DOI:10.19651/j.cnki.emt.2417764

基于拉普拉斯金字塔的特征融合深度估计算法

李铭汇 范哲意 朱艺璇
(北京理工大学集成电路与电子学院 北京 100081)

摘 要:在计算机视觉领域,单目深度估计在自动驾驶、场景重建等应用中的重要性引起了广泛的关注。然而,现有的自监督单目深度估计方法未能充分利用底层特征,导致了物体轮廓深度估计效果较差。为了解决这一问题,本文提出了一种多尺度特征融合解码方法,将原始 RGB 图像逐步高斯下采样以获得各级特征图,然后对其分别进行高斯上采样,利用上/下采样过程中相同尺寸的特征图对构建拉普拉斯金字塔,在解码时从各个尺度将下采样过程中丢失的轮廓线索与编码器提取到的特征相融合,从而引导解码器生成更精确的深度图,最大限度地提升编码器底层特征的利用效率。该方法与基线方法 Monodepth2 在 KITTI 数据集上的实验结果相比,绝对相对误差 Abs Rel 降低了 1.69%,平方相对误差 Sq Rel 降低了 6.80%,均方根误差 RMSE 降低了 1.00%,表明该方法对全局深度估计精度有所提升,此外可视化分析也验证了该方法对物体轮廓的深度估计效果有明显改善。

关键词:深度估计;自监督;拉普拉斯金字塔;特征融合
中图分类号:TP391.4;TN919.8 **文献标识码**:A **国家标准学科分类代码**:520.2

The feature fusion depth estimation algorithm based on Laplacian pyramid

Li Minghui Fan Zheyi Zhu Yixuan
(School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China)

Abstract: In the field of computer vision, monocular depth estimation has garnered significant attention due to its importance in applications such as autonomous driving and scene reconstruction. However, existing self-supervised monocular depth estimation methods fail to fully exploit low-level features, resulting in poor depth estimation performance for object contours. To address this issue, this paper proposed a multi-scale feature fusion decoding method. The original RGB image is progressively downsampled using a Gaussian approach to obtain feature maps at various levels, which are then upsampled using Gaussian processes. During upsampling and downsampling, Laplacian pyramids are constructed using feature maps of the same dimensions. During decoding, the lost contour cues from downsampling are fused with the features extracted by the encoder at each scale, guiding the decoder to generate more accurate depth maps and maximizing the utilization of low-level features from the encoder. Compared with the experimental results of the baseline method Monodepth2 on the KITTI dataset, this method reduced the absolute relative error Abs Rel by 1.69%, the squared relative error Sq Rel by 6.80%, and the root mean square error RMSE by 1.00%, indicating that this method has improved the accuracy of global depth estimation, and the visual analysis also verified that the method has significantly improved in the depth estimation effect of object contours.

Keywords: depth estimation; self-supervised; Laplace pyramid; feature fusion

0 引 言

利用二维图像进行深度估计获取三维世界的深度信息在自动驾驶、三维重建以及人机交互等领域有着广泛的应用[1-2]。在自动驾驶领域,深度估计可以识别障碍物距离从而实现安全驾驶;在机器人导航领域,通过深度估计使机器人对周围的三维空间进行感知从而实现自主导航和任务执行;在人脸识别领域,通过深度估计避免照片欺骗可以增强识别系统的安全性与可靠性[3]。激光雷达等硬件设备的发展在一定程度上推动了深度传感技术的进步,使用深度学习方法计算给定场景的深度信息估计因其具备硬件成本更低和潜在工作范围约束更少的优势,而逐渐成为计算机视觉领域的研究热点。

有监督深度学习方法在深度估计领域已经取得了较大

收稿日期:2024-12-30

的进展，使用真实深度作为监督信号，网络从输入图像中提取特征并学习每个像素点对应的深度。Eigen 等[4]首先采用有监督方法设计深度网络从单个图像中估计深度，设计了一个多尺度网络来组合全局粗深度图和局部细深度图。随后许多团队在此项工作的基础上引入了条件随机场(conditional random field，CRF)提高了深度估计的准确率[5-6]。Laina 等[7]提出使用新的上采样模块和反向 Huber 损失来改进训练。然而为有监督学习收集大型多样且具有真实深度的训练数据是一项艰难的挑战，自监督深度学习方法使用立体对或单目序列来训练深度估计模型，解决了有监督方法存在的难题。Garg 等[8]将深度估计作为一种视图合成问题，采用立体对作为自监督信号，提出最小化左图像与合成右图像之间的光度损失。Zhou 等[9]训练了一个单独的多视图姿态网络来估计两个连续帧之间的姿态。为了对动态场景进行建模，一些工作引入了多任务学习，例如光流估计[10]；或者引入了额外的约束，例如不确定性估计[11]。

