 F_d 的分辨率和通道数都相同的特征图, 然后与 ASPP 模块中多分支并行空洞卷积的输出结果通过 Concat 拼接, 最终通过 1×1 卷积降维输出结果. 图 2 中其余输入特征图均做相同的操作.

2.4 损失函数

深度估计任务的损失函数是考虑真实深度图 y 和网络模型的预测深度图 $\hat{y}$ 的差异. 我们采用与 Alhashim 等[22]提出的方法相同的损失函数, 选择图像深度损失、梯度损失以及结构相似性损失的加权和作为损失函数, 如式 (3) 所示:

$$L(y,\hat{y}) = \lambda L_{\text{depth}}(y,\hat{y}) + L_{\text{grad}}(y,\hat{y}) + L_{\text{SSIM}}(y,\hat{y}) \quad (3)$$

其中, 定义 λ 为 L_{depth} 的权重参数, 并设置 λ=0.1, L_{depth} 深度损失, 是像素级深度值的 L1 损失, 如式 (4) 所示:

$$L_{\text{depth}}(y,\hat{y}) = \frac{1}{n}\sum_{p}^{n}\left|y_p - \hat{y}\right| \quad (4)$$

其中, n 为样本个数, p 为样本, y_p 为样本的真实深度值, $\hat{y}_p$ 为样本的预测深度值. 考虑到深度图中高频失真的部分, 比如场景中物体边界模糊的问题, 我们采用梯度损失 L_{grad} 对边缘信息进行约束. 梯度损失如式 (5) 所示:

$$L_{\text{grad}}(y,\hat{y}) = \frac{1}{n}\sum_{p}^{n}\left|g_x(y_p,\hat{y})\right| + \left|g_y(y_p,\hat{y})\right| \quad (5)$$

其中, $g_x(\cdot)$ 和 $g_y(\cdot)$ 分别计算在 x 分量和 y 分量上 y_p 和 $\hat{y}_p$ 的图像梯度. 最后 L_{SSIM}[23]是结构相似性损失, 从亮度、对比度以及结构 3 个方面考虑真实图与预测图之间的相似性, 惩罚与真实图不相似的预测图, 如式 (6) 所示:

$$L_{\text{SSIM}}(y,\hat{y}) = \frac{1-\text{SSIM}(y,\hat{y})}{2} \quad (6)$$

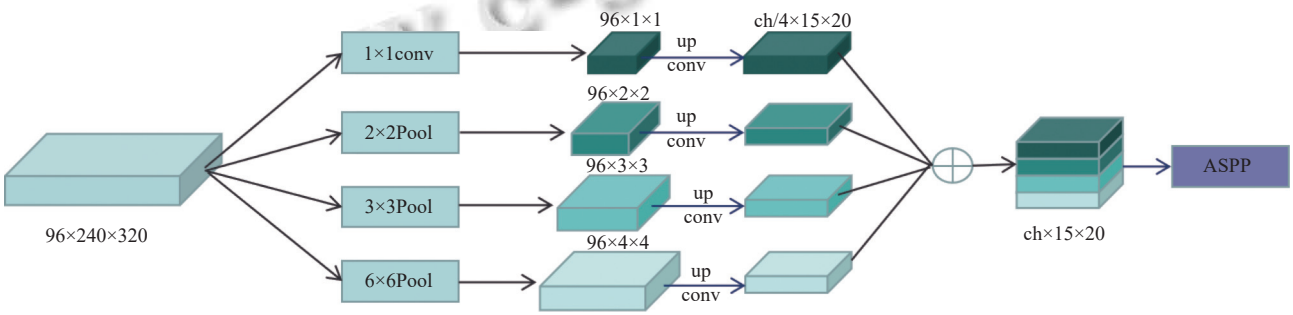


图 3 融合浅层特征信息结构

3 实验结果与分析

3.1 数据集

本文采用 NYU-DepthV2 数据集, 该数据集提供在 640×480 分辨率下的室内场景 RGB 图像和深度图. 该数据集包含 464 种不同室内场景共 12 万张, 249 个场景的 5 万张图片作为训练样本, 其余 215 个场景的 654 张图片作为测试样本. 在训练阶段, 我们的网络模型以原始分辨率的图像输入网络, 将数据集中的真实深度图下采样到 320×240, 与模型输出最终深度图的大小一致. 在测试阶段, 我们将模型预测深度图进行 2 倍上采样以匹配数据集中真实深度图的分辨率, 同时对预测深度图的精度进行评估.

