文章编号:1674-2869(2024)01-0085-06

基于多尺度深度图自适应融合的单目深度估计

郑 游,王 磊*,杨紫文
武汉工程大学电气信息学院,湖北 武汉 430205

摘 要:深度估计网络通常具有较多的网络层数,图像特征在网络编码和解码过程中会丢失大量信息,因此预测的深度图缺乏对象结构细节且边缘轮廓不清晰。本文提出了一种基于多尺度深度图自适应融合的单目深度估计方法,可有效保留对象的细节和几何轮廓。首先,引入压缩与激励残差网络(SE-ResNet),利用注意力机制对不同通道的特征进行编码,从而保留远距离平面深度图的更多细节信息。然后,利用多尺度特征融合网络,融合不同尺度的特征图,得到具有丰富几何特征和语义信息的特征图。最后,利用多尺度自适应深度融合网络为不同尺度特征图生成的深度图添加可学习的权重参数,对不同尺度的深度图进行自适应融合,增加了预测深度图中的目标信息。本文方法在NYU Depth V2数据集上预测的深度图具有更高的准确度和丰富的物体信息,绝对相对误差为0.115,均方根误差为0.525,精确度最高达到99.3%。

关键词:单目深度估计;注意力机制;多尺度特征融合网络;多尺度深度自适应融合网络
中图分类号:TP391.41 **文献标识码**:A **DOI**:10.19843/j.cnki.CN42-1779/TQ.202306025

Monocular depth estimation based on adaptive fusion of multi-scale depth maps

ZHENG You, *WANG Lei**, *YANG Ziwen*
School of Electrical and Information Engineering, Wuhan Institute of Technology, Wuhan 430205, China

Abstract: Depth estimation networks usually have a large number of layers, which may lose substantial image information in the process of image feature encoding and decoding, so the predicted depth maps lack detailed object structures and the maps' edges are not clear. In this paper, a monocular depth estimation method was proposed based on the adaptive fusion of multi-scale depth maps and it can better preserve the object details and geometric contours. First, the squeeze-and-excitation residual network was employed and the attention mechanisms were utilized to encode the feature maps from different channels, and more details of the long-distance depth maps can be reserved. Second, a multi-scale feature fusion network was adopted to fuse the feature maps of different scales, which produced feature maps with rich geometric and semantic information. Third, a multi-scale adaptive depth fusion network was used to add learnable weights to the depth maps generated by feature maps of different scales. Finally, the depth maps of different scales can be adaptively fused and the object information in the predicted depth maps increases. Experiments on NYU Depth V2 dataset demonstrate that the absolute relative error is 0.115, the root mean square error is 0.525, and the accuracy can reach 99.3%. The depth map predictions of the proposed method have higher accuracy and rich object information.

Keywords: monocular depth estimation; attention mechanism; multi-scale feature fusion network; multi-scale depth adaptive fusion network

收稿日期:2023-06-29
基金项目:武汉工程大学校内科学研究基金(21QD23)
作者简介:郑 游,硕士研究生。Email:898651804@qq.com
***通信作者**:王 磊,博士,讲师。Email:wanglei@wir.edu.com
引文格式:郑游,王磊,杨紫文. 基于多尺度深度图自适应融合的单目深度估计[J]. 武汉工程大学学报,2024,46(1):85-90.

单目深度估计是在单一视角下采集的彩色图像上估计图像中每个像素点相对于拍摄源的距离[1]，是计算机视觉中的一个重要课题，广泛应用于机器人导航、三维重建、自动驾驶等领域。近年来，基于深度学习的单目深度估计已经取得了很大的进步。最早是Eigen等[2]通过实验证明可以利用卷积神经网络获得图像的相对深度。此外，He等[3]提出使用不同速率的卷积同时提取特征图，然后通过通道将所有特征连接起来，得到一个融合的特征图。Wu等[4]将空间金字塔池应用于深度估计任务，使用不同大小的空洞卷积核来获取不同尺度的特征信息，取得了很好的估计效果。Wu等[5]提出了多级上下文和多模态融合网络融合多尺度多级上下文特征图，并从深度信息中学习对象边缘。尽管大多数基于深度学习的单目深度估计都表现出了很好的性能，但仍存在着一些挑战。