近年来，自监督深度估计已经取得了显著进展，然而仍存在物体间遮挡区域深度难以计算、物体边界深度图模糊等问题[12]。Godard 等[13]提出了 Monodepth2，它使用最小重投影损失来减轻遮挡对深度估计带来的障碍，并使用自动掩蔽损失过滤掉与相机具有相同速度的移动对象。Monodepth2 在自监督深度估计领域取得了突破性进展，成为该领域的经典算法，后续许多方法都在 Monodepth2 的基础上进行研究。但是 Monodepth2 存在边界区域的深度估计不准确的问题，由于其采用的是全局特征来进行深度估计，虽然在大部分区域能够取得较好的效果，但对于图像中的边界区域(如物体的边缘或远离视线的部分)估计不够准确；另外 Monodepth2 假设输入图像是静态的，因此它无法有效处理动态场景或相机快速移动的情况。针对这些问题，该实验团队后续提出 Depth Hints[14]通过引入合成深度提示优化训练过程，提升深度预测的稳定性和精度，以及提出 manydepth[15]引入多帧输入自适应融合机制，结合单目和立体(多帧)信息，提升动态场景下的深度估计精度。而 HR-Depth[16]重新设计了深度网络中的跳跃连接，以获得更好的高分辨率特征从而提升了深度图的高频细节和边缘清晰度。

本文提出一种多尺度特征融合的方法来改善物体边界深度图模糊的问题。该方法以 Monodepth2 作为基线方法，利用拉普拉斯金字塔保存因下采样而损失的残差信息，并在解码过程中进行特征融合，从而在全局和局部提升深度估计准确率。实验结果表明，所提出方法在 KITTI 公开数据集上取得了较好的性能，有效缓解了物体间边界模糊问题，提升了深度估计的准确率。

1　基本原理

近年来，自监督单目深度估计仍在快速发展，自监督算法利用未标记的数据进行训练，因此可以节省大量的标注成本和时间。使用有效的编码器可以提取更有价值的特征，从而提高深度估计的效果，因此多数方法都在编码器部分进行改进，而忽略了在解码器部分进行准确率提升的探索。本文在自监督单目深度估计的经典算法 Monodepth2 的基础上，解码过程中加入多尺度特征融合。该方法将拉普拉斯金字塔的各个层级分解应用于解码过程，从而更加精确地计算深度图的局部细节即物体轮廓并提升全局的深度估计准确率。具体而言，本文利用拉普拉斯金字塔保存在下采样过程中损失的残差信息，并将其在解码过程中与中间深度图特征融合，引导解码器生成更加准确的深度图。

1.1　自监督单目深度估计

在计算机视觉领域，深度估计是理解图像内容的关键步骤之一。单目深度估计是通过单一的视角图像来推断深度信息。相比于双目或多目深度估计，单目深度估计除了在成本上有了很大节约，其最主要的优势在于从单一视角下进行深度估计推断场景的三维效果具有更强的泛化能力和鲁棒性。

自监督单目深度估计使用接近目标帧 I_t 时间的相邻帧 $It+n$ 来训练自监督深度网络，其中 $n \in \{-1,1\}$。使用深度网络 DepthNet 估计的深度 D_t 和位姿网络 PoseNet 估计的相对相机位姿 $T_t \to t+n$ 在与 I_t 相同的视点合成新的视频帧，合成视频帧为：

$$I_t+n \to t = I_t+n\langle proj(Dt, T_t \to t+n, \boldsymbol{K})\rangle \quad (1)$$

