基于多尺度特征混合注意力的连续帧深度估计

郑宇航[1],曹雏清[1,2]
1. 安徽工程大学 计算机与信息学院,安徽 芜湖 241000
2. 长三角哈特机器人产业技术研究院,安徽 芜湖 241000

摘 要:目的 估计获取拍摄物体到相机之间距离的深度信息是单目视觉 SLAM 中获取深度信息的方法,针对无监督单目深度估计算法出现精度不足以及误差较大的问题,提出基于多尺度特征融合的混合注意力机制的连续帧深度估计网络。**方法** 通过深度估计和位姿估计的两种编码器解码器结构分别得到深度信息和 6 自由度的位姿信息,深度信息和位姿信息进行图像重建与原图损失计算输出深度信息,深度估计解码器编码器结构构成 U 型网络,位姿估计网络和深度估计网络使用同一个编码器,通过位姿估计解码器输出位姿信息;在编码器中使用混合注意力机制 CBAM 网络结合 ResNet 网络提取四个不同尺度的特征图,为了提升估计的深度信息轮廓细节在提取的每个不同尺度的特征中再进行分配可学习权重系数提取局部和全局特征再和原始特征进行融合。**结果** 在 KITTI 数据集上进行训练同时进行误差以及精度评估,最后还进行了测试,与经典的 monodepth2 单目方法相比误差评估指标相对误差、均方根误差和对数均方根误差分别降低 0.034、0.129 和 0.002,自制测试图片证明了网络的泛化性。**结论** 使用混合注意力机制结合的 ResNet 网络提取多尺度特征,同时在提取的特征上进行多尺度特征融合提升了深度估计效果,改善了轮廓细节。

关键词:单目视觉;连续帧深度估计;混合注意力机制;多尺度特征融合
中图分类号:TP391.4 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2024.0004.013

Continuous Frame Depth Estimation Based on Multi-scale Feature Mixed Attention Mechanism

ZHENG Yuhang[1], CAO Chuqing[1,2]
1. School of Computer and Information, Anhui University of Engineering, Anhui Wuhu 241000, China
2. Yangtze River Delta HIT Robot Technology Research Institute, Anhui Wuhu 241000, China

Abstract: Objective Estimating the depth information to obtain the distance between the photographed object and the camera is the method to obtain the depth information in monocular vision SLAM. As unsupervised monocular depth estimation algorithms suffer from insufficient accuracy as well as large errors, a continuous frame depth estimation network based on a hybrid attention mechanism with multi-scale feature fusion was proposed. **Methods** Information on depth and 6 degrees of freedom of pose were obtained by two encoder-decoder structures for depth estimation and pose estimation, respectively. The depth information and the pose information were used for image reconstruction with the original image loss calculation to output the depth information. The decoder encoder structure for depth estimation formed a U-shaped

收稿日期:2023-03-20 **修回日期:**2023-05-18 **文章编号:**1672-058X(2024)04-0104-08
基金项目:国家自然科学基金面上项目(62073101);高校优秀青年人才支持计划项目(019YQQ023);安徽省教育厅科学研究重点项目(KJ2020A0364);国家重点研发计划“智能机器人”重点专项(2018YFB1308900).
作者简介:郑宇航(1997—),男,安徽亳州人,硕士研究生,从事单目深度估计研究.
通讯作者:曹雏清(1982—),高级工程师,博士,从事智能机器人视觉感知研究. Email: caochuqing@hitrobot.com.cn.
引用格式:郑宇航,曹雏清. 基于多尺度特征混合注意力的连续帧深度估计[J]. 重庆工商大学学报(自然科学版),2024,41(4):104—111.
ZHENG Yuhang, CAO Chuqing. Continuous frame depth estimation based on multi-scale feature mixed attention mechanism[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2024, 41(4): 104—111.

network, and the same encoder was used for both the pose estimation network and the depth estimation network, and the pose information was output through the pose estimation decoder. The feature maps at four different scales were extracted in the encoder using a hybrid attention mechanism CBAM network combined with a ResNet network. For the enhancement of the estimated depth information contour details, the extracted features of each different scale were then assigned learnable weight coefficients to extract local and global features and then fused with the original features. **Results** Evaluation of error and accuracy was performed on the KITTI dataset, and finally, testing was also performed. Compared with the classical monodepth2 monocular method, the relative error, root mean square error, and log root mean square error in the error evaluation metrics were reduced by 0.034, 0.129, and 0.002, respectively, and self-made test images demonstrated the generalizability of the network. **Conclusion** The multiscale features are extracted using a ResNet network combined with a hybrid attention mechanism, while multiscale feature fusion on the extracted features enhances the depth estimation and improves the contour details.