大多数深度估计模型都是基于深度神经网络提取图像的特征。然而，经过多次卷积和池化操作后会导致信息的丢失，特别是小物体的深度信息和结构细节。虽然减少卷积层可以保留较多的对象信息，但却无法获得高层次语义信息。为了解决这些问题，文献[6-7]引入了跳跃连接，为解码器模块添加低尺度特征，但大多数情况下，跳跃连接只包括单一尺度的特征图。此外，很多前期工作参考U-NET[8]架构来设计网络，然而逐步解码的方式使得U-NET在多尺度特征融合任务中表现不佳。

针对以上问题，本文在编码阶段引入注意力机制[9]，使用压缩与激励残差网络（squeeze-and-excitation residual network，SE-ResNet）进行编码，提取不同尺度的特征图，并增加了特征之间关联度。通过多尺度特征融合网络（mlti-scale feature fusion network，MSFFN）对不同尺度的特征图进行融合，以获得信息丰富的特征图。通过多尺度自适应深度图融合网络（multi-scale adaptive depth map fusion network，MADFN）对特征图进行解码，并通过可学习的权重对不同尺度的深度图进行加权融合，得到包含丰富场景信息的深度图。

1 网络结构设计

1.1 总体框架

整体网络结构如图1所示，主要由MSFFN和MADFN两个模块组成，网络整体采用编码器-解码器的结构。

1.1.1 MSFFN 在特征提取过程中，低尺度层特征图提供了大量的细节信息，但存在背景噪声。高尺度层特征图具有丰富的语义信息和抽象特征，但丢失了许多对象信息，导致深度预测的结果不够准确。针对这一问题，本文利用MSFFN将不同尺度的特征图与拉普拉斯残差特征图融合，丰富了特征图的语义信息和几何特征。