其中，$\langle\rangle$ 是采样算子，$proj()$ 是计算 I_{t+n} 投影在 Dt 方向中的二维坐标，$\boldsymbol{K}$ 是相机内置矩阵。在训练过程中的整体损失为：

$$L = \mu L_p + \lambda L_s \quad (2)$$

其中，L_p 是光度重投影损失，L_s 是边缘感知平滑损失，μ 与 λ 分别为 L_p 与 L_s 在整体损失中所占权重。L_p 光度重投影损失的作用是，在训练过程中对于每个像素，通过将光度重投影损失最小化来优化最佳匹配源图像。

$$L_p = \min_n pe(I_t, I_t+n \to t) \quad (3)$$

$$pe(I_a, I_b) = \frac{\alpha}{2}(1-SSIM(I_a, I_b)) + (1-\alpha)\| I_a - I_b \| \quad (4)$$

其中，pe 为光度重建误差例如像素空间的 L_1 距离[13]，I_a 和 I_b 分别为不同视频帧，$SSIM$ 为像素相似度[17]。

L_s 边缘感知平滑损失的作用是阻止所估计的深度的收缩[18]：

$$L_s = |\partial x d_t^*| e^{-|\partial x It|} + |\partial y d_t^*| e^{-|\partial y It|} \quad (5)$$

其中，$d_t^* = d_t / \overline{d_t}$ 是平均归一化逆深度，I_t 为当前视频帧。

1.2　基于拉普拉斯金字塔的特征融合深度估计方法

1)基线方法 Monodepth2

基线方法 Monodepth2 由深度网络 DepthNet 和位姿网络 PoseNet 构成。其中，深度网络 DepthNet 采用标准的

全卷积 U-Net 来预测深度,使用常用的 ResNet18 作为 DepthNet 编码器进行特征提取[10,19],网络结构如图 1 所示,位姿网络 PoseNet 将一对彩色图像作为输入,使用预训练的 ResNet18 作为位姿编码器,具有 4 个卷积层的位姿解码器来估计相邻图像之间的相应 6-DoF 相对位姿。

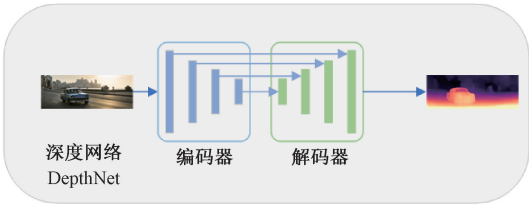


图 1 Monodepth2 深度网络 DepthNet 结构

Fig. 1 The network architecture of Monodepth2 DepthNet

作为单目深度估计领域的里程碑工作,Monodepth2 引入多视图外观匹配损失,有效缓解了动态物体遮挡导致的深度不连续问题;针对运动物体与静态场景的相机运动假设冲突,提出了一种自动掩蔽(auto-masking)方法来排除违反极几何约束的干扰像素;构建了一种多尺度外观匹配损失,在输入分辨率上执行所有图像采样,从而减少深度伪影。通过这 3 个步骤的改进,Monodepth2 在当时取得最佳的性能,其简洁的编码器-解码器架构和可扩展的训练策略更使其成为后续研究的基准模型。

2)所提出方法网络结构

本文方法的总体网络架构如图 2 所示。该网络由编码器和解码器构成的深度网络 DepthNet 和一个估计两相邻帧之间相机运动的位姿网络 PoseNet 构成,PoseNet 生成的相机位姿估计 $Tt \rightarrow t+n$ 作为 DepthNet 的监督信号,指导 DepthNet 在自监督框架下学习更准确的深度映射。

针对 Monodepth2 深度网络 DepthNet 编码器下采样过程中物体边缘特征退化的问题,本文提出在 DepthNet 中增加拉普拉斯金字塔残差模块,在解码过程中拉普拉斯金字塔残差特征图与编码器所获得的特征进行多尺度特征融合,引导解码过程恢复各个尺度空间的局部细节,改进后的结构如图 2 深度网络 DepthNet 所示,其中编码器采用 ResNet18,位姿网络 PoseNet 保留与基线实验相同的网络。

1.3 深度网络 DepthNet

1)拉普拉斯金字塔

在图像处理领域,一幅图像经过下采样操作后,直接进行上采样通常无法精确恢复至原始分辨率。这是因为在下采样的过程中丢失了部分细节信息,导致上采样时无法获得图像的全部特征。这种信息丢失导致在颜色和纹理的精度下降、物体轮廓模糊等,在深度估计等需要精确细节的任务中会显著影响结果。而这些被丢失的信息正是拉普拉斯金字塔的核心构成。拉普拉斯金字塔因其在下采样过程中能够保留局部信息,在图像处理、场景理解等多个领域中获得了广泛应用。特别是在高分辨率图像的恢复过程中,拉普拉斯金字塔展现了其重要作用。拉普拉斯金字塔通过逐