Keywords: monocular vision; continuous frame depth estimation; hybrid attention mechanism; multiscale feature fusion

1　引　言

虽然现在使用激光雷达和双目相机的深度估计技术已经较为成熟,但在特殊环境下难以进行工作,且深度估计技术复杂、难度大和过高的硬件要求等问题依然存在。而单目图像深度估计没有过高的硬件要求,具有更广泛的应用范围和场景,针对视觉 SLAM 中使用 RGB-D 相机成本高及有效测距范围小和使用双目相机深度信息误差大及计算复杂等缺点,提出使用单目相机来进行深度估计的视觉 SLAM。随着单目视觉 SLAM 技术的发展,单目深度估计成为视觉 SLAM 研究领域的热点之一,单目深度估计分为有监督和无监督深度估计。而无监督深度估计又分为双目图像深度估计(又称立体图像对)和连续帧深度估计。

随着单目图像深度估计问题的进一步研究,有监督网络被用于深度估计并取得了不错的结果[1]。EIGEN 等[2]首次使用全局粗略估计深度,局部细化估计深度的多尺度深度预测技术。接着,LONG 等[3]提出使用全卷机网络(FCN)用于语义分割并广泛用于图像预测任务,其中包括图像深度预测。随后,Lai 等[4]提出 DCPNet——一种密集连接金字塔网络,融合图像金字塔结构多个阶段的多尺度特征。为了充分利用不同尺度的特征,网络不仅可以进行相邻阶段之间的特征融合,还可以进行非相邻阶段之间的特征融合,实现更细化的特征学习。Choi 等[5]则采用拉普拉斯金字塔和局部平面引导技术来引导物体更复杂的边界,实现获得更复杂的深度信息,比如遮挡、边缘轮廓。这些方法都采取了基于全连接层的编码器解码器的方法。

随着深度学习在有监督深度估计上的发展,网络模型也在不断精准,对标签数据集的要求也越来越高,而标签数据集获取难度比较大,获取成本很高[6]。无监督(也称自监督)深度估计也发展开来,无监督深度估计分为立体图像深度估计和连续帧深度估计。在基于连续帧估算深度的方法中,Zhou 等[7]率先提出从无标签的视频序列中联合训练深度网络和相机姿态估计网络,给定场景的一个输入视图,利用网络估计的深度图,合成从不同相机位置看到的场景的新视图。Xie 等[8]率先提出 2D 视频到 3D 信息转换的深度离散化模型。Casser 等[9]引入几何结构,对场景和单个目标进行建模来估计单目视频深度信息。Pillai 等[10]提出超分辨率网络进行高质量的合成提升深度信息精度。Godard 等[11]在原来研究的基础上进行了改进并加入了连续帧的深度估计,他们考虑遮挡和动态目标的影响,接受一对彩色图像预测深度和单个 6-DoF 的相对位姿,使用 Auto-Masking 过滤出外观上从一帧到下一帧不变化的像素来排除动态物体。随后,戴仁月等[12]使用 SLAM 算法优化位姿估计,与原有深度信息估计网络结合进行图像重构推理深度估计信息;叶星余等[13]在基于自注意力的基础,使用生成式对抗网络应用于深度估计网络和位姿估计网络来改善深度和位姿估计信息。在深度估计和位姿估计改善方面,罗志斌等[14]使用拥有视差上采样、深度一致性约束和多种掩膜的优化 UnDEMoN 网络模型来改善整体的估计效果。前面基于连续帧估计深度的方法没有考虑对提取到的特征多尺度融合,忽略了全局和局部特征的利用信息。因此,在此基础上,使用基于混合注意力机制的连续帧深度估计进行多尺度特征融合的无监督深度估计方法。

2　网络模型框架

现介绍使用的 ResNet 残差网络[15]和混合注意 CBAM[16],然后介绍整体网络结构。

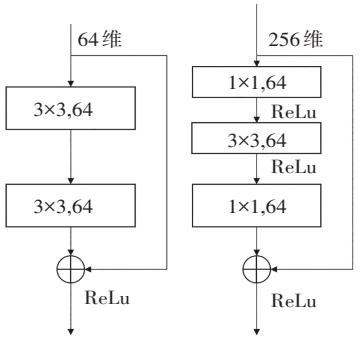
2.1　ResNet 网络

过去的卷积神经网络为了提取描述更准确和具有详细细节的特征会通过增加网络层的深度来提高提取

特征的能力,但这种做法会导致过拟合、梯度消失或者梯度爆炸等问题。过拟合指复杂的网络模型在训练数据表现优异,但是在测试数据上表现很差的现象。其实际原因是复杂的网络模型过度拟合了训练数据中的噪声和细节,从而导致学习过程中学习了大量的参数从而忽略了一些关键特征。这会使得整个神经网络的性能由最初的逐渐增强到达顶峰然后迅速下降,最终导致丧失一部分关键特征从而导致测试效果比较低。而梯度消失或者梯度爆炸更会使得神经网络训练缓慢甚至无法收敛。

为了解决神经网络层数加深出现的这些问题,He 提出了具有 short-connection 连接的 ResNet 网络,该网络可以将输入特征与输出特征相加,使得深层次的网络层也可以比较容易地训练捕捉到原始特征和提取特征之间的差异,避免了过拟合、梯度消失或者爆炸问题的出现。ResNet 网络在视觉领域的出现引起众多学者在基于原网络或者在原网络改变基础上用于处理大量的图像任务。如 Chen 等[17]使用残差网络进行语义分割任务。