首先，利用公式（1）计算输入彩色图像S的拉普拉斯残差特征图L_i。

$$L_i = S_i - U\left(S_{i+1}\right) \quad (i = 1,2,3,4) \tag{1}$$

其中，i表示拉普拉斯金字塔中的层级索引，S_i是通过原始输入图像下采样到$1/2^{i-1}$尺度来获得，S_i的

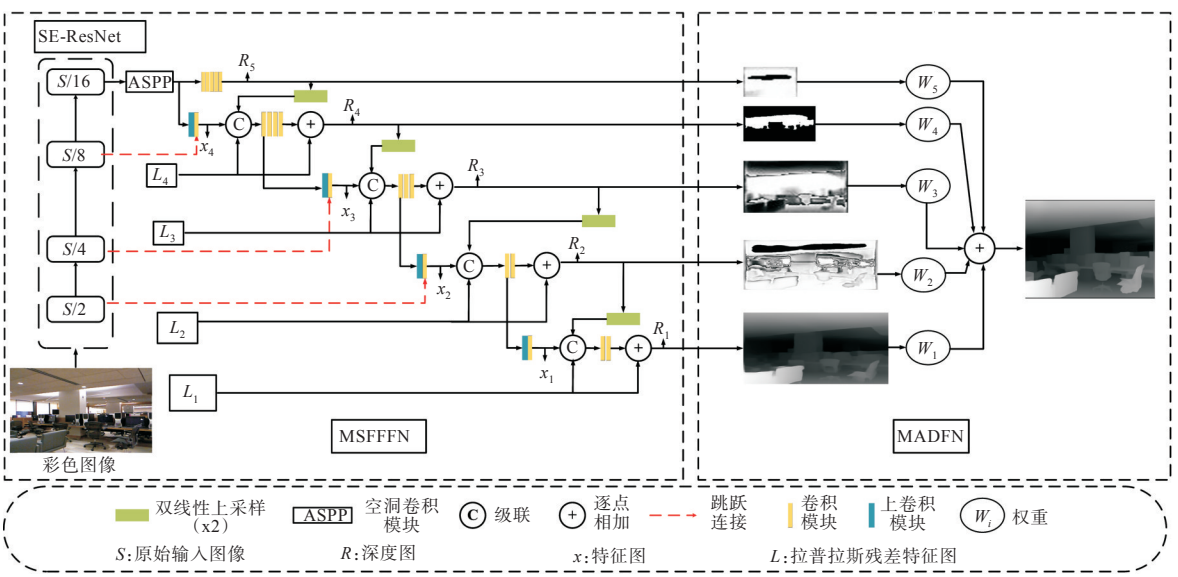

图1 网络总体结构

Fig. 1 Overall network structure

大小为 $S/2^{i-1}$，$U(\cdot)$ 表示采样倍数为2的上采样函数，在该方法中所有缩放过程都采用双线性插值。

然后，利用公式(2)将卷积得到的特征图 x_i、残差特征图 L_i 和从第 i+1层的深度图上采样的结果级联，将级联的特征输入堆叠的卷积块 B_i 中，并将输出结果与 L_i 逐像素相加，得到第 i 层深度图 R_i。其中，R_5 经过空洞空间卷积池化金字塔(atrous spatial pyramid pooling，ASPP)处理得到。

$$R_i = B_i\left[L_i, x_i, U\left(R_{i+1}\right)\right] + L_i \quad (i = 1,2,3,4) \tag{2}$$

通过这种策略，可以准确还原各个尺度下的局部特征细节。

1.1.2　MADFN　传统的深度图融合大多采用逐级相加的方式，在多尺度深度融合中表现较差[10]。受文献[11]的启发，本文提出了一种MADFN模块，通过上采样将多尺度的深度图调整到相同大小，并为每个深度图 d_k 设置一个可学习的权重 w_k。自适应融合后的深度图 D_{f} 可以表示为式(3)：

$$D_{\mathrm{f}} = \sum_{k \in l} w_k \cdot d_k \tag{3}$$

其中，k 表示尺度层级，$l = 1,2,3,4,5$。图像中坐标 (i,j) 位置的深度 $Y_{i,j}$ 可以表示为式(4)：

$$Y_{i,j} = \alpha_{i,j} y_{i,j}^1 + \beta_{i,j} y_{i,j}^2 + \gamma_{i,j} y_{i,j}^3 + \varepsilon_{i,j} y_{i,j}^4 + \delta_{i,j} y_{i,j}^5 \tag{4}$$

其中，$y_{i,j}^k$ 为深度图 d_k 中坐标 (i,j) 位置的深度值，$\alpha_{i,j}, \beta_{i,j}, \gamma_{i,j}, \varepsilon_{i,j}, \delta_{i,j}$ 分别为 $y_{i,j}^k$ 的权重。此外，本文对权重设置了 $\alpha_{i,j}, \beta_{i,j}, \gamma_{i,j}, \varepsilon_{i,j}, \delta_{i,j} \subset [0,1]$ 和 $\alpha_{i,j} + \beta_{i,j} + \gamma_{i,j} + \varepsilon_{i,j} + \delta_{i,j} = 1$ 两个归一化约束，通过将权重归一化增强网络模型的泛化能力，并且加速神经网络的训练收敛，减少训练时间和资源消耗。同时，归一化约束将权重限制在一定的范围内，减少了权重值的波动，避免了梯度爆炸或消失的问题，保证了训练的稳定性。

首先，输入彩色图像，如图2(a)所示。然后，经过MSFFN处理后得到不同尺度下的深度图，如图2(b–f)所示。最后，经过MADFN融合后得到的深度图如图2(g)所示。