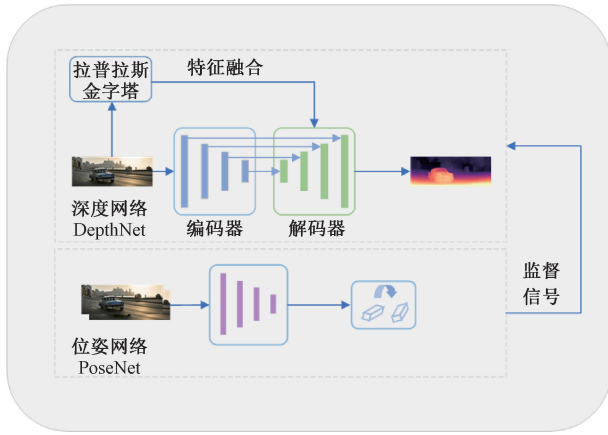


图 2 总体网络架构

Fig. 2 Overview of the proposed network architecture

层表示图像中各级别的细节差异,从而帮助恢复丢失的分辨率细节。其定义可以表示为:

$$L_i = G_i - upsample(G_i + 1) \tag{6}$$

其中,L_i 表示拉普拉斯金字塔中的第 i 层,G_i 表示高斯金字塔中的第 i 层,因此拉普拉斯金字塔中的第 i 层,等于高斯金字塔中的第 i 层与高斯金字塔中的第 $i+1$ 层的上采样结果之差[20]。如图 3 所示,展示了高斯金字塔和拉普拉斯金字塔的对应关系。

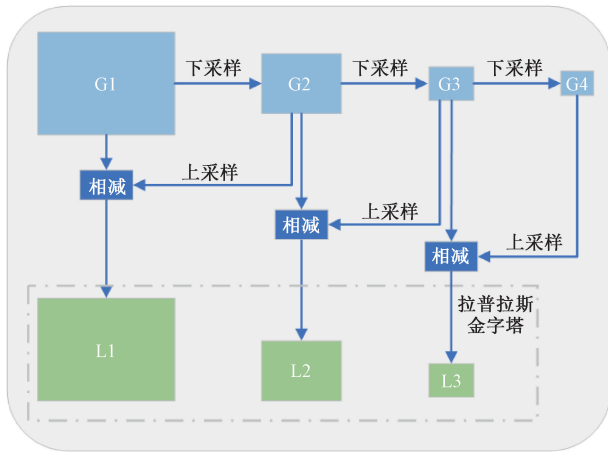


图 3 高斯金字塔与拉普拉斯金字塔对应关系

Fig. 3 The relationship between the Gauss pyramid and the Laplace pyramid

2)深度网络解码器

受到拉普拉斯金字塔能够有效保存在下采样过程中信息损失的启发,本文在深度估计任务中提出了一种基于多尺度特征融合的深度图恢复方法。具体而言,将拉普拉斯金字塔的每一层分别与编码器获得的特征图进行多尺度特征融合,通过结合每个金字塔级别的深度残差,从粗尺度到细尺度逐步恢复深度图,这种融合机制有助于提高深度图边界信息预测的准确性。本文使用 It 时刻的 RGB 图像作为输入,生成拉普拉斯金字塔残差特征 $L_5 \sim L_1$。

$$L_k = I_k - upsample(I_k + 1), k = 1,2,3,4,5 \quad (7)$$

其中，k 表示拉普拉斯金字塔中的级别索引，Ik 通过将原始输入图像下采样到 $1/2^{k-1}$ 尺度来获得，$upsample(\cdot)$ 表示上采样操作(即最近邻插值)。

这种残差特征图逐渐与中间深度图相结合，解码器中所有卷积层的滤波器大小设置为 3×3，这种设置保证了网络能够在捕获局部特征的同时保留较高的计算效率，所提方法的体系结构细节如表 1 所示。

