

doi: 10.12052/gdutxb.190140

结合注意力与无监督深度学习的单目深度估计

岑仕杰, 何元烈, 陈小聪

(广东工业大学 计算机学院, 广东 广州 510006)

摘要: 针对当前的无监督单目深度估计方法边界模糊的问题, 提出了一种基于双重注意力模块的网络架构。这种架构能有效利用图像特征的远程上下文信息解决深度估计中的边界模糊问题。整个框架使用基于视图合成的无监督方法训练, 模型框架包括深度估计网络与位姿估计网络, 同步估计深度和相机位姿变换。双重注意力模块嵌入在深度估计网络中, 包含位置注意力模块和通道注意力模块, 能表示远程空间位置 and 不同特征图间的上下文信息, 从而使网络估计出细节更好的深度信息。在KITTI数据集以及Make3D数据集上的实验结果表明, 本文的方法能有效提高单目深度估计的精度和解决深度估计边界模糊问题。

关键词: 深度估计; 无监督学习; 深度学习; 注意力; 机器人技术
中图分类号: TP249 **文献标志码:** A **文章编号:** 1007-7162(2020)04-0035-07

A Monocular Depth Estimation Combined with Attention and Unsupervised Deep Learning

Cen Shi-jie, He Yuan-lie, Chen Xiao-cong

(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: To solve the problem of boundary blurring of current unsupervised monocular depth estimation method, a network architecture is proposed based on dual attention module. This architecture can effectively solve the problem of boundary blurring of depth estimation by using long-range context information of image features. The model framework that includes depth estimation network and pose estimation network is trained by an unsupervised method based on view synthesis and estimation depth and camera pose transformation at the same time. The dual attention module is embedded in the depth estimation network, including position attention module and channel attention module. This module can represent the long-range spatial location and the context information between different feature maps, so that the network can estimate the depth information with better details. The experimental results on the KITTI dataset and the Make3D dataset show that our method can effectively improve the accuracy of the monocular depth estimation and can solve the depth estimation boundary blur problem.

Key words: depth estimation; unsupervised learning; deep learning; attention; robotics

深度信息对于理解三维场景有着重要的作用, 它能应用于各种机器人技术, 例如三维重建、三维目标检测和同时定位与构图(Simultaneous Localization and Mapping, SLAM)^[1]。从图像中获取深度信息的任务被称为图像深度估计, 由于摄像头轻便、廉价等特性, 通过图像来恢复像素级深度在计算机视觉领域中越来越受关注。

随着深度学习技术的快速发展, 许多工作使用监督深度学习从图像中推断出深度信息。然而监督学习所需要的真值数据的获取并不容易, 因此最近的工作试图使用无监督学习解决深度估计问题。在缺少真实标注的情况下, 要学习从像素到深度的映射, 模型需要附加其他约束条件。其中一种形式的无监督深度估计是使用同步的双目图像对进行训练。

收稿日期: 2019-11-18
基金项目: 国家自然科学基金资助项目(61876043)
作者简介: 岑仕杰(1992-), 男, 硕士研究生. 主要研究方向为计算机视觉、深度学习和深度估计
通信作者: 何元烈(1976-), 男, 副教授. 主要研究方向为计算机视觉、深度学习和智能机器人, E-mail:heyuanlie@163.com

同步双目图像对仅在训练期间使用,模型估计左右图像视差或图像深度,从而重建图像,通过比较图像之间的差异来训练模型。Xie等^[2]提出了一种具有离散深度的模型,用于新视图合成的问题,随后Garg等^[3]通过估计连续视差值来扩展这种方法。Godard等^[4]通过使用左右深度一致性项来产生优于当前部分监督方法的结果。另一种约束较少的无监督形式是使用单目视频数据来训练模型,将图像重建损失用作监督信号来训练网络。这种无监督的训练方式除了估计深度之外,网络还必须估计帧间的相机姿势。Zhou等^[5]开创性地仅使用单目视频训练了深度估计网络以及单独的位姿估计网络,为了处理非刚性场景运动,他们提出了使用网络学习解释掩模,允许模型忽略违反刚性场景假设的特定区域。Mahjourian等^[6]使用更明确的几何损失来联合学习深度和相机运动的刚性场景。Yin等^[7]在文献[5]研究的基础上添加了一个细化网络,以估计残差光流。这些方法仅仅利用单目视频序列或者双目图像对就可以完成训练任务,并在户外场景中产生比部分监督方法更好的结果。

