基于采样汇集网络的场景深度估计

谢 昭 [1] 马海龙 [1] 吴克伟 [1] 高 扬 [1] 孙永宣 [1]

摘 要 针对现有场景深度估计方法中, 由于下采样操作引起的复杂物体边界定位不准确, 而造成物体边界处的场景深度估计模糊的问题, 受密集网络中特征汇集过程的启发, 本文提出一种针对上/下采样过程的汇集网络模型. 在下采样过程中, 使用尺度特征汇集策略, 兼顾不同尺寸物体的估计; 在上采样过程中, 使用上采样反卷积恢复图像分辨率; 同时, 引入采样跨层汇集策略, 提供下采样过程中保存的物体边界的有效定位信息. 本文提出的采样汇集网络 (Sampling aggregate network, SAN) 中使用的尺度特征汇集和采样跨层汇集, 都可以有效缩短特征图到输出损失之间的路径, 从而有利于避免模型的参数优化时陷入局部最优解. 在公认场景深度估计 NYU-Depth-v2 数据集上的实验说明, 本文方法能够有效改善复杂物体边界等干扰情况下的场景深度估计效果, 并在深度估计误差和准确性上, 优于当前场景深度估计的主流方法.

关键词 采样汇集网络, 场景深度估计, 尺度特征汇集, 上采样

引用格式 谢昭, 马海龙, 吴克伟, 高扬, 孙永宣. 基于采样汇集网络的场景深度估计. 自动化学报, 2020, **46**(3): 600−612

DOI 10.16383/j.aas.c180430

Sampling Aggregate Network for Scene Depth Estimation

XIE Zhao[1] MA Hai-Long[1] WU Ke-Wei[1] GAO Yang[1] SUN Yong-Xuan[1]

Abstract State-of-the-art approaches for scene depth estimation are built on downsampling strategy, which can lead to inaccurate location and ambiguous depth estimation for complicated boundary. Inspired with feature aggregation in DenseNets, we propose a novel feature aggregation strategy for upsample/downsample in our sampling aggregate network (SAN). Firstly, scale feature aggregation is used in downsample process to consider various scale object boundaries. Secondly, transposed convolution is applied in upsample process to restore image resolution. Thirdly, sample skip connection and aggregation is devoted to extract effective location of object boundary from downsample module with the same resolution. We adopt scale feature aggregation and sample skip aggregation to shorten the path from feature map to output loss, in order to avoid local optimal solution of our sampling aggregate network. Experiments in the recognized NYU-Depth-v2 database of scene depth estimation show that our model can improve the depth estimation result under complecated object boundaries and other disturbances. Our sampling aggregate network outperforms the state-of-the-art methods in error and accuracy evaluations.

Key words Sampling aggregate network (SAN), scene depth estimation, scale feature aggregate, upsampling

Citation Xie Zhao, Ma Hai-Long, Wu Ke-Wei, Gao Yang, Sun Yong-Xuan. Sampling aggregate network for scene depth estimation. *Acta Automatica Sinica*, 2020, **46**(3): 600−612

单目图像的场景深度估计, 关注于如何从单目图像中获得场景深度信息. 在 Marr 奠定的计算机视觉理论中, 将单目图像的场景深度估计作为人类视觉的一项重要任务. 场景深度信息, 对于许多其他任务提供了重要信息, 例如, 语义分割[1]、目标检测[2]、姿态估计[3]、3D 重建[4]、即时定位与地图构建[5]等. 随着深度传感器技术的成熟, 含有场景深度信息的 RGBD 数据集被构建, 拓展了单目图像的场景深度估计的研究领域. 但是, 由于在真实世界的不同场景中, 视觉信息含有大量的复杂干扰因素, 场景深度估计仍然是一个不明确的病态问题.

近年来, 单目图像的场景深度估计, 被视为场景深度值的连续回归问题, 其使用的基本假设是外观特征差异与场景深度的不连续性具有对应关系. 卷积神经网络由于具有准确地图像特征提取能力, 受到场景深度估计研究人员的广泛关注[6−10], 借助场景深度数据集, 卷积神经网络可有效实现场景深度模型的训练. 然而, 现有方法中仍然存在着以下几大挑战: 1) 场景深度恢复任务需要像素级的预测结果, 卷积神经网络下采样过程会丢弃部分图像像素, 从而导致场景深度估计精度不足; 2) 随着卷积神经网络模型深度的增加, 梯度退化现象严重, 造成场景深度估计模型学习能力降低; 3) 卷积神经网

收稿日期 2018-06-15 录用日期 2019-02-13
Manuscript received June 15, 2018; accepted February 13, 2019
国家自然科学基金 (61503111, 61273237) 资助
Supported by National Natural Science Foundation of China (61503111, 61273237)
本文责任编委 吴毅红
Recommended by Associate Editor WU Yi-Hong
1. 合肥工业大学计算机与信息学院 合肥 230601
1. School of Computer and Information, Hefei University of Technology, Hefei 230601

络中跨层方式和特征组合方式的多样性, 造成场景深度估计模型的复杂性和预测精度之间难以平衡.

针对现有场景深度估计方法中, 由于下采样操作引起的复杂物体边界定位不准确, 而造成物体边界处的场景深度估计模糊的问题; 受密集神经网络中特征汇集过程的启发[11], 本文提出一种针对上/下采样过程的汇集神经网络模型. 首先, 模型使用层次卷积和下采样策略描述图像中不同层次物体的基本结构; 其次, 采用反卷积和上采样策略, 恢复场景深度分辨率, 避免卷积神经网络对图像分辨率的损失. 最终, 针对采样神经网络训练过程中的梯度退化问题, 通过分析上/下采样过程中物体边缘保持的对应关系, 引入相同尺度采样约束下的跨层连接, 实现高精度的场景深度估计. 本文主要贡献如下:

1) 通过分析下采样分辨率损失对复杂边界精度估计的影响, 引入相同尺度采样约束下的跨层连接, 并使用上采样反卷积过程逐层还原图像分辨率, 提出一种采样汇集网络 (Sampling aggregate network, SAN) 模型.

2) 使用尺度特征汇集策略, 兼顾不同尺寸物体的深度估计; 同时, 受密集神经网络中特征汇集过程的启发, 尺度特征汇集和采样跨层汇集一样, 也有效缩短了特征图到输出层的路径, 避免了模型梯度过小陷入局部最优解.

3) 通过分析不同尺度采样下的场景深度估计结果, 确定深度卷积神经网络的最佳层次结构, 在 NYU-Depth-v2 场景深度公认数据集中, 本文提出采样汇集网络模型, 能够提供更准确的场景深度估计结果.

1 场景深度估计现状

解决单目场景深度估计问题过程中, 利用的基本线索是物体的外观特征, 除此以外, 场景几何、物体语义、运动、3D 位置和方向都可以实现对场景深度的约束. Su 等[12] 对场景深度的外观模式, 使用自然场景统计获得局部深度模式字典, 构建多变量高斯混合似然模型估计场景深度. Liu 等[13] 同时分析语义分割和场景几何约束对深度估计的影响. Karsch 等[14] 采用非参数采样方法, 使用局部运动和光流保持时间约束上的场景深度一致性. Saxena 等[4] 在马尔科夫随机场 (Markov random field, MRF) 框架下分析超像素的 3D 位置和 3D 方向对场景深度重建的影响. 但是, 上述模型存在两个主要问题, 1) 忽略了场景中内容之间的深度相互约束关系. 2) 手工特征在描述复杂外观模式上的局限性.

针对深度相互约束关系, 条件随机场 (Conditional random field, CRF) 模型具有统一深度特征和上下文深度约束的建模能力, 具体来说包括层内建模和层次间的建模. Batra 等[15] 使用 Laplacian 形式定义 CRF 层内中团势函数, 并使用最大边界模型对 CRF 参数进行求解. Saxena 等[16]针对非结构化室外场景, 构建层次化的多尺度 MRF, 实现全局和局部场景深度的融合. 上述模型解决了深度约束, 但是受到一元函数求解精度的限制, 因此, 研究的主要方向转向深度学习模型及其在深度学习模型基础上构建的图模型.