表 1　解码器详细架构

Table 1　detailed architecture of the decoder

Level	Block	Filter size	Stride	Channel	In	Out
Level5	upconv0	3×3	2	512/256	S/32	S/32
	upsample	—	2	256/256	S/32	S/16
	upconv1	3×3	2	515/256	S/16	S/16
Level4	upconv0	3×3	2	256/128	S/16	S/16
	upsample	—	2	128/128	S/16	S/8
	upconv1	3×3	2	259/128	S/8	S/8
	disconv	3×3	2	128/1	S/8	S/8
Level3	upconv0	3×3	2	128/64	S/8	S/8
	upsample	—	2	64/64	S/8	S/4
	upconv1	3×3	2	131/64	S/4	S/4
	disconv	3×3	2	64/1	S/4	S/4
Level2	upconv0	3×3	2	64/32	S/4	S/4
	upsample	—	2	32/32	S/4	S/2
	upconv1	3×3	2	67/32	S/2	S/2
	disconv	3×3	2	32/1	S/2	S/2
Level1	upconv0	3×3	2	32/16	S/2	S/2
	upsample	—	2	16/16	S/2	S
	upconv1	3×3	2	19/16	S	S
	disconv	3×3	2	16/1	S	S

图 4 所示为解码器详细结构，解码器中的每一个阶段都与相应层的拉普拉斯金字塔残差特征图相结合，从粗尺度到细尺度逐层恢复图像的深度信息。在每一个融合步骤中，金字塔残差图像包含了在下采样过程中丢失的局部细节信息，在解码过程中与编码器所提取的特征图进行融合，能够帮助网络更好地捕获和恢复这些细节。

2　实验结果与分析

实验中，本文在 KITTI 数据集[21]上验证了基于拉普拉斯金字塔的多尺度特征融合深度估计方法对整体深度估计准确性的提高，同时与近年来提出的其他算法进行对比，进一步证明了该方法的有效性。所提出的方法在 PyTorch 中实现，使用 Adam 训练了 20 个 epoch，batch 大小为 12。

2.1　数据集与评价指标

KITTI 数据集包括从自动驾驶场景获取的各种道路环境，有城市、马路、住宅、校园和行人 5 类场景，继续采用 Eigen 等[22]对 KITTI 数据集的数据拆分，39 810 张图像作为训练集，4 424 张图像作为测试集。实验使用单目序列进行训练，遵循 Zhou 等[9]的预处理来去除静态帧。对所有的图像使用相同的内部函数，将相机主轴设置在图像的中心，焦距设置为 KITTI 所有焦距的平均值。

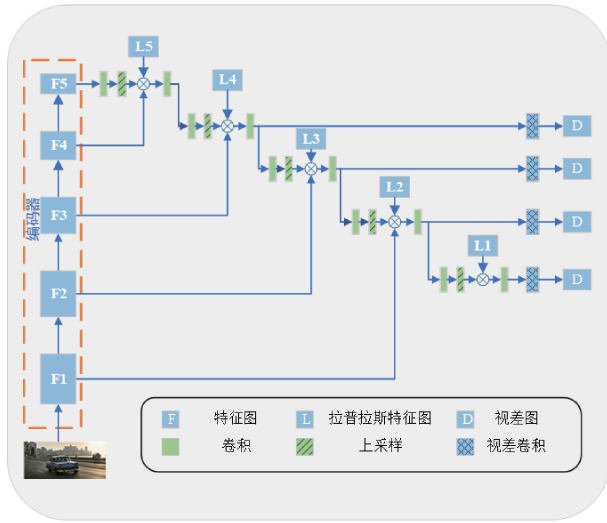


图 4　解码器网络结构

Fig. 4　Architecture of decoder

本文选用 Eigen 提出的 7 个常用实验评估指标来反映深度估计的精度，分别是 $Abs\ Rel$、$Sq\ Rel$、$RMSE$、$RMSE\ log$、$\delta < 1.25$、$\delta < 1.25^2$、$\delta < 1.25^3$，这些指标已经广泛应用于单目视频深度估计的性能评估，定义如下：

$$Abs\ Rel = \frac{1}{T}\sum \left| \frac{y_{pre} - y_{gt}}{y_{gt}} \right| \quad (8)$$

$$Sq\ Rel = \frac{1}{T}\sum (y_{pre} - y_{gt})^2 \quad (9)$$

$$RMSE = \sqrt{\frac{1}{T}\sum (y_{pre} - y_{gt})^2} \quad (10)$$

$$RMSE\ log = \sqrt{\frac{1}{T}\sum (\log(y_{pre}) - \log(y_{gt}))^2} \quad (11)$$