然而,以上方法都没有利用好场景中的上下文信息。Huang等^[8]研究了自然场景的深度图像的统计量,表明深度图像可以分解为分段的平滑区域,彼此之间几乎没有依赖性,并且通常存在尖锐的不连续性。因此,场景深度的变化与场景中“对象”的概念息息相关,而不是像颜色、纹理、照明等一些底层特征。当前部分研究^[5]采用基于边缘感知的平滑损失来约束模型,从而产生在“对象”内较为平滑的深度图像。但是基于图像梯度的边缘图并不能很好地表现出对象的边界。为了解决该问题,本文提出了利用Fu等^[9]在语义分割领域所提出的双重注意力模块改进深度估计网络,通过注意力机制更有效地利用对象内和对象间的上下文信息,增强模型的特征提取能力。本文的方法在KITTI数据集和Make3D数据集上的验证结果表明了注意力机制对提高深度估计精度的有效性。

1 相关研究

1.1 问题描述

图像记录的是三维世界在成像平面上的投影信息,图像的深度估计是从图像信息中估计场景深度的问题,在计算机领域属于三维重建的一部分,这个问题用数学语言来描述为 $D = F(I)$,其中 D 为深度、 I 为图像,而 F 为图像到深度的映射函数。由于尺度的

模糊性,单目深度估计是一种不适定(ill-posed)问题,所以几乎不能直接求解 F 。许多学者开始使用监督深度学习进行深度估计,但由于获取大规模的真实标签的数据费时且昂贵,因而最近许多研究关注无监督的深度学习方法。

1.2 视图重建作为监督信号

利用视图重建作为监督信号是一种无监督方法,其核心思路是以深度和位姿作为中间量,结合对极几何进行视图重建。假设观察场景是静止的,给出两张不同视角下拍摄的视图 I_t 和 I_s ,若已知视图 I_t 的深度图 D_t 和 I_t 到 I_s 的坐标变换矩阵,则 I_t 和 I_s 之间的像素映射关系为

$$p_s = K T_{t \rightarrow s} D_t K^{-1} p_t \quad (1)$$

其中 K 为相机内参, $T_{t \rightarrow s}$ 为 I_t 到 I_s 的坐标变换矩阵, p_t 、 p_s 分别为两张视图的像素坐标。通过网络模型可学习出每个像素的深度以及帧间位姿变换,因此根据式(1)的映射关系可以利用插值算法(如双线性插值)对不同视角下的图像进行合成并与目标视图比较,从而无监督训练模型实现估计深度和位姿变换。

2 系统模型框架

本文模型框架如图1所示,模型框架分为深度估计网络和位姿变换估计网络两部分。深度估计网络以单张彩色图像作为输入,不同于之前的一些工作^[5,7],本文深度估计网络输出的是稠密的深度图,由于直接估计深度比估计视差少了取逆操作,整个系统的训练更容易收敛。位姿估计网络的输入为两帧图像,输出为6-DoF位姿变换。模型估计的深度图与位姿变换用于视图重建,重建的视图与目标视图的对比误差作为损失来训练神经网络,训练过程无需真实深度和实际相机运动的位姿变换标注。

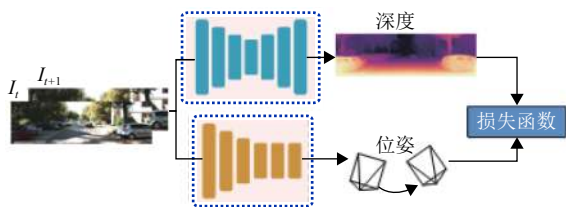


图1 模型框架

Fig.1 Model framework

2.1 网络结构概述

本文模型框架包含深度估计网络和位姿变换估计网络,两个网络均为全卷积网络。深度估计网络基于U-Net架构,并且包含跳跃连接与注意力模块,在

提取深层特征的同时又能表示远程的上下文信息,结构如图2所示。为了提取鲁棒的图像特征,本文将ResNet18^[10]作为用于提取RGB图像特征的编码器。与现有工作中使用DispNet^[5]和基于ResNet50的模型相比,本文的编码器参数更少,运行速度更快。本文把在ImageNet上预训练的权重用作编码器权重的初始化,实验表明与从头开始训练相比,这种初始化方式能提高模型的准确率。

由于输入图像通过编码器提取特征图是下采样过程,因此需要进行上采样操作对特征图分辨率进行还原。深度估计网络的解码器由5个上采样模块组成,除了输出处,解码器其他地方均使用指数线性单元(Exponential Linear Unit, ELU)作为激活函数。与文献[5]中基于反卷积的上采样模块不同,本文的上采样模块由卷积运算层以及最近邻插值算法组成,结构如图2中虚线标注区域所示。为了建模远程的上下文信息,增强特征之间的关联性,本文在深度估计网络的解码器部分加入了注意力模块。解码器的前两