卷积神经网络用于提高外观特征建模的准确性, Eigen 等[17] 使用深度神经网络, 分别对局部和全局场景深度建模, 实现尺度不变的场景深度估计. 在卷积神经网络基础上, Roy 等[6] 使用随机森林构建层次化的场景深度估计模型. Fu 等[7] 提出回归分类级联网络, 同时预测低分辨率和高分辨率的场景深度. 在卷积神经网络的场景深度描述能力基础上, CRF 模型进一步对场景深度的局部不一致性进行优化, 包括对多尺度 CRF 建模和求解, 以及不同线索下的二元约束问题. 在多尺度 CRF 建模和求解方面, Liu 等[18] 将单目深度估计问题, 定义为离散 – 连续优化的 CRF 问题, 对超像素进行连续编码, 对超像素之间的关系进行离散编码, 使用粒子置信度传播算法来推理求解. Liu 等在超像素基础上, 使用卷积神经网络提取场景深度特征, 并构建像素池化的 CRF 模型[6, 19]. Xu 等[20] 构建深度序列卷积神经网络模型, 并将卷积后的多尺度输出, 构建层次化的 CRF 模型实现场景深度估计. 此外, 场景全局布局和表面法向量约束, 可以用于构建 CRF 二元约束. Zhuo 等[21] 使用场景全局结构, 将场景内容分层, 使用 CRF 对多层次的场景深度进行编码和推理. Wang 等[8] 在全局布局指导下, 将图像分解为局部区域, 以卷积神经网络为基础构建层次 CRF 模型, 进行场景深度和语义预测. Yan 等[22] 使用 CRF 模型添加物体的表面法向量的约束, 对超像素级别和像素级别的多层次场景深度估计. 可以看出卷积神经网络对一元函数的建模提高了场景深度建模的准确性, 然而, 深度学习模型自身的演化, 必将带动场景深度估计的再次突破.

随着卷积神经网络模型深度的增加, 存在严重的梯度退化问题, 该现象被场景深度估计研究者关注. Cao 等[10] 首先将场景深度进行离散化, 并将场景深度估计视为分类任务, 使用残差神经网络求解. Laina 等[23] 对全卷积残差网络, 采用多尺度上卷积和上投影策略实现重叠特征映射. 此外, 左右视差一致性[24] 和场景深度的空间上下文[25], 同样被用

于残差网络, 以解决局部深度不一致性问题. 与残差网络模型思想一致, 汇集网络[11] 也通过特征汇集策略, 使得特征图与输出损失之间路径变短从而避免模型陷入局部最优解. Sharma 等[26] 对预训练的 Denseblock 模型进行反卷积处理, 同时考虑使用均方根误差 (Root mean square error, RMSE) 和 berHu 两种损失项, 重新设计深度估计损失函数. Zhu 等[27] 在上采样反卷积过程中, 使用 denseblock 模块并尝试引入同尺度跨层特征共享策略, 应用于像素级的图像光流估计任务, 在其上采样过程中使用特征累积, 并没有考虑特征的冗余性. 通过分析发现, 上述模型并没有关注采样过程中场景深度误差的产生的原因, 尤其是上采样过程中同尺度特征共享和冗余是否会干扰场景深度估计的损失. 现有方法结果中存在物体边界处的场景深度值出现模糊的情况, 造成这种情况的主要原因是下采样操作引起的复杂物体边界定位不准确. 为了解决这些问题, 受到密集神经网络中特征汇集过程的启发, 本文提出一种针对上/下采样过程的汇集神经网络模型.

2 采样汇集网络

2.1 采样汇集网络模型结构

针对现有深度神经网络模型不能解决卷积下采样引起的场景深度估计损失问题, 图 1 给出了本文提出的采样汇集网络 (SAN) 模型. 本文的主要创新点包括 3 个方面: 1) 引入反卷积上采样模块 (图 1 中 US (Up sampling) 模块), 实现对场景深度分辨率的恢复; 2) 基于特征汇集思想, 对相同尺度的场景深度估计引入跨层误差传递, 图 1 中灰色虚线, 通过缩短误差计算的路径, 提高模型的收敛精度. 本文提出的采样汇集网络模型, 通过上述的上采样策略和采样跨层误差传递, 从而实现场景深度估计精度的提高; 3) 在汇集网络模块内部 (图 1 中 AB (Aggregate block) 模块), 使用尺度特征汇集策略, 进一步缩短特征图到输出损失的路径, 有利于模型的参数优化.

基于本文提出采样汇集网络 (SAN) 模型, 场景深度估计问题可以描述为对场景深度值的回归估计, 即通过学习 RGB 特征和场景深度值之间的映射关系, 并使用深度模型学习具有层次化的局部结构特征, 从而实现场景深度值的回归估计. SAN(x, w) 是测试时, 本文使用的深度网络模型, x 其中是输入图像, w 是采样汇集网络中每层中的参数集合. 为了学习本文模型中的参数, 在训练过程中, 本文模型的目标函数 $\Omega(x, w)$ 可以定义为

$$\Omega(x, w) = \|\text{SAN}(x, w) - y_{gt}\|_2^2 + \lambda\|w\|_2^2 \quad (1)$$

其中, y_{gt} 是真实测量的场景深度值, 采用逐像素方式比较, 并采用 2 范数的平方描述预测值和真实值之间的损失. λ 是回归模型的正则化参数, 以保证采样汇集网络中参数尽可能小, 避免过拟合现象. 在本文模型的预处理模块, 模型中将图像的 RGB 通道分离, 并使用 3D 卷积层 (图 1 中 CL (Convolutional layer 模块) 对其进行特征预处理, 可以记作 $z_1 = f(x, w_1)$, 其中, w_1 是第 1 块网络的滤波器参数, 其中使用的 3D 滤波器的尺寸为 3 × 3 × 3, 使用 64 个 3D 滤波器 ($m = 64$), 获得预处理模块的 64 层特征, 每层特征图与原始图像大小一致.

根据图 1 的说明, 本文的采样汇集网络从输入到输出共包括 13 个模块, 即, 1 个预处理卷积层模块, 5 个基于局部汇集网络的下采样模块, 1 个局部汇集网络转换模块, 5 个基于局部汇集网络的上采样模块, 和 1 个线性回归模块. 最后一层的回归模块中, 使用 1 × 1 卷积模型, 等价实现线性回归单元 (图 1 中 LR (Linear regression) 模块), 获得场景深度估计. 通过优化求解整个网络滤波器的权重, 恢复出场景的深度信息.

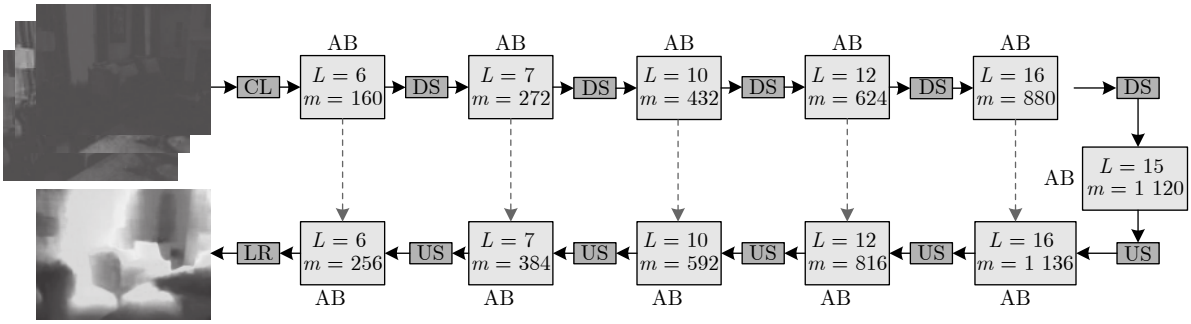


图 1 基于采样汇集网络的场景深度估计

Fig. 1 Sampling aggregate network for scene depth estimation

2.2　尺度特征汇集的下采样网络

本文模型使用下采样 (Down sampling, DS) 的主要原因是, 1) 使用下采样可以降低图像分辨率, 在较小的分辨率中, 每个像素对应到原始图像中的感受野较大, 这样可以描述更大尺度上的场景深度的分布; 2) 使用下采样可以降低图像分辨率, 同时降低了图像滤波过程的计算代价. 但是, 图像下采样过程的负面作用是, 在重建和原始图像相同分辨率的场景深度时, 产生了预测精度上的损失.