从图2中可以看出，经过MADFN处理使融合后得到的深度图能够预测出较多的物体，且物体结构信息清晰。

1.1.3　损失函数　为了缓解数据集中近距离深度数据密集和远距离深度数据稀疏产生的不平衡问题，对文献[2]中使用的损失函数求平方根，得到损失函数的计算公式如下：

$$L(y, y^*) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2 - \frac{\lambda}{n^2}\left(\sum_{i=1}^{n} d_i\right)^2} \tag{5}$$

其中，$d_i = \log y_i - \log y_i^*$，$y_i$ 和 y_i^* 分别为第 i 个像素点的预测深度值和真实深度值，n 为有效像素的总数。平衡因子 λ 为超参数，用于权衡损失函数中误差方差和误差加权平方均值的重要性。本文将平衡因子 λ 设置为0.85，即平衡了误差方差和误差加权平方均值，同时也强调减小误差方差的重要性，以实现更加稳定和高质量的深度估计训练。

1.2　实　验

1.2.1　实验设置　本文所设计的网络结构使用Pytorch框架实现。硬件配置为16 GiB内存的i7-11800H处理器和12 GiB显存的NVIDIA GeForce RTX3060显卡。在训练阶段，批处理大小设置为8，初始学习率设置为0.000 1，权重衰减率为0.000 5，

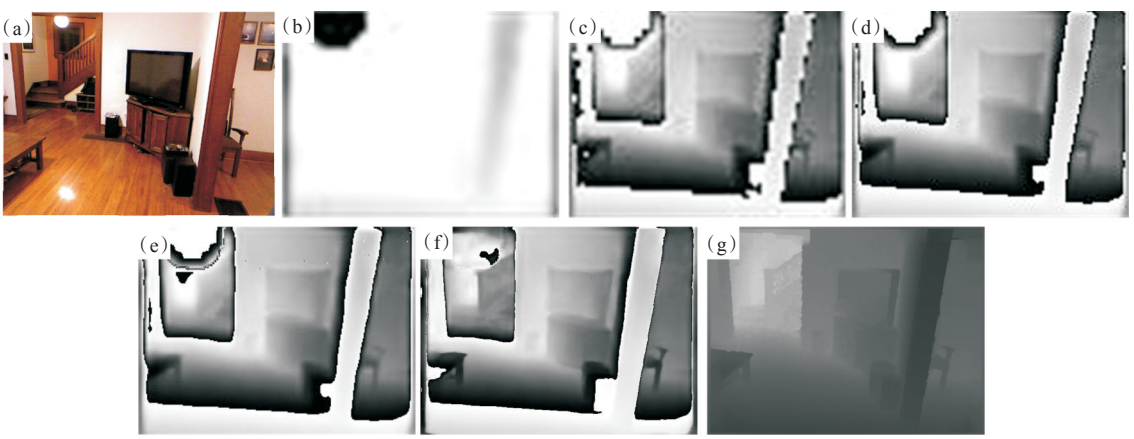

图2　不同尺度深度图与自适应深度融合结果：(a)输入图像，(b)第5尺度得到的深度图，(c)第4尺度得到的深度图，(d)第3尺度得到的深度图，(e)第2尺度得到的深度图，(f)第1尺度得到的深度图，(g)自适应融合得到的深度图

Fig. 2　Depth maps at different scales and adaptive depth fusion results: (a) input image, (b) depth map obtained at the 5th scale, (c) depth map obtained at the 4th scale, (d) depth map obtainedthe at 3rd scale, (e) depth map obtained at the 2nd scale, (f) depth map obtained at the 1st scale, (g) depth map obtained by adaptive fusion

网络参数采用动量为0.9的AdamW[12]优化器进行优化。

1.2.2 数据集 本文使用的深度估计数据集NYU Depth V2[13]，包含249个训练场景以及215个测试场景的RGB图像和深度图像。通过对训练场景中的原始图片做左右翻转和尺度变换等预处理，选择了36 253个样本进行训练，并从测试场景中选择654个样本进行测试。