如图 1 所示,ResNet 网络包含两种残差块:basicblock 和 bottleneck,图 1(a)展示了 basciblock,其包含两个卷积层和一个 shortcut 连接的简单残差块,适用于较浅的网络结构。而图 1(b)bottleneck 包含 3 个卷积层和一个 shortcut 连接的残差块,可以降低深层网络的计算量适用于更深的网络结构。Bottleneck 中的第一个卷积层用于降维,将输入的特征映射到一个较小的维度,第二个卷积层用于学习特征,第三个卷积层用于升维,将特征维度升到原始维度。通过组合不同数量和不同种类的残差块来构建 ResNet 网络,使用不同数量的 basicblock 残差块可以搭建 resnet-18 和 resnet-34,同样地,使用不同数量的 bottleneck 可以搭建 resnet-50、resnet-101 以及 resnet-152。残差块的引入使得神经网络可以添加更深层次的网络,在图像领域取得了很好的效果。



(a) basicblock (b) bottleneck

图 1 残差单元

Fig. 1 Residual unit

2.2 混合注意力机制 CBAM

混合注意力机制 CBAM 模型结构(Convolutional Block Attention Module, CBAM),它包括通道注意力模块 CAM 和空间注意力模块 SAM。CBAM 的模型结构如图 2 所示,它对输入的特征图,首先进行通道注意力模块处理;得到的结果,再经过空间注意力模块处理,最后得到调整后特征。

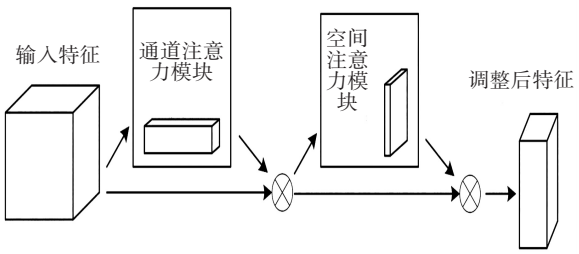


图 2 网络卷积注意力模块

Fig. 2 Network convolutional attention module

通道注意力模块 CAM 的输入是特征图,维度设为 $H\times W\times C$;其中 H 是指特征图的高度,W 是指宽度,C 是指通道数。它的思路流程是:首先,对输入的特征图,进行全局池化和平均池化(在空间维度进行池化,压缩空间尺寸;便于后面学习通道的特征);然后,将得到全局和评价池化的结果,送入到多层感知机中 MLP 学习(基于 MLP 学习通道维度的特征,和各个通道的重要性);最后,将 MLP 输出额结果,进行“加”操作,接着经过 Sigmoid 函数的映射处理,得到最终的“通道注意力值”。

空间注意力模块的输入是通道注意力模块输出的特征图,首先对输入的特征图,进行全局池化和平均池化(在通道维度进行池化,压缩通道大小,便于后面学习空间的特征);然后将全局池化和平均池化的结果,按照通道拼接,得到特征图维度是 $H\times W\times 2$;最后对拼接的结果进行卷积操作,得到特征图维度是 $H\times W\times 1$;接着通过激活函数处理。

2.3 整体网络模型

使用编码器与解码器结构来搭建连续帧深度估计网络框架,整体框架如图 3 所示。输入一组连续帧彩色图,通过源图像 I_T 输入深度估计网络推断当前帧的深度信息,同时将连续帧彩色图片(可以是当前帧与上一帧也可以是当前帧与下一帧)输入位姿估计网络进行位姿估计,接着将得到的深度信息和位姿信息进行当前帧图像重构,重构后的图像与当前帧通过图像相似性进行损失计算,反向传播用于更新各个参数,之后输出当前帧的深度信息。

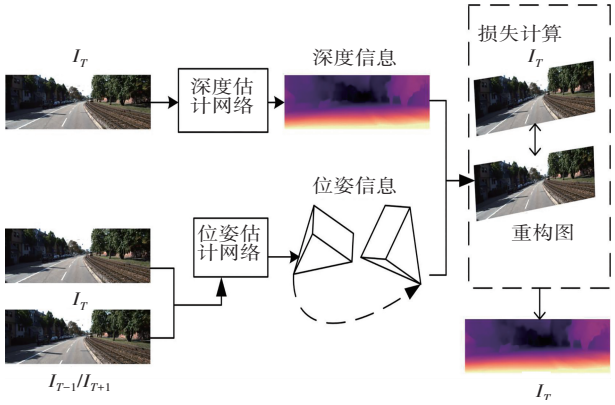


图 3　连续帧深度估计网络框架

Fig. 3　Network framework for continuous frame depth estimation

2.3.1　深度信息估计网络

在深度信息估计网络中,借助基于 Unet 网络模型[18]的概念构搭建编码器和解码器。编码器是提取特征部分,解码器是上采样部分。Unet 提出的初衷是为了解决医学图像分割的问题,是使用一种 U 型网络结构来获取上下文和位置的信息。Unet 使用全卷积网络,其主要特点是使用对称编码器和解码器,外观看起来是一个 U 型网络结构,因此称为 Unet。Unet 编码器部分使用卷积层和池化层提取输入图像的特征,解码器部分由反卷积层和上采样层组成,将图像的特征维度恢复到原始维度大小。在编码器和解码器之间使用多个跳跃连接来连接编码器和解码器相同维度的层数,通过各个尺度的特征融合获取图像更详细的特征信息。