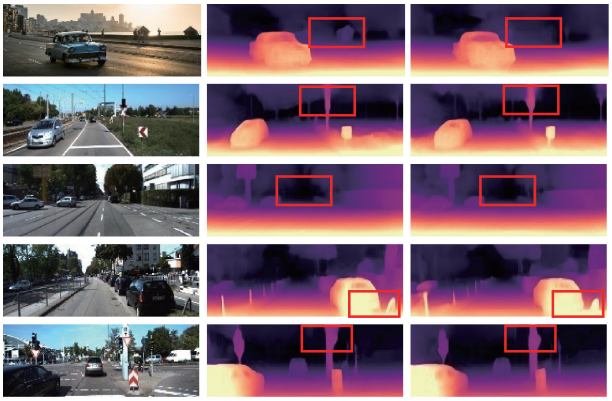
$$\delta x: \%ofypres.t.max\left(\frac{y_{pre}}{y_{gt}}, \frac{y_{gt}}{y_{pre}}\right) = \delta < thr, thr = 1.25^1, 1.25^2, 1.25^3 \quad (12)$$

其中，y_{pre} 和 y_{gt} 分别表示预测的深度图和真实深度，T 是真实深度中有效像素的总数，δx 表示深度估计在不同精度下的准确率。

2.2　实验结果

为了证明方法的有效性，在 KITTI 数据集上对该方法与其他方法进行了定性与定量的对比。首先，与基线方法的定性结果比较如图 5 所示。具体来说，以往大多数的无监督单目深度估计方法都难以准确估计物体边界的深度信息，例如图 5 方框中强光背景下或光线不足等条件下房屋、

人、指示牌等物体在基线方法的深度估计结果边界模糊,本文的方法在深度图物体边界有明显优势。



(a) RGB原图 (a) RGB (b) 所提出方法 (b) Proposed (c) 基线方法 (c) Baseline

图 5 KITTI 数据集深度估计定性结果

Fig. 5 Qualitative results on KITTI

基于前文所提到的指标,将本文方法与基线方法 Monodepth2 以及其他先进方法在 KITTI 数据集上进行了比较,相应的结果如表 2 所示。可以看出,本文的方法由于在深度网络 DepthNet 中增加拉普拉斯金字塔特征融合模块使底层特征利用的更充分,改善了物体轮廓深度估计模糊的问题,相比于基线实验有了明显提升。同时可以与最近的方法 Liu 等[23](ResNet18)与 GC-Depth[24] 相竞争。SC-SfMLearner 是在 Monodepth2 框架基础上进行改进与优化的深度估计方法。Liu 等[23]提出的基于 ResNet18 架构的模型以及 GC-Depth 网络设计均与 Monodepth2 保持结构相似性,三者均采用 U-Net 作为深度网络的基本架构,并通过单帧图像输入完成特征提取过程。具体而言,这些方法在编码器-解码器结构的设计理念上具有继承性,但通过引入不同的改进策略实现性能提升。其中 Liu 等(ResNet18)的方法由于加入了光流估计,在某些指标有更优秀的效果。

表 2 KITTI 数据集上的定量对比

Table 2 Quantitative results on KITTI

方法	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25^1$	$\delta<1.25^2$	$\delta<1.25^3$
Monodepth2	0.118	0.927	4.893	0.195	0.872	0.958	0.981
SC-SfMLearner(2021)[25]	0.119	0.857	4.950	0.197	0.863	0.957	0.981
Liu 等(ResNet18)(2024)[23]	0.118	0.903	4.809	0.190	0.870	0.959	0.982
GC-Depth(2024)[24]	0.122	0.868	4.986	0.200	0.857	0.953	0.980
Ours	0.116	0.864	4.844	0.195	0.874	0.959	0.981

3 结 论

本文提出了一种基于拉普拉斯金字塔的多尺度特征融合方法,该方法通过在解码器中引入拉普拉斯金字塔残差图像,将不同尺度的特征逐步融合,准确恢复各个尺度空间的局部细节。多尺度特征融合的策略能够有效缓解下采样过程中信息的丢失问题,尤其是在复杂场景下提高了物体边界深度信息的预测性能。实验结果表明,该方法能够在深度估计任务中提供更准确的预测。

参考文献

[1] SHOTTON J, GIRSHICK R, FITZGIBBON A, et al. Efficient human pose estimation from single depth images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(12): 2821-2840.