1.2.3 评价指标 为了评价不同方法的深度估计效果，通常采用以下4个度量指标[2]：

(1)绝对相对误差：

$$E_{\mathrm{A}}=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i^*-y_i\right|}{y_i} \tag{6}$$

(2)均方根误差：

$$E_{\mathrm{R}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|y_i^*-y_i\right\|^2} \tag{7}$$

(3)对数均方根误差：

$$E_{\mathrm{L}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\|\log y_i^*-\log y_i\right\|^2} \tag{8}$$

(4)精确度：

$$A=\sum_i\frac{y_i}{n},y_i\ \mathrm{s.t.}\max\left(\frac{y_i^*}{y_i},\frac{y_i}{y_i^*}\right)=\delta<T \tag{9}$$

其中，公式(9)表示，取$\left(\frac{y_i}{y_i^*},\frac{y_i^*}{y_i}\right)$的最大值$\delta$，统计所有当$\delta$小于精确度阈值$T$时的像素点$y_i$数量占总像素点数量$n$的百分比。其中，$T$分别取1.25，$1.25^2$和$1.25^3$。精确度表示预测深度值和真实深度值的比值在一定阈值内的像素个数百分比，越接近1表示深度估计的效果越好。

2 结果与讨论

在NYU Depth V2数据集上将本文方法与其他单目深度估计方法进行对比实验，实验结果如表1所示。其中，误差指标数值越小，表示深度估计误差越小；精确度指标数值越接近1，表示深度估计越准确。

表1 NYU Depth V2数据集上的评价指标表

Tab. 1 Evaluation indicators on NYU Depth V2 dataset

方法	误差			精确度		
	E_{A}	E_{R}	E_{L}	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
文献[2]	0.215	0.907	0.285	0.611	0.887	0.971
文献[14]	0.213	0.759	0.087	0.604	0.885	0.973
文献[15]	0.209	0.749	0.091	0.701	0.921	0.973
文献[16]	0.118	0.527	0.058	0.858	0.971	0.989
文献[17]	0.121	0.581	0.057	0.863	0.970	0.991
文献[18]	0.115	0.509	0.051	0.828	0.965	0.992
本文方法	0.115	0.525	0.056	0.864	0.972	0.993

从表1中可以看出，本文方法的E_{A}为0.115，E_{R}为0.525，精确度最高为0.993，与大部分算法相比，具有更小的误差，以及更高的精确度。其中，文献[18]中采用序数回归的损失函数训练网络，得到了更小的误差值，但导致了计算量比本文方法要大，占用内存资源更多与耗时更久。

NYU Depth V2数据集上深度图可视化结果对比图如图3所示，从图3中可以看出，在室内数据集的表现上，本文方法预测的场景深度更连续，并且场景中物体的边缘轮廓也更加清晰。

为了进一步分析网络中各模块的作用，在NYU Depth V2数据集上进行了消融实验，以传统的编码-解码网络为基准(Baseline)，主要分析多尺度特征融合模块和多尺度自适应深度融合模块的效果。实验结果如表2所示。

由表2可以看出，单独加入MSFFN模块，E_{A}、

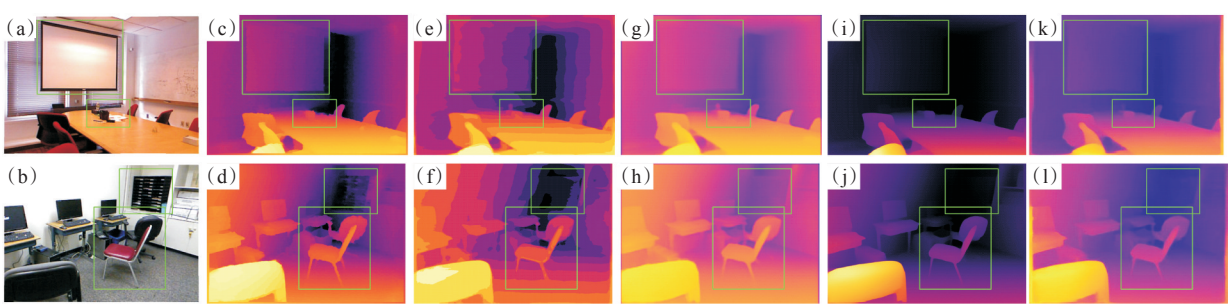

图3 NYU Depth V2数据集上深度图可视化结果对比图：(a、b)输入的2幅图像，(c、d)真值深度图，(e、f)文献[18]预测的深度图，(g、h)文献[19]预测的深度图，(i、j)文献[20]预测的深度图，(k、l)本文方法预测的深度图