3.2 实验环境与参数

本文提出的网络模型使用 PyTorch 深度学习框架. Python 环境版本为 3.7, Cuda 环境版本为 10.0.130. 网络在显存为 16 GB 的 Tesla P100 上进行训练, 在 GTX 960M 上进行测试评估. 编码器端主干网络是在 ImageNet 上训练的 DenseNet-161, 解码器权重随机初始化. 优化器采用 Adam 优化算法, 学习率为 0.000 1, β_1=0.9, β_2=0.999. Batch 大小为 2, 迭代次数 epoch 为 10.

3.3 评估指标

本文采用 Eigen[1]中提出的评价指标, 从 4 个方面对模型进行评估, 这 4 个评估指标分别为: 平均相对误差 (*AbsRel*)、均方根误差 (*RMSE*)、均方根对数误差 ($RMSE_{\log_{10}}$)、不同阈值下的准确率 (δ_1, δ_2, δ_3), 定义式分别如式 (7)–式 (10).

$$AbsRel = \frac{1}{n}\sum_{p}^{n}\frac{\left|y_p - \hat{y}_p\right|}{y} \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{p}^{n}\left(y_p - \hat{y}_p\right)^2} \quad (8)$$

$$RMSE_{\log_{10}} = \frac{1}{n}\sum_{p}^{n}\left|\log_{10} y_p - \log_{10}\hat{y}_p\right| \quad (9)$$

$$\delta_i = \max\left(\frac{y_p}{\hat{y}_p}, \frac{\hat{y}_p}{y_p}\right) < thr,\ thr = 1.25, 1.25^2, 1.25^3 \quad (10)$$

其中, y_p 和 $\hat{y}_p$ 分别是真实深度图 y 和预测深度图 $\hat{y}$ 的像素点, n 是图像的像素点总数.

3.4 实验结果分析

通过第 2 节设计的网络结构, 在 NYU-DepthV2 数据集上验证模型性能, 并进行不同算法的对比实验, 根据第 3.3 节提出的评估指标计算, 定量实验结果如表 1、表 2 所示.

从表 1 的评估指标来看, 本文算法结果均较为理想, 在 δ_1、$RMSE$、$AbsRel$、$RMSE_{\log 10}$ 指标上来看相较一些主流算法取得了更好的效果.

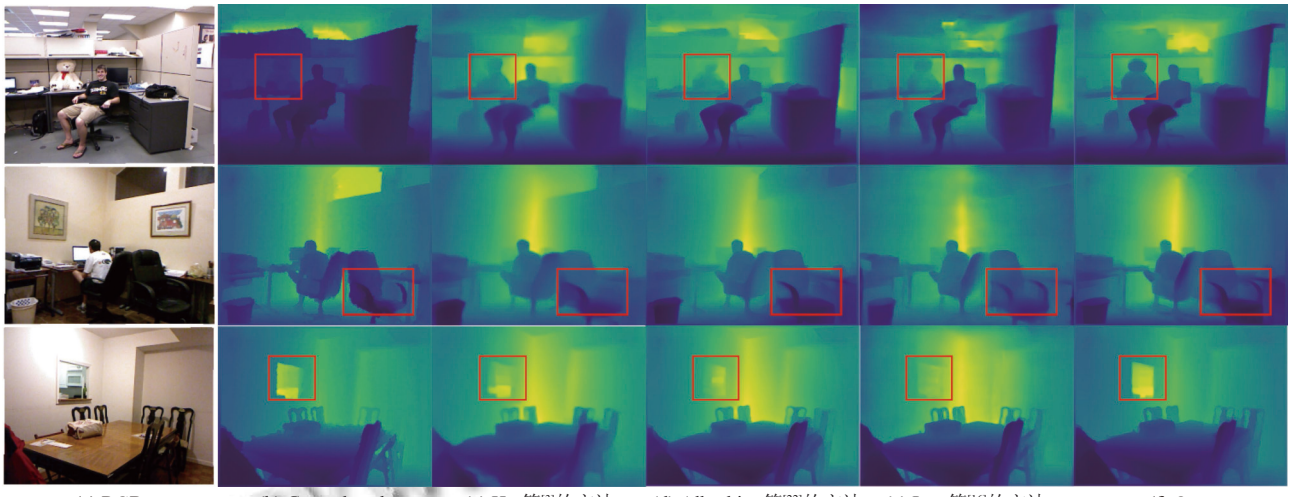
图 4 给出了本文方法与不同方法的深度图对比, 第 1 列为原始 RGB 图像, 第 2 列为真实深度图, 第 3 列为 Hu 等[3]方法结果, 第 4 列为 Alhashim 等[22]方法结果, 第 5 列为 Lee 等[16]方法结果, 最后一列为本文深度估计结果. 可以看出, 本文的方法在物体边缘的预测效果更好, 对人物的轮廓的预测要比 Lee 等[16]方法更加细致, 在图 4 红框中, 玩具熊、桌椅、窗口位置容易出现与周围场景深度一致的问题有所改善.