[2] 余萍,胡旭欣. 基于单目深度估计和校准参数的距离测算方法[J]. 电子测量技术, 2022, 45(20): 88-94.
YU P, HU X X. Distance measurement method based on monocular depth estimation and calibration parameters[J]. Electronic Measurement Technology, 2022, 45(20): 88-94.

[3] 刘安旭,黎向锋,刘晋川,等. 改进卷积空间传播网络的单目图像深度估计[J]. 电子测量技术,2021,44(23):78-85.
LIU AN X, LI X F, LIU J CH, et al. Monocular image depth estimation of improved convolutional spatial propagation network [J]. Electronic Measurement Technology, 2021, 44(23): 78-85.

[4] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network[C]. 28th International Conference on Neural Information Processing Systems, 2014, 2: 2366-2374.

[5] LI B, SHEN CH H, DAI Y CH, et al. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1119-1127.

[6] LIU F Y, SHEN CH H, LIN G SH, et al. Learning depth from single monocular images using deep convolutional neural fields[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(10): 2024-2039.

[7] LAINA I, RUPPRUCHT C, BELAGIANNIS V, et

al. Deeper depth prediction with fully convolutional residual networks [C]. 2016 Fourth International Conference on 3D Vision (3DV). IEEE, 2016: 239-248.

[8] GARG R, BG V K, CARNEIRO G, et al. Unsupervised cnn for single view depth estimation: Geometry to the rescue[C]. Computer Vision-ECCV 2016: 14th European Conference, 2016: 740-756.

[9] ZHOU T H, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1851-1858.

[10] YIN ZH CH, SHI J P. Geonet: Unsupervised learning of dense depth, optical flow and camera pose[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 1983-1992.

[11] CASSER V, PIRK S, MAHJOURIAN R, et al. Unsupervised monocular depth and ego-motion learning with structure and semantics[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019:381-388.

[12] 曲熠,陈莹.基于稳定光度损失的无监督单目深度估计[J].电子测量与仪器学报,2024,38(11):158-167.
QU Y, CHEN Y. Unsupervised monocular depth estimation based on stable photometric loss [J]. Journal of Electronic Measurement and Instrumentation,2024,38(11):158-167.

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

[14] WATSON J, FIRMAN M, BROSTOW G J, et al. Self-supervised monocular depth hints [C]. IEEE/CVF International Conference on Computer Vision, 2019: 2162-2171.

[15] WATSON J, MAC AODHA O, PRISACARIU V, et al. The temporal opportunist: Self-supervised multi-frame monocular depth[C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 1164-1174.

[16] LYU X Y, LIU L, WANG M M, et al. HR-depth: High resolution self-supervised monocular depth estimation [C]. AAAI Conference on Artificial Intelligence, 2021, 35(3): 2294-2301.

[17] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: From error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[18] GODARD C, MAC AODHA O, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2017: 270-279.

[19] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. ArXiv preprint arXiv:1409.1556, 2014.

[20] BURT P J, ADELSON E H. The Laplacian Pyramid as a compact image code[J]. IEEE Transactions on Communications, 1983, 31(4):532-540.

[21] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The kitti dataset [J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[22] EIGEN D, FERGUS R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture [C]. IEEE International Conference on Computer Vision, 2015: 2650-2658.

[23] LIU X L, SHEN F R, ZHAO J, et al. Self-supervised learning of monocular 3D geometry understanding with two-and three-view geometric constraints[J]. The Visual Computer, 2024, 40(2): 1193-1204.

[24] XIONG M K, ZHANG ZH H, LIU J Y, et al. Monocular depth estimation using self-supervised learning with more effective geometric constraints[J]. Engineering Applications of Artificial Intelligence, 2024,128:107489.

[25] BIAN J W, ZHAN H Y, WANG N Y, et al. Unsupervised scale-consistent depth learning from video[J]. International Journal of Computer Vision, 2021, 129(9): 2548-2564.

作者简介

李铭汇,硕士研究生,主要研究方向为基于自监督深度学习的单目深度估计。

E-mail:liminghui0303@163.com

范哲意,博士,高级实验师,硕士生导师,主要研究方向为智能图像分析与识别。

E-mail:funye@bit.edu.cn

朱艺璇(通信作者),硕士,助理实验师,主要研究方向为智能图像分析与识别。

E-mail:zhuyixuan@bit.edu.cn