Fig. 3 Comparison of depth map visualization results on NYU Depth V2 dataset: (a,b) two input images, (c,d) ground truth depth map, (e,f) depth map predicted by reference[18], (g,h) depth map predicted by reference[19], (i,j) depth map predicted by reference[20], (k,l) depth map predicted by the proposed method in this paper

E_R和E_L分别减少了约9.9%、5.0%和10.7%；阈值分别取1.25，1.25^2和1.25^3时，精确度分别提高了约3.2%、0.4%和0.2%。单独加入MADFN模块，E_A、E_R和E_L分别减少了约8.4%、5.3%和8.9%；精确度分别提高了约2.3%、0.5%和0.4%。2个模块都添加后，E_A、E_R和E_L分别减少了约12.2%、7.8%和10.7%；精确度分别提高了约4%、0.5%和0.6%。由此可见本文提出的2个网络模块起到了显著作用。

消融实验的可视化结果如图4所示。从图4中可以看出，加入了MSFFN和MADFN模块后，预测的深度图质量有明显提升。例如：第一个场景中两个相邻座椅之间的空隙也能精确地预测；第二个场景中书桌上的罐子与座椅上的扶手等细节地方轮廓也十分清晰。

表2　各模块性能分析

Tab. 2　Performance analysis of each module

模块	误差			精确度		
	E_A	E_R	E_L	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Baseline	0.131	0.565	0.056	0.835	0.967	0.987
Baseline+MSFFN	0.118	0.537	0.050	0.862	0.971	0.989
Baseline+MADFN	0.120	0.535	0.051	0.855	0.972	0.991
Baseline +MSFFN+MADFN	0.115	0.524	0.050	0.868	0.972	0.993

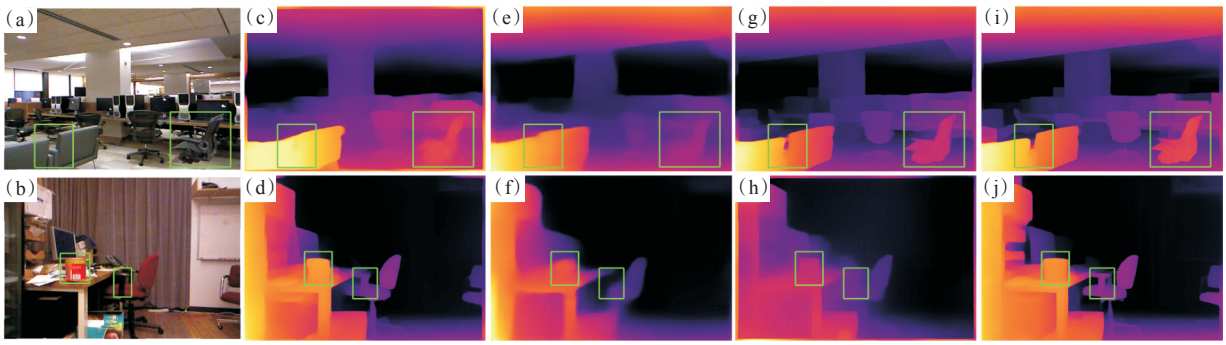

图4　消融实验结果图：(a、b) 输入的2幅图像，(c、d) Baseline预测的深度图，(e、f) 添加MSFFN模块预测的深度图，(g、h)添加MADFN模块预测的深度图，(i、j) 添加MSFFN和MADFN模块预测的深度图