使用 Unet 网络的好处是能够实现高精度的图像分割;Unet 网络模型采用了跳跃连接机制,可以利用多尺度特征图像进行信息提取,从而提高数据利用率;Unet 网络模型可以通过添加或删除卷积层和池化层来进行扩展或压缩,可以适应不同的应用场景,有比较强的通用性。

使用的编码器网络基于 resnet18 轻量级网络,提取关键信息能力不足,因此考虑使用混合注意力机制 CBAM 来提升网络性能。具体使用方法:将深度信息估计网络先加载预训练模型后,通过编码器中残差网络结合混合注意力机制 CBAM 作为编码器,具体网络框架如图 4 所示。在编码器模块中选用 ResNet18 层残差网络,同时 CBAM 中的通道注意力机制和空间注意力机制加入在残差网络的第一层。输入的 RGB 图片通过编码器分成 64、128、256 和 512 不同尺度的特征图后,在解码器中,相同尺度的特征图经过特征融合后,通过上采样成为上一层尺度的特征图,再与编码器相同尺度的特征度进行融合,这样的过程持续 3 次,即经过三层解码器块后输出深度信息。

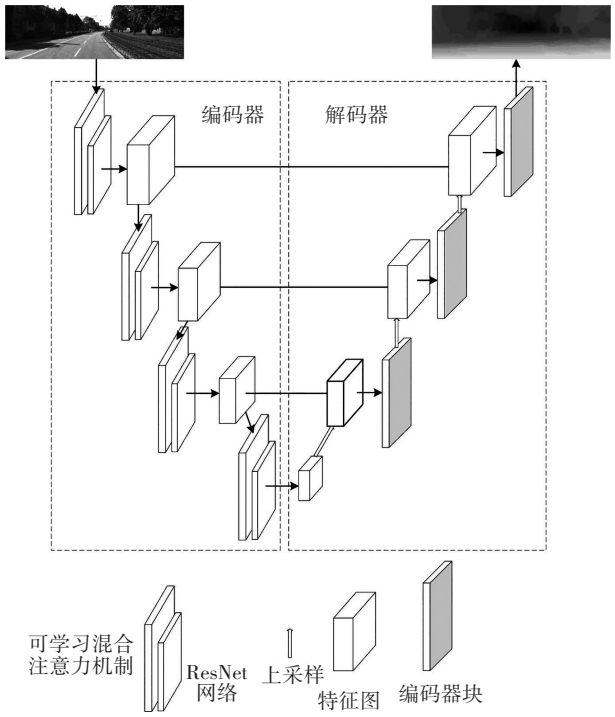


图 4　深度估计网络

Fig. 4　Depth estimation network

虽然 Unet 网络模型在每个尺度的特征上通过拼接的方式来做到特征融合,但依然会浪费提取的特征信息,因此后续的 Unet++[19]和 Unet3+[20]都是在对特征融合进行改善,但由于采用的轻量级编码器解码器结构不再加深网络层,因此在编码器提取阶段利用基于 pytorch 的可学习参数,将提取的特征进行全局特征提取、局部特征以及原始特征提取再根据权重自适应特征融合来利用提取的特征改善轮廓边缘信息,特征融合公式如式(1)所示:

$$\boldsymbol{F}_{\mathrm{ff}}=\alpha_1\boldsymbol{F}_{\mathrm{id}}+\alpha_2\boldsymbol{F}_{\max}+\alpha_3\boldsymbol{F}_{\mathrm{avg}} \tag{1}$$

$$\alpha_i=\frac{\mathrm{e}^{w_i}}{\sum_j \mathrm{e}^{w_j}}(i=1,2,3;j=1,2,3)$$

其中,α_i 为归一化权重,$\sum\alpha_i=1$,W_i 为初始化权重系数,本文设置为 1。$\boldsymbol{F}_{\mathrm{ff}}$ 为融合的特征矩阵;$\boldsymbol{F}_{\mathrm{id}}$ 是残差分支将提取到的特征图跳跃连接直接加过去保留的原始特征;$\boldsymbol{F}_{\max}$ 是进行最大池化提取的局部特征;$\boldsymbol{F}_{\mathrm{avg}}$ 是平均池化提取的局部特征。

2.3.2　位姿估计网络

如图 5 所示,位姿估计网络接收一对连续帧彩色图像作为输入,其中连续帧图像分为源视图和目标视图,重构图像为目标视图,和深度信息估计网络共用同一个编码器。将从解码器得到的特征图进行连续上采样,每个不同尺度的特征图之间采用跳跃连接和 relu