图 2 进一步给出了本文采样汇集网络模型的下采样网络结构, 其中每次下采样过程, 包括一次局部汇集网络和一次下采样网络. 每个局部汇集网络中包含了若干的卷积层, 图 2 中给出了 2 个不同深度的局部汇集网络, 其中每一个 CL 矩形是一个 3D 卷积层. 本文通过特征通道的汇集操作实现特征前向的跨层传递, 以便误差反向传播时, 能够进行跨层形式的传递. 本文模型中第 2 块到第 6 块为包含局部汇集网络的下采样模块. 每个局部汇集网络 (AB) 的参数可以记为 $w_i = [w_{i,1}, w_{i,2}, w_{i,l_i}, w_{i,d}]$, 其中 l_i 是当前第 i 块局部汇集网络 (AB) 中具有的 3D 卷积层的数量, $w_{i,d}$ 是下采样操作过程中使用的 1×1 滤波器参数 (图 2 中 DS 模块). 局部汇集网络 (AB) 和下采样网络 (DS) 的前向推理过程可以记为

$$z_i = ds\left(f([z_{i-1}, f^1(z_{i-1}, w_i), f^2(z_{i-1}, w_i), \cdots, f^{l_i}(z_{i-1}, w_i)], w_{i,d})\right) \tag{2}$$

其中, $ds(\cdot)$表示下采样过程, z_i 表示第 i 块局部汇集网络的特征图, z_{i-1}表示第 $i-1$ 块局部汇集网络的特征图, $f^1(z_{i-1}, w_i)$ 表示对输入的第一次 3D 卷积网络的特征输出, 由于局部汇集网络各卷积层采用串行级联方式前向推理, 因此, 每经过一个 3D 卷积层就叠加一次卷积过程, 到该局部汇集网络的最后一层时, 共经历 l_i 层卷积层, 所以记作 $f^{l_i}(z_{i-1}, w_i)$.

如果不考虑下采样 $ds(\cdot)$ 和其中 $w_{i,d}$ 卷积过程, 而单独考虑每个 3D 卷积层中的滤波器前向计算过程, 我们可以将式 (2) 中的卷积过程记为

$$\begin{aligned} zt_i = &[f^1(z_{i-1}, w_i), f^2(z_{i-1}, w_i), \cdots, f^{l_i}(z_{i-1}, w_i)] = \\ &[f^1(z_{i-1}, w_{i,1}), f(f(z_{i-1}, w_{i,1}), w_{i,2}), \cdots, \\ &f(\cdots(f(z_{i-1}, w_{i,1}), w_{i,2}), \cdots, w_{i,l_i})] \end{aligned} \tag{3}$$

其中, $f^{l_i}(z_{i-1}, w_i)$ 包含了第 i 块局部汇集网络中每一层的滤波器卷积过程. 图 2 中下方给出各特征通道汇集过程的示意图, 局部汇集网络中每个卷积层输出的特征汇集到一起 (即图 2 中的圆形节点), 并与前一层输入的特征汇集. 在局部汇集网络中, 为了保证每个 3D 卷积层的输出都为 16 层特征, 在局部汇集网络的特征输入时, 预先采用 3D 卷积处理转化为 16 层的特征宽度.

根据上述特征汇聚过程, 可以推理出每层特征通道中包含的特征来源和特征图数量, 例如, 第 1 次下采样过程中, 使用的局部汇集网络的参数为 $L = 6$, $m = 160$, 其代表的含义为局部汇集网络有 6 个 3D 卷积层 (图 2 中 CL 矩形), 每个 3D 卷积层输出 16 层特征, 同时浅层网络中输入特征层为 $m = 64$, 因此, 第 1 次下采样过程的输出特征层数为 $16 \times 6 + 64 = 160$. 随后的下采样网络中, 继续执行一次 2×2 的最大池化下采样, 得到长宽各为原始图像分辨率一半的图像, 继续前向传递计算.

2.3　采样汇集跨层的上采样网络

在下采样过程中, 随着网络深度的增加, 特征图数量在增加, 但是特征图的空间分辨率随之下降. 为了恢复空间分辨率, 本文模型中引入了上采样反卷积操作, 并引入跨层连接, 组成上采样路径. 每个上采样模块与下采样模块一一对应, 每个上采样模块包括局部汇集网络和上采样网络. 由于上采样过程中, 引入了反卷积滤波器, 因此, 其参数形式与下采样模块不同, 上采样网络的模型参数可以记作 $w_j = [w_{j,1}, w_{j,2}, \cdots, w_{j,l_i}, w_{j,u}]$, 其中 $w_{j,u}$ 为上采样反卷积滤波器的参数.

图 3 给出了上采样反卷积的执行过程, 包含

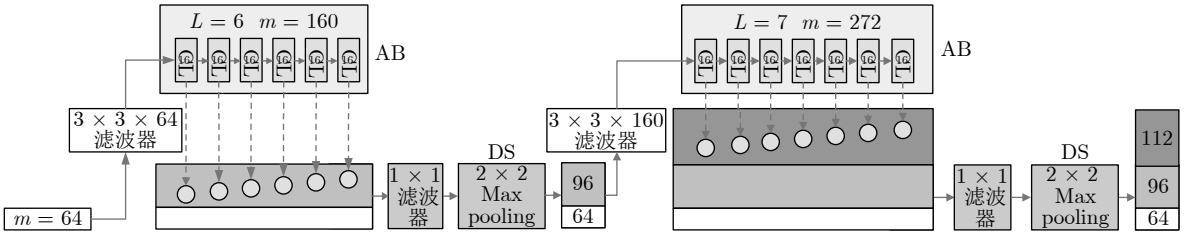


图 2　尺度特征汇集的下采样网络

Fig. 2　Downsampling network with scale feature aggregation

2 个主要步骤: 1) 进行空间分辨率 2 倍的上采样, 并将新增的像素初始化为 0; 2) 对 2 倍上采样的图像进行 3 × 3 滤波器卷积, 并保持图像的分辨率不变 (如图 3 所示), 从而实现对 0 像素位置场景深度的重新估计. 由于本文模型采用多层的特征图, 在不同的特征图上使用各自的 3 × 3 滤波器参数独立前向推理, 从图 3 中可以看出不同滤波器具有不同的边缘效应, 反卷积过程会将滤波器自身包含的边缘信息添加到上采样输出中, 从而实现分辨率细节的恢复. 图 4 描述了采样汇集跨层 (图 1 中向下虚线) 的网络模型结构, 其中采样汇集跨层是指从相同分辨率的下采样模块到上采样模型的特征图传递 (图 4中向下虚线), 并与前向传递的上采样反卷积特征图进行特征汇集, 从而产生后续的特征图.

本文模型的上采样网络模块中, 对于低分辨率的特征图, 先进行一次上采样反卷积, 随后执行一次局部汇集处理. 由于下采样过程中的分辨率损失, 仍然受到上采样反卷积滤波器参数的局限, 因此, 根据图像分辨率的对应关系, 将同分辨率的卷积特征图进行关联, 引入采样同层跨层约束, 使用所有可用的特征来参与上采样计算. 本文模型中第 8 块到第 12 块为包含局部汇集网络的上采样模块, 上采样网络的前向推理过程可以记为

$$z_j = f(us([zt_{14-j}, f^1(z_{j-1}, w_j), f^2(z_{j-1}, w_j), \cdots, f^{l_j}(z_{j-1}, w_j)]), w_{j.u}) \tag{4}$$

其中, zt_{14-j} 表示与上采样第 j 块对应的下采样模块, 从图 1 中可知, 其模块编号为 $14-j$. 注意到为了避免特征层数无限增加, 因此, 在上采样过程中, 其特征通道仅保留对应的下采样特征图, 以及该块局部汇集网络自己产生的特征图. 上采样模块中的内部 3D 卷积层数与下采样过程一一对应, 根据图 4 可以看出本文模型特征随着上采样过程的进行, 特征图数量逐步减少.