Fig. 4　Images of ablation experiment results: (a, b) two input images, (c, d) depth map predicted by Baseline, (e, f) depth map predicted by adding the MSFFN module, (g, h) depth map predicted by adding the MADFN module, (i, j) depth map predicted by adding the MSFFN and MADFN modules

3　结　论

本文提出了一种多尺度深度图自适应融合的单目深度估计方法。通过在编码过程中引入注意力机制，提取到了更多特征细节。同时利用多尺度特征融合网络，融合不同尺度的特征图，得到具有丰富几何特征和语义信息的特征图。最后，利用多尺度自适应深度融合网络，自适应地融合不同尺度的深度图，增加了预测深度图中的目标信息，解决了预测深度时丢失物体信息和边缘深度估计不准确的问题。在NYU Depth V2数据集上测试的大量实验结果证明，本文提出的模型预测的深度图与其他方法相比具有更多的对象信息、更低的误差和更为准确的深度精度。后续研究中，考虑增加网络的泛化能力，提高在不同场景下预测深度的准确性。

参考文献

[1] 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.

[2] 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 (NIPS). Cambridge Ma: MIT Press, 2014: 2366-2374.

[3] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904-1916.

[4] WU K W, ZHANG S R, XIE Z. Monocular depth prediction with residual DenseASPP network [J]. IEEE

Access,2020,8:129899-129910.

[5] WU J W, ZHOU W J, LUO T, et al. Multiscale multilevel context and multimodal fusion for RGB-D salient object detection[J]. Signal Process, 2021, 178: 107766.

[6] QI F, LIN C H, SHI G M, et al. A convolutional encoder-decoder network with skip connections for saliency prediction [J]. IEEE Access, 2019, 7: 60428-60438.

[7] ZHAO S Y, ZHANG L, SHEN Y, et al. Super-resolution for monocular depth estimation with multi-scale sub-pixel convolutions and a smoothness constraint [J]. IEEE Access, 2019, 7: 16323-16335.

[8] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation [C]//Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241.

[9] QIU C C, ZHANG S Y, WANG C, et al. Improving transfer learning and squeeze-and-excitation networks for small-scale fine-grained fish image classification [J]. IEEE Access, 2018, 6: 78503-78512.

[10] LI L F, FANG Y, WU J, et al. Encoder-decoder full residual deep networks for robust regression and spatiotemporal estimation [J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(9): 4217-4230.

[11] PARK J S, JEONG Y, JOO K, et al. Adaptive cost volume fusion network for multi-modal depth estimation in changing environments [J]. IEEE Robotics and Automation Letters, 2022, 7(2): 5095-5102.

[12] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization [OL]. (2019-01-04) [2023-06-29]. https://arxiv.org/abs/1711.05101.

[13] LOZA A, MIHAYLOVA L, BULL D, et al. Structural similarity-based object tracking in multimodality surveillance videos [J]. Machine Vision and Applications, 2009, 20(2): 71-83.

[14] LIU F Y, SHEN C H, LIN G S, et al. Learning depth from single monocular images using deep convolutional neural fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2024-2039.

[15] 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.

[16] XU X F, CHEN Z, YIN F L. Monocular depth estimation with multi-scale feature fusion[J]. IEEE Signal Processing Letters, 2021, 28: 678-682.

[17] KIM D, LEE S, LEE J, et al. Leveraging contextual information for monocular depth estimation[J]. IEEE Access, 2020, 8: 147808-147817.

[18] FU H, GONG M M, WANG C H, et al. Deep ordinal regression network for monocular depth estimation [C]//Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2018: 2002-2011.

[19] YUAN W H, GU X D, DAI Z Z, et al. NeW CRFs: neural window fully-connected CRFs for monocular depth estimation[C]//Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE, 2022: 3906-3915.

[20] 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]. (2021-09-23) [2023-06-29]. https://arxiv.org/abs/1907.10326v5

本文编辑:霍淑利