激活函数,上采样结束后恢复到原来的尺度,输出 6 自由度的位姿变换信息。

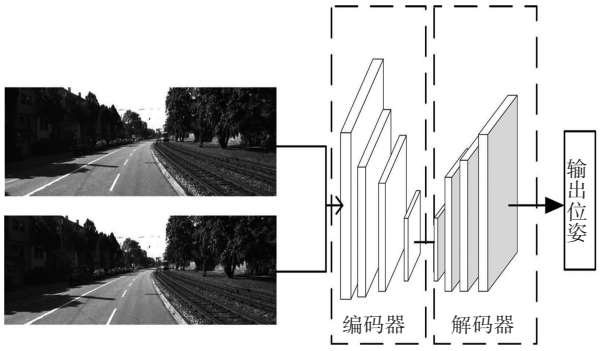


图 5 位姿估计网络

Fig. 5 Pose estimation network

3 损失函数设计

损失函数参考 monodepth2 的损失函数设计,首先是像素级最小重投影损失,在计算多个源图像的重投影误差时,常见的自监督深度估计方法将每个可用源图像的重投影误差平均在一起,会导致像素在部分图像上不可见。如果网络预测了这样一个不可见像素的正确深度,则源图像中的对应颜色很可能与目标不匹配,从而导致高光度误差惩罚。产生这样问题的像素可分为两类:一类是由于图像边界的自我运动而超出视野的像素,另一类是被遮挡的像素。传统方法是通过在重投影损失中掩盖这些像素来减少视域外像素的影响,但这不能处理去遮挡,其中平均重投影会导致模糊的深度不连续。Mondepth2 的解决方法是只使用最小值而不是平均所有源图像的光度误差,如式(2)所示:

$$L_p = \min_{t'} pe(I_t, I_{t'\to t}) \tag{2}$$

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

其中,$I_{t'\to t}$ 中的 t' 为源图像,t 为目标图像,$I_{t'\to t}$ 为利用相机位姿和深度对源图像进行采样的图像,$pe(\,)$ 函数如式(3),$proj(\,)$ 函数如式(4)所示:

$$pe(I_a, I_b) = \frac{\alpha}{2}(1-SSIM(I_a, I_b)) + (1-\alpha)\,\|I_a - I_b\|_1 \tag{3}$$

$$proj(D_t, \boldsymbol{T}_{t\to t'}, K) = \varphi(K[\boldsymbol{T}_{t\to t'} D_a(p_a) K^{-1}(h(p_a))]) \tag{4}$$

其中,P_a 为二维像素点坐标$(x,\ y)$,$h(p_a)$ 表示 p_a 的齐次坐标,K 为相机内参,$D_a(p_a)$ 为 p_a 处的深度,$\boldsymbol{T}_{t'\to t}$ 为旋转矩阵,α 是权重系数设置为 0.85,SSIM 是结构相似性,是一种衡量两幅图像相似度的指标,指标越大表示图像越相似,$I_{a,b}$ 表示图像上的一个像素点,其中 $\varphi([x,y,z]) = \left[\frac{x}{z}, \frac{y}{z}\right]$。

使用一种检测相机间有无运动的标记方法,剔除图像序列中未发生运动的像素点,如式(5)所示:

$$\mu = [\min_{t'} pe(I_t, I_{t'\to t}) < \min_{t'} pe(I_t, I_{t'})] \tag{5}$$

其中,μ 为掩码,属于二进制即 $\mu \in \{0,1\}$,通过对比目标图像 I_t 与源图像 $I_{t'}$ 之间的像素值差,并取差异的最小值作为阈值,假如上一个亮度守恒损失函数的值小于此阈值,则标记此像素点为存在运动的像素点,如不成立,则标记此像素点在图像序列之间静止,并将此像素点的损失值在整个损失函数中移除。最后一个损失函数视差平滑损失,计算公式如式(6)所示:

$$L_s = |\partial_x d_t^*|\, e^{-|\partial_x I_t|} + |\partial_y d_t^*|\, e^{-|\partial_y I_t|} \tag{6}$$

其中,∂_x、∂_y 表示视差图在 x,y 区域的梯度,将较低分辨率的深度图(从中间层)上采样到输入图像分辨率,在这个较高输入分辨率下重新投影、重新采样和计算误差 pe。最终损失函数如式(7)所示:

$$L = \mu L_p + \lambda L_s \tag{7}$$

4 实验结果与分析

4.1 实验数据集

KITTI 数据集是被广泛使用的数据集,由德国卡尔斯鲁厄理工学院和丰田汽车公司联合发布,在汽车驾驶场景下由激光雷达、灰度相机、彩色摄像头以及 GPS 惯性导航系统等传感器收集制作而成。数据集主要包括 22 个驾驶场景的图像序列,总共约 4 万张图像;包含 22 个驾驶场景的激光点云数据;包含 22 个驾驶场景的 3D 物体检测和跟踪数据;包含 22 个驾驶场景的姿态估计数据。其被广泛用于自动驾驶、目标检测和深度估计等领域的研究和应用。

4.2 数据增强

随着深度学习的发展,深度神经网络参数也越来越多,对数据集的容量要求也越来越高。但数据集的制作比较复杂,而深度学习中神经网络在学习中对于图像物体中一点微小的改变都会认为是一种新的特征,在此基础上就可以使用数据增强方式来扩张数据集到原来的几倍规模。实验采用基于 Pytorch 的图像预处理 transforms 模块进行数据增强,基于随机概率 50%的水平翻转、随机亮度、对比度、饱和度和色调抖动,调节范围分别为±0.2、±0.2、±0.2 和±0.1。