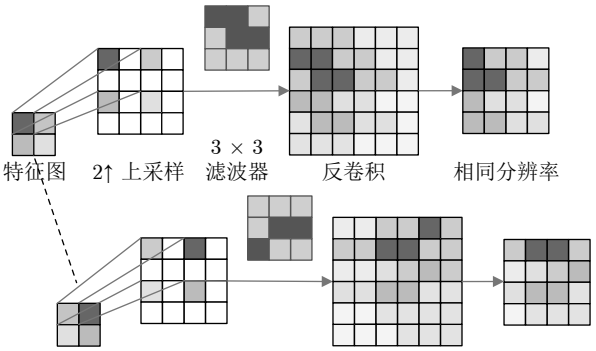


图 3 上采样反卷积过程

Fig. 3 Upsampling network with deconvolution process

2.4 采样汇集网络的参数学习

图 1 中给出了本文模型的基本参数设置, 本文模型包含预处理模块 (1 层), 包含局部汇集网络的下采样模块 (56 层, 其中每次下采样后进行一次 1 × 1 卷积), 转换模块 (15 层), 上采样模块 (56 层, 其中每次反卷积算作一次卷积层) 和线性回归模块 (1 层), 共计 129 层卷积神经网络.

为了避免每一层的数据分布不同, 在每个局部汇集块前使用批规一化 (Batch normalization, BN) 进行预处理, 随后使用 ReLu 激活函数, 3 × 3 滤波器模块, 进行无分辨率损失的滤波操作. 每个下采样模块采用批规范化, 进行预处理, 使用 ReLu 激活函数和 1 × 1 滤波器, 采用 2 倍的最大池化下采样方式降低分辨率. 每个上采样模块, 采用 3 ×

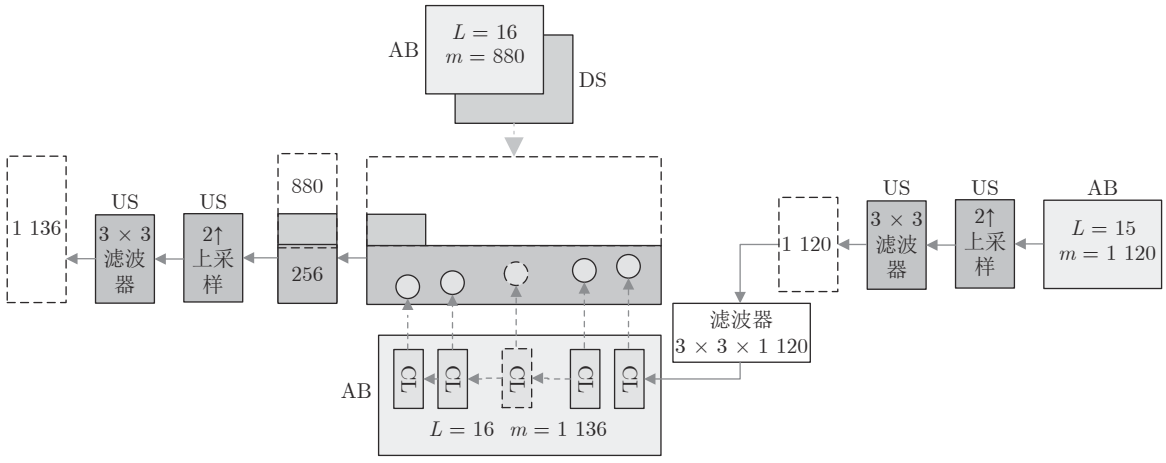


图 4 采样汇集跨层的上采样网络

Fig. 4 Upsampling network with sample skip aggregation

3 滤波器进行反卷积. 本文模型中最后一个模块是线性回归模块, 采用 1 × 1 滤波器实现, 输入 $m = 256$ 层的特征图, 进行场景深度数值的线性回归.

本文模型的目标函数如式 (1) 所示, 根据图 1, $\mathrm{SAN}(x, w)$ 的场景深度预测值就是模型的第 13 块的输出, 即 $\mathrm{SAN}(x, w) = z_{13}$. 式 (2) 和式 (4) 分别给出了下采样和上采样的前向计算过程, 用于模型的场景深度值预测过程. 参数学习的执行过程可以记为

$$w^* = \arg\min_w(\|\mathrm{SAN}(x, w) - y_{gt}\|_2^2 + \lambda\|w\|_2^2) \quad (5)$$

本文模型使用 Torch[28] 深度学习开源平台训练网络. 实验工作站配置为 CoreX i7-6800k 6 核 3.4 GHz CPU, 2 块 NIVDIA GTX1080 8 GB 显卡. 本文模型不使用任何预训练模型, 而是对所有层的参数重新训练, 本文模型的参数初始化采用 He-Uniformed 形式[29], 参数优化过程使用随机梯度下降方法. 训练过程的批处理大小为 4, 每循环一次训练集合的所有图像, 作为一轮迭代, 模型训练的最大迭代次数设置为 30. 参数学习率为 0.01, 每迭代 5 次降低 20 %. 迭代过程中, 权重衰减系数为 10^{-4}, 权重衰减用于模型正则化.

3　实验

3.1　实验设置

本文模型使用纽约大学构建的 NYU-Depth-v2 数据集进行模型的训练和测试[30], 数据库包含 1 449 幅不同类型的室内场景的 RGBD 图像, 该数据集是场景深度估计公认的大型数据集之一. 其中 depth 图像使用 Microsoft kinect 设备采集获得, 场景深度的数值从 0 米到 10 米. 实验随机选择 795 幅图片作为训练图像, 其余的 654 幅图片作为测试图像. 并对 795 幅训练图像进行扩充, 具体操作为, 根据随机条件对原始训练图像进行变换, 最终产生 48 k 合成 RGBD 图像对用于模型训练. 随机条件包括: 1) 尺度缩放, 尺度缩放因子的取值范围为 [1,1.5]; 2) 旋转变换, 旋转角度的取值范围为 [–5, 5]; 3) 颜色变换, 对图像的亮度, 饱和度和对比度, 分别进行线性变换, 线性变换因子的取值范围为 [0.6,1.4]; 4) 图像左右翻转, 左右翻转的随机概率为 0.5. 在训练和测试过程中图像采用相同的分辨率, 为了分析输入图像分辨率对实验结果的影响, 在网络层次结构不变的情况下, 采用 2 种不同尺寸图像分辨率 304 × 228, 152 × 114 进行实验分析.

本文对比方法包括传统的字典学习模型, 结构化深度模型, 深度 CRF 模型以及残差深度模型. 具体来说: 1) Su 等[12] 使用局部模式字典估计场景深度模式. 2) 在深度模型结构化方面, Roy 等[6] 使用随机森林构建层次化深度模型; Fu 等[7] 使用回归级联形式的深度模型. 3) 在深度 CRF 方面, Wang 等[9] 在卷积神经网络基础上, 构建层次 CRF 模型; Liu 等[10] 在超像素基础上, 构建卷积池化 CRF 模型. 4) 在残差网络方面, Laina 等[23] 使用残差网络, 构建多尺度上卷积和上投影模型; Cao 等[10] 使用残差网络, 并将场景深度问题视为分类任务建模, Sharma 等[26] 使用带 denseblock 结构的反卷积网络实现深度估计. 我们通过与上述模型对比, 来分析本文方法中采用的上采样策略和尺度采样约束的功能. 本文使用评价标准[9] 具体包括: 1) 平均相对深度 (Average relative error, REL), 即预测深度与真实深度的差值的绝对值与真实深度的比值. 2) 根均方误差 (Root mean squared error, RMS), 即预测深度与真实深度的均方根误差. 3) 对数误差 ($\log_{10}$), 即对预测深度与真实深度进行 $\log_{10}$ 处理后, 计算像素上的两者之间的平均差值. 4) 阈值精度 (δ), 根据 $\max(\mathrm{SAN}(x, w^*)/y_gt, y_gt/\ \mathrm{SAN}(x, w^*))$ 求出比值误差, 并与阈值比较, 如果比值误差小于阈值 δ, 则认为深度数值预测正确, 本文实验中阈值参数设置为 $\delta_1 < 1.25$, $\delta_2 < 1.25^2$, $\delta_3 < 1.25^3$.