表 1 实验结果对比

方法	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	$RMSE\downarrow$	$AbsRel\downarrow$	$RMSE_{\log 10}\downarrow$
Eigen等[1]	0.769	0.950	0.988	0.641	0.158	—
Laina等[2]	0.811	0.953	0.988	0.573	0.127	0.055
Hao等[24]	0.841	0.966	0.991	0.555	0.127	0.053
Fu等[25]	0.828	0.965	0.992	0.509	0.115	0.051
Hu等[3]	0.866	0.975	0.993	0.530	0.115	0.050
Alhashim等[22]	0.846	0.974	0.994	0.465	0.123	0.053
Lee等[16]	0.885	0.978	0.994	0.392	0.110	0.047
Xu等[26]	0.884	**0.979**	—	0.398	0.108	0.047
Yang等[13]	0.864	0.972	0.993	0.525	0.115	0.050
Wu等[17]	0.841	0.969	0.994	0.430	0.136	0.054
Ours	**0.807**	0.978	**0.995**	0.389	0.099	**0.043**

表 2 消融实验结果对比

方法	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	$RMSE\downarrow$	$AbsRel\downarrow$	$RMSE_{\log 10}\downarrow$
DenseNet	0.889	0.979	0.994	0.404	0.106	0.045
DenseNet-AMFF	0.893	0.978	0.994	0.398	0.104	0.044
DenseNet-AMFF-ASPP	0.893	0.979	0.995	0.392	0.101	0.044
DenseNet-ASPP_MSFI	0.892	0.979	0.995	0.395	0.100	0.043
DenseNet-AMFF-ASPP_MSFI	**0.897**	0.978	**0.995**	**0.389**	**0.099**	**0.043**



(a) RGB (b) Ground truth (c) Hu 等[3]的方法 (d) Alhashim 等[22]的方法 (e) Lee 等[16]的方法 (f) Ours

图 4 深度图结果对比

3.5 消融实验

为验证本文方法中的各个模块在深度估计中的性能, 在 NYU-DepthV2 数据集上进行了消融实验, 主要分析多尺度特征自适应融合模块、ASPP 模块与浅层特征信息融合机制. 设置两个相融方式: 一是模型是否嵌入特征自适应融合模块, 二是模型是否使用 ASPP 模块并融合浅层特征信息. 由表 2 可以看出, 在对模型嵌入特征自适应融合模块之后在 δ_1、$RMSE$、$AbsRel$、$RMSE_{\log 10}$ 指标上取得更好的结果, 在图 5 的第 3 列、第 4 列以及第 5 列对比下, 提升了在场景物体细节方面的预测结果. 具体来看, 红框中, 使用特征自适应融合模块后, 在第 1 场景下可以预测出灯杆的下半部分, 并且有一个整体的轮廓; 在第 2 个场景下, 可以预测出椅子把手的细节; 第 3 个场景下梯子的细节更为清晰; 白框中, 角落处深度更加准确, 并且变化趋势更加明显. 由此我们可以认为对编码器的特征进行自适应融合后通过跳层连接到解码器可以提高深度图的精度、减少误差.