4.3 实验设置

为了验证优化算法,通过定量和定性的评估,将模型和现有的算法进行比较,实验在 Ubuntu18.04 系统环

境下搭配 V40 显卡,用 Pytorch 来构建神经网络。实验软件环境为 Python3.7;NVIDIA cuda11.1; cudnn11.1。

实验中,采用经过 KITTI 数据集数据增强的 39 810 张图片训练,4 424 张原始图片用于评估。在数据集训练时长约为 11 h,训练学习批次(batch size)的大小设置为 12;使用带有 Adam 优化器的单周期学习策略,Epoch 设置为 20,在前 15 个 epoch 中学习率的大小设置为 0.000 1,后面 5 个 epoch 学习率下降到 0.000 01。

4.4　评价指标

评估之前,将预测深度和深度真值进行尺度对齐,如式(8)所示:

$$\hat{D}=\lambda D,\lambda=\frac{median(D^*)}{median(D)} \tag{8}$$

其中,$\hat{D}$、D 和 D^* 分别表示对齐后的预测深度、初始预测深度和深度真值。通过以下指标来评价本文深度估计方法的效果:绝对相对误差 A_{AbsRel}、平方相对误差 S_{SqRel}、均方根误差 R_{RMSE}、对数均方根误差 L_{RMSElog} 及精确度 δ($\delta<1.25$、$\delta<1.25^2$、$\delta<1.25^3$)。具体定义如下:

$$A_{\mathrm{AbsRel}}=\frac{1}{|D^*|}\sum_{d\in\hat{D},d^*\in D^*}\frac{|d-d^*|}{d^*}$$

$$S_{\mathrm{SqRel}}=\frac{1}{|D^*|}\sum_{d\in\hat{D},d^*\in D^*}\frac{(d-d^*)^2}{d^*}$$

$$R_{\mathrm{RMSE}}=\sqrt{\frac{1}{|D^*|}\sum_{d\in D,d^*\in D^*}(d-d^*)^2}$$

$$L_{\mathrm{RMSELog}}=\sqrt{\frac{1}{|D^*|}\sum_{d\in\hat{D},d^*\in D^*}(\log d-\log d^*)^2}$$

$$\delta=\frac{\left|\left\{d\,|\,d\in\hat{D},\max\left(\frac{d^*}{d},\frac{d}{d^*}\right)<1.25^t\right\}\right|}{|D^*|}\times100\%$$

其中,$t\in(1,2,3)$。

如表 1 所示,上述所有算法都是在连续帧下的单目估计算法,其中算法-C 表示无 CBAM 注意力机制,-MF 表示没有对提取的特征进行最大池化以及平均池化后融合。通过误差评价指数(误差数据越小越好)和精度指数(精度数据越高越好)可以看出:本文与现有其他方法对比在精度和误差方面有明显优势,特别是在与传统算法 SFMlearner 对比,在精度上分别提高了 0.142、0.057 和 0.024。与 Luo 等[14]算法对比,在精度上也分别提高了 0.006、0.011 和 0.007。同时,算法在没有 CBAM 注意力机制下,误差分别上升 0.08、0.04、0.099 和 0.003,精度也下降 0.023 和 0.003。综上可见,算法在搭建 Unet 网络的基础上使用混合注意力机制和多尺度特征融合提升了总体算法性能。