3.2　定量分析

3.2.1　采样汇集网络的消融分析

本文的消融因素包含两个, 即采样汇集和尺度特征汇集. 为了验证本文采样汇集的有效性, 采用如下的方式进行消融分析: 1) 对本文模型中的采样汇集跨层进行删除, 保留尺度特征汇集过程, 将该消融模型称为尺度特征汇集网络 (Scale feature aggregate network, SFAN); 2) 对本模型中的尺度特征汇集过程删除, 即去除前层输入的特征, 保留下采样对上采样的汇集过程, 将该消融模型称为窄采样网络 (Narrow sampling network, NSN). 表 1 给出了图像分辨率为 304 × 228 情况下, 采样汇集网络的消融分析的定量结果, 图 5 给出了该情况下的采样汇集网络的消融分析的定性结果.

从表 1 中可以看出, 本文模型中两个消融因素在多种评价指标中都具有明显作用. 从图 5 定性结果中可以看出, 采样汇集网络 (图 5(e)), 在尺度和采样像素上的准确性高于其他消融模型. 1) 采样汇集比尺度特征汇集策略, 对整个描述的正确性影响更大, 以平均相对深度 (REL) 来说, 采样汇集可以提供 0.007 的贡献, 而尺度特征汇集提供 0.005 的贡献. 2) 观察图 5(c) 第 2 行中, 预测的物体轮廓模糊的情况, 可以说明采样汇集的作用主要在于下采

样的跨层, 可以利用下采样前高分辨率的特征, 来保持物体外围边界位置的准确, 避免产生局部最优解, 使得物体轮廓模糊. 3) 观察图 5(d) 第 1 行中, 预测的场景背景中散落的杂乱信息情况, 可以说明尺度特征的作用主要在于描述不同感受野大小的观测, 以保证预测在不同尺度上的一致性, 避免产生局部最优解, 使得出现琐碎伪物体的估计结果. 通过消融模型分析发现, 本文模型通过考虑尺度特征和采样特征在场景估计中各自的优势, 设计出新的深度模型结构, 实现了更鲁棒的场景深度估计.

3.2.2 采样汇集网络的感受野范围分析

通过消融模型分析发现尺度特征的感受野大小对场景估计中有明显的影响. 场景深度估计任务中采用下采样的主要意义在于, 下采样能够产生不同尺度的特征图, 其中各像素对应原始图像中的感受野大小不同, 从而发现不同尺度下的场景深度模式, 但是, 同时也注意到下采样降低了空间分辨率, 可能导致更大的误差, 而且随着下采样次数的增加, 网络模型加深, 整体网络模型的参数增加, 训练难度和测试时间都会增加, 从模型的计算成本出发, 因此, 需要讨论网络结构中的下采样次数.

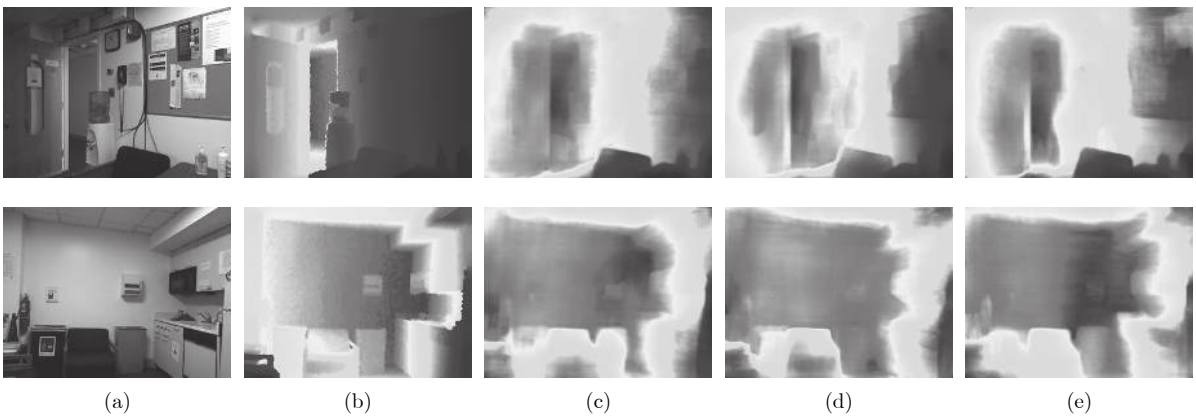
根据图 1 所示, 本文采用 5 块汇集网络和下采样网络, 针对本文模型设计变形模型, 即分别采用 1 层下采样, 即本文模型只使用第 1 次下采样后直接进行转换模块和对应的一次上采样, 此时模型为 31 层, 记作 SAN-31. 同理, 我们分别讨论不同层次的下采样次数和对应的变形模型, 分别记作 SAN-47, SAN-69, SAN-95. 表 2, 给出了图像分辨率 304×228 情况下, 不同下采样次数下的模型变形定理分析, 图 6 给出了该情况下, 不同下采样次数下的定性结果.

通过表 2 和图 6 的下采样次数分析发现: 1) 下采样次数越多, 感受野范围变化越多, 场景深度估计的准确性越准确; 2) 下采样次数进一步增多, 所带来的场景深度估计的贡献逐渐减小, 可以理解为场景中的主要物体尺寸集中在中小物体尺寸, 进一步增加下采样次数带来的增益有限; 3) 下采样次数过多带来的计算成本和储存成本提高. 在实验工作环境下, SAN-95 的测试单幅图像的平均运行时间为 0.06 s, 而 SAN-129 的测试单幅图像的平均运行时间为 0.11 s, 同时, 由于显存大小的限制, 难以训练更大深度的 SAN 模型; 4) 从图 6 中可以看出, 小尺寸感受野的情况下, 出现了大量的杂乱估计, 这

表 1 采样汇集网络的消融分析

Table 1 Ablation analysis of sampling aggregate network

消融模型	Error			Accuracy (%)		
	REL	$\log_{10}$	RMS	δ_1	δ_2	δ_3
SFAN-129	0.165	0.072	0.586	75.70	93.70	98.10
NSN-129	0.163	0.070	0.583	76.00	94.10	98.20
SAN-129	0.158	0.067	0.567	77.60	95.20	98.80



(a) (b) (c) (d) (e)

图 5 采样汇集网络的消融模型对比实例图. (a) 原始图像; (b) 真实场景深度; (c) SFAN-129 结果; (d) NSN-129 结果; (e) SAN-129 结果

Fig. 5 Contrasting examples of ablation models for sampling aggregate network. (a) RGB image; (b) GT depth; (c) result of SFAN-129; (d) result of NSN-129; (e) result of SAN-129

是因为小尺寸对边缘敏感, 但是不对物体级别的区域敏感; 5) 只有在下采样尺寸达到 3 以上, 才能出现与真实场景物体分布相似的估计. 因此, 结合下采样次数的定量和定性分析, 以及工作环境和成本的限制, 本文模型最后确定下采样次数为 5 次.

3.2.3　采样汇集网络的输入图像分辨率分析

表 3 给出了不同图像分辨率情况下采样汇集网络的定量分析结果. 本文方法目的在于重构出于输入图像分辨率相同的场景深度图像, 其中, 不同的输入图像分辨率, 会改变模型中每层特征图的分辨率, 也会影响模型参数规模, 从而影响模型最终的参数学习结果. 实验采用 2 种不同尺寸的分辨率图像, 讨论该参数对结果的影响. 对 304 × 228 的训练和测试图像, 采用间隔为 2 的下采样方法获得对应 152 × 114 的训练和测试图像. 通过表 3 中的实验结果可知, 在相同的模型结构和参数学习条件下, 使用缩小后图像训练的模型, 会在各项预测指标上都有所提高. 这是因为较大图像分辨率中包括较多的局部细节结构, 这些相对精细结构需要更复杂优化算法找出模型中的卷积参数. 而对于使用较小分辨率图像训练的情况, 由于较小图像分辨率已经丢弃了部分局部细节结构, 可以认为局部细节结构的复杂度有所降低, 模型中的卷积参数已经能够有效描述存在的局部模式, 从而在指标上有所提高. 但是, 较小分辨率预测的缺点是, 对于具有深度范围变化的物体, 会无法准确提取物体较大分辨率上的深度值. 因此, 本文同时给出 304 × 228 和 152 × 114 分辨率下的预测结果.