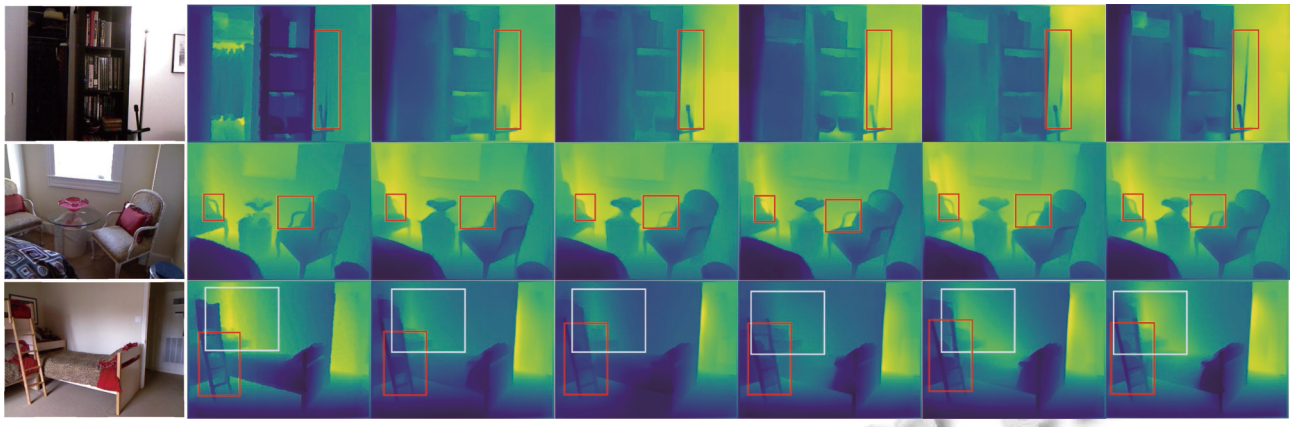


图 5 消融实验深度图对比

在仅增加 ASPP 模块的情况下, 由图 5 可知, 在第 1 场景下的灯杆的预测结果更加准确, 在第 2 场景下的左侧红框中的预测结果相比未使用 ASPP 模块的模型对椅子把手下的空白区域预测更加准确, 但是椅子把手的预测仍存在不连续的问题. 在此基础上, 使用 ASPP 模块并融合浅层特征信息之后取得最好效果, 由表 2 可以看出在 δ_1、*RMSE*、*AbsRel*、$RMSE_{\log 10}$ 指标上取得最好的结果, 并且提升了小物体以及细长物体的预测, 比如在第 1 场景中的灯杆、第 2 场景中的椅子把手和第 3 场景的梯子的预测结果中均比之前要清晰, 确保椅子把手的连续性, 并且对第 3 场景中墙角处的深度预测相比之前更为准确. 由此可以表明使用 ASPP 模块融合浅层特征信息可以减少空洞卷积下像素信息的丢失, 提升对场景中物体细节的预测精度.

4 结论

本文在编码器-解码器结构的网络模型基础上, 提出一种充分利用编码器阶段不同尺度特征的方法, 对不同尺度特征进行自适应融合并连接到解码器端, 实现解码端对编码端的信息合理且高效的利用, 并减少下采样过程中细粒度的深度信息的丢失对密集预测任务的影响. 此外在使用 ASPP 模块对深层特征进行整体深度信息的提取时, 融合浅层特征信息, 在保证整体深度信息准确的基础上, 优化细节深度信息.未来的研究方向可以建立更轻量级的网络模型, 在精度可以接受范围内提高模型预测速度, 继续推进单目图像深度估计任务的研究.