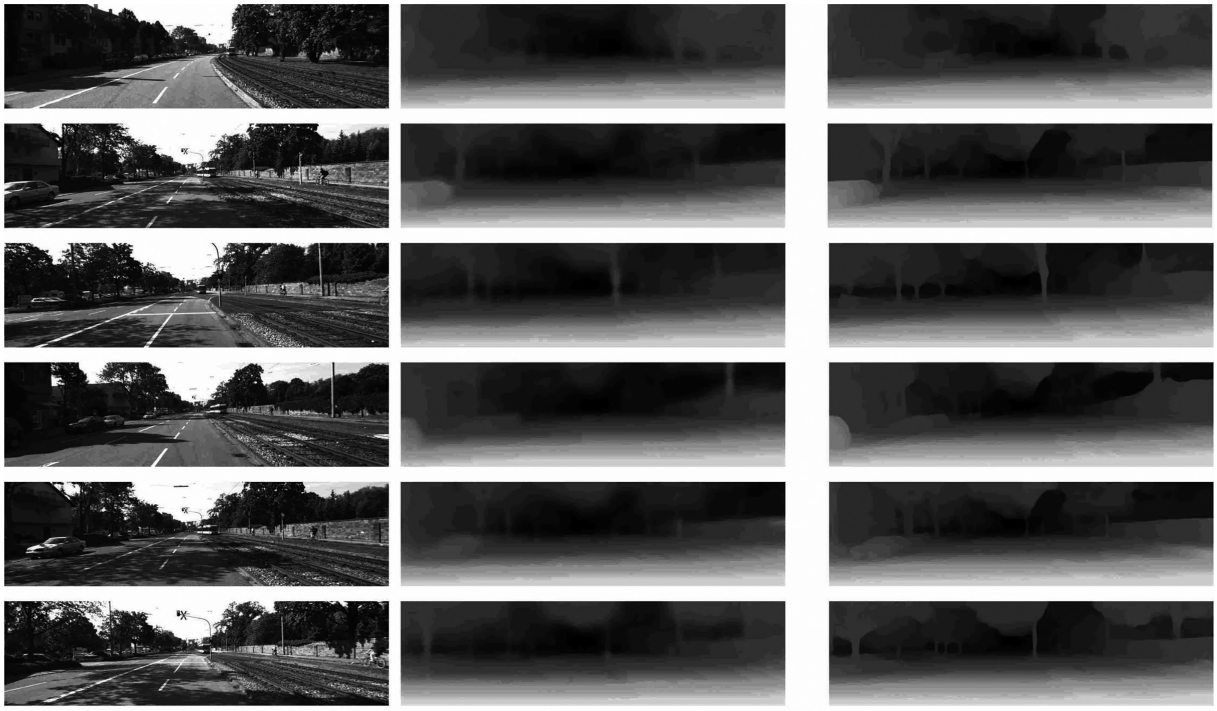
表 1　KITTI 数据集上的深度估计精度对比实验

Table 1　Comparative experiments on depth estimation accuracy on the KITTI dataset

方　法	AbsRel	SqRel	RMSE	RMSElog	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
SFMLearner[7]	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Struct2Depth[9]	0.141	1.036	5.294	0.215	0.816	0.945	0.979
SuperDepth[10]	0.112	0.875	4.958	0.207	0.852	0.947	0.977
MonoDepth2[11]	0.115	0.905	4.863	0.193	0.877	0.959	0.981
Da[12]	0.189	1.592	6.432	0.268	0.714	0.911	0.963
Ye[13]	0.151	1.200	5.560	0.234	0.806	0.933	0.971
Luo[14]	0.1038	0.940	4.963	0.203	0.870	0.948	0.976
本文方法(-C,-MF)	0.126	0.861	4.756	0.193	0.874	0.952	0.981
本文方法(-C)	0.124	0.840	4.835	0.194	0.853	0.956	0.983
本文方法	0.116	0.836	4.736	0.191	0.876	0.959	0.983

4.5　可视化结果与分析

使用可学习的多尺度注意力机制可以更好地获得特征信息,为了更直观地表达深度信息估计算法的优势,进行可视化实验。如图 6 所示,选用一组 KITTI 数据集图片进行深度信息测试,并将测试结果可视化与 monodepth2 算法单目测试结果对比。通过对比结果看出:算法中的大树和汽车等轮廓边缘相较于 monodepth2 算法更加清晰,且更好地保留了图像里物体的几何结构,体现了连续帧单目深度估计算法的优势。



RGB图片 Monodepth2算法 本文算法

图6 可视化结果对比

Fig. 6 Comparison of visualization results

为验证提出的网络的泛化性,使用 RealSence 相机中的单目视觉传感器拍摄彩色图片并进行真实场景测试。使用在 KITTI 数据集上训练所获得的网络模型对测试图片进行深度估计,测试效果如图 7 所示。结果分为两组,每一组上面是真实场景,下面是测试结果灰度图,物体越近,越呈现白色。模型在真实场景时获得的深度信息是连续的,且深度信息估计有明显的物体轮廓边缘。由真实场景可视化结果可知:该模型具有优良的泛化性,可以重复应用在其他真实场景中。

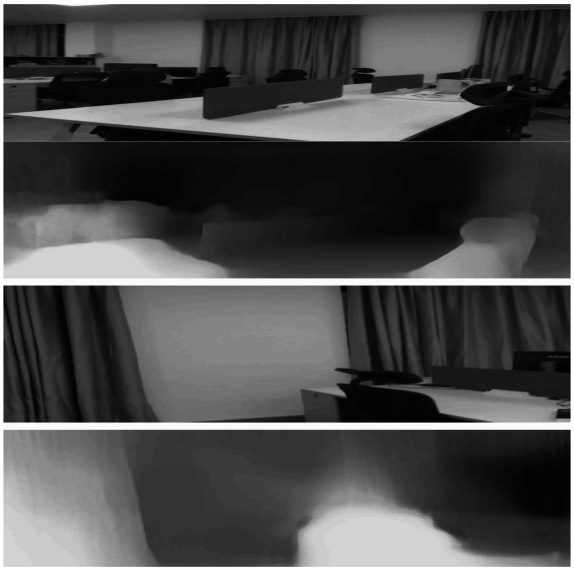


图7 测试结果可视化

Fig. 7 Visualization of test results

5 结论与展望

5.1 结 论

提出了一种基于多尺度特征融合与混合注意力结合的连续帧单目深度估计算法。在公开数据集 KITTI 上进行了评估和测试,通过评价指标对比其他的连续帧单目深度估计算法,可以发现:算法的绝对相对误差达到了 0.116,平方相对误差达到了 0.861,均方根误差达到了 4.736,对数均方根误差达到了 0.191 精确度 δ ($\delta<1.25$、$\delta<1.252$、$\delta<1.253$) 分别为 0.876、0.959 和 0.983,并且还通过在 KITTI 数据集上与 monodepth2 算法进行可视化结果对比,结果发现轮廓边缘细节更优,最后在自制的测试数据集上也证明了网络的泛化性。