3.2.4　对比方法

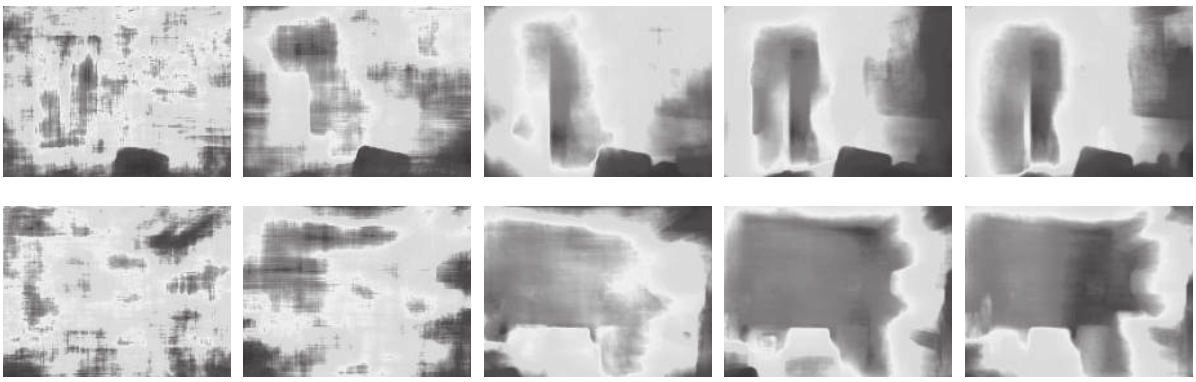
本文训练过程采用整幅图像像素级的监督信息, 对各像素的场景深度值进行回归处理, 这种处理的有效性在于: 1) 与 Fu 等[7] 方法离散化的像素深度值预测相比, 本文模型可以获得连续性的深度预测值, 避免场景中出现相邻像素深度值的阶梯效应; 2) 与使用预分割的区域标记方法[9, 12] 相比, 本文模型直接使用端对端的方式分析边缘两侧深度的连续性, 从而避免使用预分割过程中存在的误分割标记.

表 4 展示了现有主流对比方法, 对比方法包括局部模式字典, 深度模型优化策略, 深度 CRF 模

表 2　采样汇集网络中下采样次数定量分析

Table 2　Quantitative analysis of downsampling times in sampling aggregate network

采样汇集网络模型	Error			Accuracy (%)		
	REL	$\log_{10}$	RMS	δ_1	δ_2	δ_3
SAN-31	0.311	0.129	1.012	46.20	77.10	91.80
SAN-47	0.250	0.107	0.830	55.70	84.90	95.50
SAN-69	0.194	0.083	0.672	68.00	90.70	97.70
SAN-95	0.169	0.073	0.608	73.60	93.10	98.30
SAN-129	0.158	0.067	0.567	77.60	95.20	98.80



(a)　(b)　(c)　(d)　(e)

图 6　采样汇集网络中下采样次数定性结果, 图 6 的原图和真实场景深度与图 5 中对应. (a) SAN-31 结果; (b) SAN-47 结果; (c) SAN-69 结果; (d) SAN-95 结果; (e) SAN-129 结果

Fig. 6　Qualitative results of downsampling times in sampling aggregate network. Fig. 6 and Fig. 5 have the same RGB images and GT depth images. (a) result of SAN-31; (b) result of SAN-47; (c) result of SAN-69; (d) result of SAN-95; (e) result of SAN-129

型, 深度残差网络. 从表 4 中的实验发现: 1) 本文模型比局部模式字典的方法[12], 准确性有显著提高, 这主要归功于深度学习特征对自然场景中复杂边缘结构的捕获能力; 2) 建立在深度学习基础上的级联优化过程[6−7], 有助于准确性提高, 本文模型使用多尺度下采样和采样汇集策略, 提供了更准确的结果; 3) CRF 模型有助于琐碎区域的平滑[8−9], 本文模型对于琐碎区域的解决思路是, 考虑较大尺度的感受野, 以保证小物体区域在不同感受野下深度估计的一致性, 从而提高了准确性; 4) 残差网络模型[10, 23]通过减少特征图到输出层的路径长度, 可以有效地改善参数优化过程, 避免模型中各层特征图梯度过小, 陷入局部最优的情况. 本文模型使用多尺度特征汇集和采样跨层汇集策略, 实现了不同特征层次到输出损失层的更短的路径, 从而提高了模型训练效果. 本文模型在所有 Error 评价上优于所有方法, 在 Accuracy 评价上, 本文模型在小误差范围内能得到更好的效果.

基于表 4 的实验对比, 本文方法部分指标上也有不足之处. 1) Fu 等[7] 方法在大范围精度的评价标准上略高于本文 SAN 方法, 但是, 首先这种误差已经接近于 1.95 倍 (1.25^3 约等于 1.95), 在实际场景深度应用中会产生较多的后续错误, 此外, 大范围精度的准确性达到 99 %, 反映出的是场景整体分布的范围, 而不反映局部区域的特性, 因此, 对于小物体的估计参考价值有限. 2) Sharma 等[26] 方法在部分指标上超出本文方法, 这主要是因为 Sharma 等使用了两种不同形式的数据损失项, RMSE 损失是一种均方根形式的损失, berHu 是一种分段函数, 原始误差在较小数值范围内是线性变换, 在较大数值范围内是平方形式变换. 因此, 我们看出 Sharma-RMSE 损失项的方法, 在 RMS 指标上能够到达所有方法中最好的结果, 此外, Sharma-berHu 在 $\log_{10}$ 和 RMS 上超出本文方法. 通过对比 Sharma 等方法和本文 SAN-129 在 304 像素 × 228 像素分辨率上的表现, 说明 Sharma 等使用的损失项有利于部分指标, 但并不能兼顾所有指标的提高. 3) 同时, 我们注意到 Sharma 等方法[26] 使用的图像分辨率为 175 像素 × 127 像素, 与本文方法的 304 像素 × 228 像素不同, 为了进一步分析图像分辨率对结果的影响, 我们对 304 × 228 图像进行间隔 2 下采用获得训练集合, 重新训练模型. 进一步观察本文方法 SAN-129 在 152 像素 × 114 像素分辨率上的实验效果, 仍然能够发现 Sharma等使用的损失项在 $\log_{10}$ 和 RMS 指标上是有优势的, 但是本文方法在 Accuracy 指标上均超过 Sharma 等方法, 同时还获

表 3 采样汇集网络中输入图像分辨率定量分析

Table 3 Quantitative analysis of image resolution in sampling aggregate network

采样汇集网络模型	图像分辨率	Error			Accuracy (%)		
		REL	$\log_{10}$	RMS	δ_1	δ_2	δ_3
SAN-129	304 × 228	0.158	0.067	0.567	77.60	95.20	98.80
SAN-129	152 × 114	0.149	0.064	0.562	79.95	95.23	98.80

表 4 本文采样汇集网络与现有方法定量对比

Table 4 Quantitative analysis of our sampling aggregate network with state-of-the-art methods

对比方法	Error			Accuracy (%)		
	REL	$\log_{10}$	RMS	δ_1	δ_2	δ_3
Su等[12]	0.302	0.128	0.937	–	–	–
Laina等[23]	0.215	0.083	0.790	62.90	88.90	97.10
Liu等[9]	0.213	0.087	0.759	65.00	90.60	97.60
Wang等[8]	0.210	0.094	0.745	60.50	89.00	97.00
Roy等[6]	0.187	0.078	0.744	–	–	–
Cao等[10]	0.187	0.071	0.681	71.20	92.30	98.00
Fu等[7]	0.160	–	0.586	76.50	95.00	99.10
Sharma-RMSE[26]	0.159	0.064	0.549	79.10	94.60	98.40
SAN-129 @ 304 × 228	0.158	0.067	0.567	77.60	95.20	98.80
Sharma-berHu[26]	0.153	0.062	0.549	79.90	95.00	98.50
SAN-129 @ 152 × 114	0.149	0.064	0.562	79.95	95.23	98.80

得了 REL 指标的最好值. 综上所述, 本文模型由于使用了采样汇集和尺度特征汇集策略, 改善了神经网络结构, 缩短了特征图到输出层的路径, 从而实现了更准确了场景深度估计定量结果.