参考文献

1 Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 2366–2374.
2 Laina I, Rupprecht C, Belagiannis V, *et al*. Deeper depth prediction with fully convolutional residual networks. Proceedings of the 4th International Conference on 3D Vision (3DV). Stanford: IEEE, 2016. 239–248.
3 Hu JJ, Ozay M, Zhang Y, *et al*. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa: IEEE, 2019. 1043–1051.
4 Chen XT, Chen XJ, Zha ZJ. Structure-aware residual pyramid network for monocular depth estimation. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao: IJCAI.org, 2019. 694–700.
5 Song M, Lim S, Kim W. Monocular depth estimation using laplacian pyramid-based depth residuals. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(11): 4381–4393. [doi: 10.1109/TCSVT.2021.3049869]
6 Zhang AM, Ma YC, Liu JY, *et al*. Promoting monocular depth estimation by multi-scale residual Laplacian pyramid fusion. IEEE Signal Processing Letters, 2023, 30: 205–209. [doi: 10.1109/LSP.2023.3251921]
7 Miangoleh SMH, Dille S, Mai L, *et al*. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 9680–9689.
8 Ranftl R, Bochkovskiy A, Koltun V. Vision Transformers for

dense prediction. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal: IEEE, 2021. 12159–12168.

9 Cao YZH, Wu ZF, Shen CH. Estimating depth from monocular images as classification using deep fully convolutional residual networks. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(11): 3174–3182. [doi: 10.1109/TCSVT.2017.2740321]

10 Yuan WH, Gu XD, Dai ZZ, *et al*. Neural window fully-connected CRFs for monocular depth estimation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans: IEEE, 2022. 3906–3915.

11 Wu JW, Zhou WJ, Luo T, *et al*, Multiscale multilevel context and multimodal fusion for RGB-D salient object detection. Signal Processing, 2021, 178: 107766.

12 Hu J, Shen L, Albanie S, *et al*. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011–2023. [doi: 10.1109/TPAMI.2019.2913372]

13 Yang X, Chang QL, Liu XL, *et al*. Monocular depth estimation based on multi-scale depth map fusion. IEEE Access, 2021, 9: 67696–67705. [doi: 10.1109/ACCESS.2021.3076346]

14 Xu HH, Li F. An efficient monocular depth prediction network using coordinate attention and feature fusion. Journal of Information Processing Systems, 2022, 18(6): 794–802.

15 Xu Y, Yu Q. Adaptive weighted multi-level fusion of multi-scale features: A new approach to pedestrian detection. Future Internet, 2021, 13(2): 38. [doi: 10.3390/fi13020038]

16 Lee JH, Han MK, Ko DW, *et al*. From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv:1907.10326, 2021.

17 Wu KW, Zhang SR, Xie Z. Monocular depth prediction with residual DenseASPP network. IEEE Access, 2020, 8: 129899–129910. [doi: 10.1109/ACCESS.2020.3006704]

18 廖志伟, 金兢, 张超凡, 等. 基于分层压缩激励的 ASPP 网络单目深度估计. 图学学报, 2022, 43(2): 214–222.

19 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami: IEEE, 2009. 248–255.

20 Liu ST, Huang D, Wang YH. Learning spatial fusion for single-shot object detection. arXiv:1911.09516, 2019.

21 Chen LC, Papandreou G, Kokkinos I, *et al*. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834–848. [doi: 10.1109/TPAMI.2017.2699184]

22 Alhashim I, Wonka P. High quality monocular depth estimation via transfer learning. arXiv:1812.11941, 2018.

23 Wang Z, Bovik AC, Sheikh HR, *et al*. Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing, 2004, 13(4): 600–612. [doi: 10.1109/TIP.2003.819861]

24 Hao ZX, Li Y, You SD, *et al*. Detail preserving depth estimation from a single image using attention guided networks. Proceedings of the 2018 International Conference on 3D Vision (3DV). Verona: IEEE, 2018. 304–313.

25 Fu H, Gong MM, Wang CH, *et al*. Deep ordinal regression network for monocular depth estimation. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City: IEEE, 2018. 2002–2011.

26 Xu XF, Chen Z, Yin FL. Monocular depth estimation with multi-scale feature fusion. IEEE Signal Processing Letters, 2021, 28: 678–682. [doi: 10.1109/LSP.2021.3067498]

(校对责编: 张重毅)