5.2 展 望

针对连续帧单目深度估计提出了一种基于 U 型网络架构多尺度特征融合的方法,用来更好地利用多尺度特征,获取精确估计的深度信息。通过深度以及位姿编码器和解码器模块来设计网络架构,和不同的单目深度估计算法评估结果对比、在 KITTI 数据集上与 monodepth2 可视化结果对比,证明了设计网络的精确性和泛化性。由于使用的是轻量级网络,没有完全利用好提取的特征信息,所以估计精度仍有较大的提升空间。后续计划采用更深层次的网络,优化模型的特征提取能力,提高其深度估计精度。

参考文献(References):

[1] 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.
SONG Wei, ZHU Meng-fei, ZHANG Ming-hua, et al. A review of monocular depth estimation techniques based on deep learning[J]. Journal of Image and Graphics, 2022, 27(2): 292—328.

[2] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[J]. ArXiv e-Prints, 2014: arXiv: 1406.2283.

[3] SHELHAMER E, LONG J, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640—651.

[4] LAI Z, TIAN R, WU Z, et al. DCPNet: a densely connected pyramid network for monocular depth estimation[J]. Sensors, 2021, 21(20): 6780—9798.

[5] CHOI Y H, KEE S C. Monocular depth estimation using a Laplacian image pyramid with local planar guidance layers [J]. Sensors, 2023, 23(2): 845—861.

[6] 江俊君, 李震宇, 刘贤明. 基于深度学习的单目深度估计方法综述[J]. 计算机学报, 2022, 45(6): 1276—1307.
JIANG Jun-jun, LI Zhen-yu, LIU Xian-ming. Deep learning based monocular depth estimation: a survey[J]. Chinese Journal of Computers, 2022, 45(6): 1276—1307.

[7] ZHOU T, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6612—6619.

[8] XIE J, GIRSHICK R, FARHADI A. Deep3D: fully automatic 2D-to-3D video conversion with deep convolutional neural networks [C]//European Conference on Computer Vision, Cham: Springer, 2016: 842—857.

[9] CASSER V, PIRK S, MAHJOURIAN R, et al. Depth prediction without the sensors: leveraging structure for unsupervised learning from monocular videos[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8001—8008.

[10] PILLAI S, AMBRUŞ R, GAIDON A. SuperDepth: self-supervised, super-resolved monocular depth estimation[C]//Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), 2019: 9250—9256.

[11] GODARD C, MAC AODHA O, FIRMAN M, et al. Digging into self-supervised monocular depth estimation [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 3828—3838.

[12] 戴仁月, 方志军, 高永彬. 融合扩张卷积网络与 SLAM 的无监督单目深度估计[J]. 激光与光电子学进展, 2020, 57(6): 114—122.
DAI Ren-yue, FANG Zhi-jun, GAO Yong-bin. Unsupervised monocular depth estimation by fusing dilated convolutional network and SLAM[J]. Laser & Optoelectronics Progress, 2020, 57(6): 114—122.

[13] 叶星余, 何元烈, 汝少楠. 基于生成式对抗网络及自注意力机制的无监督单目深度估计和视觉里程计[J]. 机器人, 2021, 43(2): 203—213.
YE Xing-yu, HE Yuan-lie, RU Shao-nan. Unsupervised monocular depth estimation and visual odometry based on generative adversarial network and self-attention mechanism [J]. Robot, 2021, 43(2): 203—213.

[14] 罗志斌, 项志宇, 刘磊. 无监督单目图像深度和位姿估计算法优化[J]. 传感器与微系统, 2022, 41(7): 110—113.
LUO Zhi-bin, XIANG Zhi-yu, LIU Lei. Optimization of algorithm for unsupervised monocular image depth and pose estimation[J]. Transducer and Microsystem Technologies, 2022, 41(7): 110—113.

[15] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770—778.

[16] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module [C]//Proceedings of the Computer Vision-ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII. 2018: 3—19.

[17] 陈玲, 许钢, 伏娜娜, 等. 融合边缘检测的 3D 点云语义分割方法研究[J]. 重庆工商大学学报(自然科学版), 2022, 39(5): 1—9.
CHEN Ling, XU Gang, FU Na-na, et al. Research on 3D point cloud semantic segmentation method fused with edge detection[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2022, 39(5): 1—9.

[18] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation [C]//International Conference on Medical Image Computing and Computer-Assisted Intervention, Cham: Springer, 2015: 234—241.

[19] ZHOU Z, RAHMAN SIDDIQUEE M M, TAJBAKHSH N, et al. UNet++: A nested U-net architecture for medical image segmentation[M]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Cham: Springer International Publishing, 2018: 3—11.

[20] HUANG H, LIN L, TONG R, et al. UNet 3+: a full-scale connected UNet for medical image segmentation [C]//Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics Speech and Signal Processing, 2020: 1055—1059.

责任编辑:田　静