3.2.5　困难实例的定性分析

场景深度估计中存在几个主要挑战是小物体干扰、复杂边界干扰、光照干扰、深度范围干扰, 因此, 本文进一步给出上述挑战情况下的困难实例的实验结果, 以说明本文方法处理的鲁棒性. 1) 小物体的主要困难在于需要从背景中区分出物体 (图 7 第 1 行), 并避免小物体的深度被周围信息干扰, 对比图 7(c) 和图 7(e) 第 1 行可以看出, 在小尺度上小物体 (桌子) 具有一定的显著性, 但是, 随着采样尺度的增加, 场景中小物体可以更好地与周围的环境分离, 对比图 7(d) 和图 7(e) 第 1 行可以看出, 由于删除了采样汇集跨层, 图 7(d) 第 1 行图中, 小物体周围的边界较模糊, 而本文方法中桌子的轮廓较为清晰. 2) 复杂边界是指物体轮廓的形状较复杂, 而且具有的场景深度跨度较大 (图 7 第 2 行). 不同尺度下的特征图所对应的原始图像感受野不同, 如果去除特定层次的感受野, 会导致物体整体淹没在背景深度中 (图 7(c) 第 2 行), 同样, 去除采样跨层后 (图 7(d) 第 2 行), 其中的复杂边界和小物体其轮廓都更为模糊, 这主要是因为下采样中丢失了物体边界的准确位置, 造成场景深度难以恢复. 3) 光照干扰是指场景中由于存在干扰, 造成局部外观与周围外观的突变 (图 7 第 3 行). 如果丢失大尺度的特征

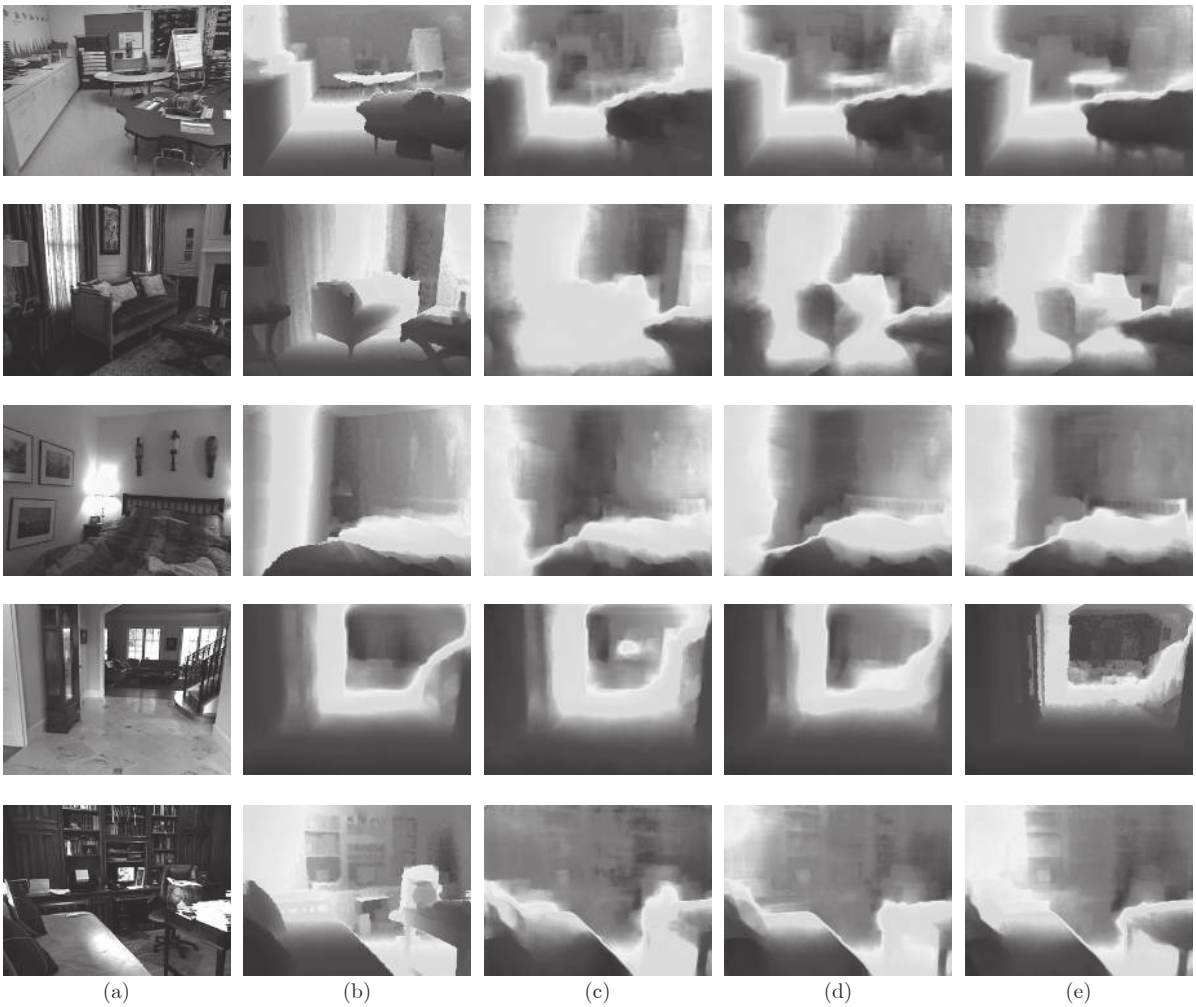


图 7　场景估计中的困难实例, 第 1 行小物体干扰, 第 2 行复杂边界干扰, 第 3 行光照干扰, 第 4 行深度范围大干扰, 第 5 行深度范围小的干扰. (a) 原始图像; (b) 真实场景深度; (c) SAN-95 结果; (d) SFAN-129 结果; (e) SAN-129 结果

Fig. 7　Challenge examples in depth estimation, including small object interference (Line 1), complex boundary interference (line 2), illumination interference (line 3), large depth range interference (line 4), small depth range interference (line 5). (a) RGB image; (b) GT depth; (c) result of SAN-95; (d) result of SFAN-129; (e) result of SAN-129

图, 在捕获场景的较大的边缘时就会更为困难, 会产生光源区域的错误估计 (图 7(c) 第 3 行). 尺度特征汇集策略 (图 7(d) 第 3 行) 其精度明显不如本文的模型 (图 7(e) 第 3 行) 的原因, 主要是环境光照干扰下进一步加剧了下采样过程对边界定位的误差. 4) 深度范围干扰是指本文模型需要兼顾处理场景深度变化大的图像, 也要兼顾处理场景深度变化小的图像. 图 7(a) 第 4 行场景的深度范围, 大于图 7(a) 第 5 行场景的深度范围, 但是基于本文方法的处理策略, 产生了明显的深度估计的改善, 使用尺度特征汇集避免了图 7(c) 第 4 行中的孔洞, 使用采样汇集提高了图 7(e) 第 4 行和第 5 行中的边界准确性. 从而说明本文方法在上述各种干扰中, 都实现了可靠的场景深度估计.

从上述定性实验中可以看出, 现有场景深度估计任务中的主要挑战在于细节信息的预测, 即精确的物体轮廓, 本文方法使用的采样汇集网络, 其成功主要在于: 1) 利用深度学习的层次卷积方式挖掘复杂局部结构, 这种结构既能反映出物体轮廓, 同时物体轮廓也能用于深度估计, 因为物体轮廓暗示了区域之间的深度不连续性; 2) 本文方法在汇集网络的基础上, 进一步讨论了引入相同尺度采样约束下的跨层连接的特征汇集, 和去除同尺度特征冗余在特征选择中的作用, 即通过不同尺度特征的分析, 找出哪些特征对于场景深度不连续性是有效的. 由于本文关注于场景深度估计, 因此其学习出的特征主要反映的是深度模式; 3) 由于精确物体轮廓也是图像分割中关注的重点, 需要进一步分析本文方法与图像分割任务之间的关系. 本文方法同样可以使用于有分割标记的训练过程, 这是因为两个任务有相同点, 都需要学习 RGB 局部结构, 都是像素级标记预测, 在 RGB 结构学习中面对的噪声干扰是相似的. 但是, 本文任务与图像分割也具有明显的不同点, 具体来说: 1) 深度估计是连续标记回归, 而分割是离散标记分类, 场景回归任务需要更高的精度; 2) 监督信号不同, 分割标记和深度值不是一一对应的, 同一个物体可以具有不同的场景深度, RGB 局部结构特征对不同任务的有效性不同, 学习过程中对 RGB 局部结构特征的选择不同; 3) 虽然, 分割中不同物体之间的轮廓可以暗示深度值的差异, 但是是否真的有差异, 以及差异程度仍然需要进一步学习.

4 结论

针对现有基于深度卷积模型的场景深度估计方法中, 由于采样分辨率损失, 引起的物体边界估计不足的问题, 受密集网络中的特征汇集策略启发, 本文提出一种针对上/下采样过程的汇集神经网络模型. 通过方法分析和实现分析可以证明: 1) 通过采样汇集跨层和上采样卷积策略, 提供了更准确的物体轮廓精度估计; 2) 通过尺度特征汇集, 有效地避免了小尺寸物体容易引起的杂乱场景深度现象; 3) 受密集神经网络中特征汇集过程的启发, 尺度特征汇集和采样汇集跨层都缩短了特征图到输出层的路径, 从而有利于本文模型的参数优化和准确性提高. 在公认的场景深度 NYU-Depth-v2 数据库实验结果中, 说明本文方法达到并在部分指标上超过了现有主流方法在深度估计误差和精度上的执行效果, 并通过对小物体干扰、复杂边界干扰、光照干扰、深度范围干扰的定性实现分析, 说明本文方法在处理实际问题真实可靠.

References

1 Jégou S, Drozdzal M, Vazquez D, Romero A, Bengio Y. The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA: IEEE, 2017. 1175–1183

2 Cheng Y H, Zhao X, Huang K Q, Tan T N. Semi-supervised learning and feature evaluation for rgb-d object recognition. *Computer Vision and Image Understanding*, 2015, **139**: 149–160

3 Borghi G, Venturelli M, Vezzani R, Cucchiara R. POSEidon: face-from-depth for driver pose estimation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA: IEEE, 2017. 5494–5503

4 Saxena A, Sun M, Ng A Y. Make3d: learning 3d scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(5): 824–840

5 Tateno K, Tombari F, Laina I, Navab N. CNN-SLAM: real-time dense monocular SLAM with learned depth prediction. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA: IEEE, 2017. 6565–6574

6 Roy A, Todorovic S. Monocular depth estimation using neural regression forest. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA: IEEE, 2016. 5506–5514

7 Fu H, Gong M M, Wang C H, Tao D. A compromise principle

in deep monocular depth estimation. arXiv preprint arXiv: 1708.08267, 2017. 1−11

8 Wang P, Shen X H, Lin Z, Cohen S, Price B, Yuille A. Towards unified depth and semantic prediction from a single image. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA: IEEE, 2015. 2800−2809

9 Liu F Y, Shen C H, Lin G S, Reid I. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, **38**(10): 2024−2039

10 Cao Y, Wu Z, Shen C H. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017, **28**(11): 1−11

11 Huang G, Liu Z, Maaten L V D. Weinberger K Q. Densely connected convolutional networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA: IEEE, 2017. 2261−2269

12 Su C C, Cormack L K, Bovik A C. Bayesian depth estimation from monocular natural images. *Journal of Vision*, 2017, **17**(5): 22−22

13 Liu B Y, Gould S, Koller D. Single image depth estimation from predicted semantic labels. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA: IEEE, 2010. 1253−1260

14 Karsch K, Liu C, Kang S B. Depth transfer: depth extraction from videos using nonparametric sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014, **36**(11): 2144−2158

15 Batra D and Saxena A. Learning the right model: efficient max-margin learning in laplacian CRFS. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA: IEEE, 2012. 2136−2143

16 Saxena A, Chung S H, Ng A Y. Learning depth from single monocular images. In: Proceedings of the 2005 Advances in Neural Information Processing Systems, Vancouver, British Columbia, Canada: MIT Press, 2005. 1161−1168

17 Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. In: Proceedings of the 2014 Advances in Neural Information Processing Systems, Montreal, Quebec, Canada: MIT Press, 2014. 2366−2374

18 Liu M M, Salzmann M, He X M. Discrete-continuous depth estimation from a single image. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA: IEEE, 2014. 716−723

19 Liu F Y, Shen C H, Lin G S. Deep convolutional neural fields for depth estimation from a single image. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA: IEEE, 2015. 5162−5170

20 Xu D, Ricci E, Ouyang W L, Wang X G, Sebe N. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA: IEEE, 2017. 161−169

21 Zhuo W, Salzmann M, He X M, Liu M M. Indoor scene structure analysis for single image depth estimation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA: IEEE, 2015. 614−622

22 Yan H, Zhang S L, Zhang Y, Zhang L. Monocular depth estimation with guidance of surface normal map. *Neurocomputing*, 2017, **280**: 86−100

23 Laina I, Rupprecht C, Belagiannis V, Tombari F, Navab N. Deeper depth prediction with fully convolutional residual networks. In: Proceedings of the 4th International Conference on 3D Vision, Stanford, CA, USA: IEEE, 2016. 239−248

24 Godard C, Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA: IEEE, 2017. 6602−6611.

25 Grigorev A, Jiang F, Rho S, Sori W J, Liu S H, Sai S. Depth estimation from single monocular images using deep hybrid network. *Multimedia Tools and Applications*, 2017, **76**(18): 18585−18604

26 Sharma S, Padhy R P, Choudhury S K, Goswami N, Sa P K. DenseNet with pre-activated deconvolution for estimating depth map from single image. In: Proceeding of the 2017 British Machine Vision Conference, London, UK: BMVA, 2017. 1−12

27 Zhu Y, Newsam S. Densenet for dense flow. In: Proceedings of the 2017 IEEE International Conference on Image Processing, Beijing, China: IEEE, 2017. 790−794

28 Collobert R, Kavukcuoglu K, Farabet C. Torch7: a matlab-like environment for machine learning. In: Proceeding of the 2011 Advances in Neural Information Processing Systems, Granada,

Spain: Springer, 2011. 1−6

29 He K M, Zhang X Y, Ren S Q, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile: IEEE, 2015. 1026−1034

30 Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from rgbd images. In: Proceedings of the 2012 European Conference on Computer Vision, Florence, Italy: IEEE, 2012. 746−760

谢 昭 合肥工业大学计算机与信息学院副研究员. 2007 年于合肥工业大学获得博士学位. 主要研究方向为计算机视觉, 图像处理, 模式识别.
E-mail: xiezhao@hfut.edu.cn
(**XIE Zhao** Associate research fellow at Hefei University of Technology. He received his Ph.D. degree from Hefei University of Technology in 2007. His research interest covers computer vison, image processing, and pattern recognition.)

马海龙 合肥工业大学硕士研究生. 主要研究方向为计算机视觉, 图像处理, 模式识别.
E-mail: mhl_hfut@163.com
(**MA Hai-Long** Master student at Hefei University of Technology. His research interest covers computer vision, image processing, and pattern recognition.)

吴克伟 合肥工业大学计算机与信息学院副研究员. 2013 年于合肥工业大学获得博士学位. 主要研究方向为计算机视觉, 图像处理, 模式识别. 本文通信作者.
E-mail: wukewei@hfut.edu.cn
(**WU Ke-Wei** Associate research fellow at Hefei University of Technology. He received his Ph.D. degree from Hefei University of Technology in 2013. His research interest covers computer vison, image processing, and pattern recognition. Corresponding author of this paper.)

高 扬 高扬合肥工业大学硕士研究生. 主要研究方向为计算机视觉, 图像处理, 模式识别.
E-mail: alto1996@163.com
(**GAO Yang** Master student at Hefei University of Technology. His research interest covers computer vision, image processing, and pattern recognition.)

孙永宣 合肥工业大学计算机与信息学院讲师. 2013 年于合肥工业大学获得博士学位. 主要研究方向为计算机视觉, 图像处理, 模式识别.
E-mail: syx@hfut.edu.cn
(**SUN Yong-Xuan** Lectiurer at Hefei University of Technology. He received his Ph.D. degree from Hefei University of Technology in 2013. His research interest covers computer vison, image processing, and pattern recognition.)