层嵌入了双通道注意力模块,包含位置注意力模块与通道注意力模块,可以学习特征之间的上下文信息且不会增加过多的计算开销。深度估计层由3×3的卷积运算以及Sigmoid激活函数组成,输出图像的深度信息。为了将输出限制在一个合理的范围,本文对输出进行了线性变换,即 $D = a + (b - a)\sigma$,这里 D 是模型输出的深度图, σ 是Sigmoid激活函数的输出, a 、 b 分别是最小和最大深度值,本文取值为0.001和100。

位姿变换估计网络使用全卷积网络,输入和输出通道数都为6。编码器部分同样是采用标准的ResNet18结构。解码器共4层卷积运算,第1、第4层卷积核大小是1×1,第2层和第3层卷积核大小为3×3,除输出层外其余层激活函数均为修正线性单元(Rectified Linear Unit, ReLU)。图像序列按批大小堆叠输入到网络,通过编码器提取出特征图,再经过后续卷积操作得出不同帧的高层特征,最后通过1×1大小的卷积输出位姿。输出位姿是一个6维的位姿变换向量,前3维代表旋转,后3维代表位移。

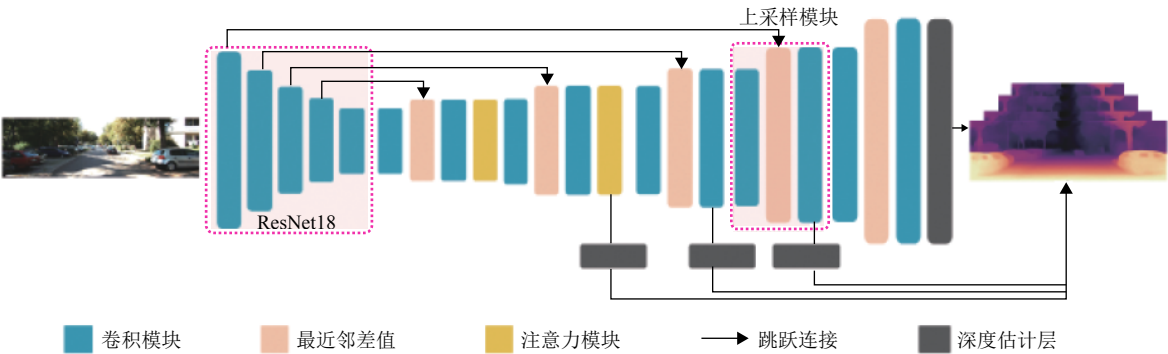


图 2 深度估计网络结构
Fig.2 Depth estimation network structure

2.2 结合双重注意力模块的深度估计网络

给定一张输入图像,图像中的对象物体在尺度、角度和光照各不相同,而卷积运算仅仅只拥有局部感受野,这容易导致一些细长的物体(例如路灯)的深度估计出现断裂现象。为了使网络更有效地利用场景的全局信息和表示特征之间的依赖关系,提高深度估计精度,本文利用双重注意力模块对深度估计网络进行改进。

双注意力模块包含两个注意力模块,即位置注意力模块和通道注意力模块。两种注意力模块分别捕获空间和通道维度的远程上下文信息。双通道注意力模块嵌入到深度估计网络的解码器中,图3为两种注意力模块的结构示意图。

2.2.1 位置注意力模块

传统全卷积网络提取的局部特征缺少全局信息,无法表示局部特征之间的关系,因此在估计深度时容易出现边缘与实际物体不相符的问题。为了对局部特征的上下文关系进行建模,本文引入了位置注意力模块。位置注意力模块如图3(a)所示。对于经过卷积层编码的特征图 $X \in \mathbf{R}^{C \times H \times W}$,首先将其输入1×1卷积层分别把通道数降维并生成两个新的特征 $Q \in \mathbf{R}^{r \times H \times W}$ 和 $K \in \mathbf{R}^{r \times H \times W}$,这里 r 本文取值为8。然后将 Q 、 K 重塑为 $Q \in \mathbf{R}^{r \times N}$ 和 $K \in \mathbf{R}^{r \times N}$,并将 Q 的转置与 K 进行矩阵乘法,这里 $N=H \times W$ 。最后将得到的结果通过softmax层来计算空间注意力图 $S \in \mathbf{R}^{N \times N}$,如式(2)所示。

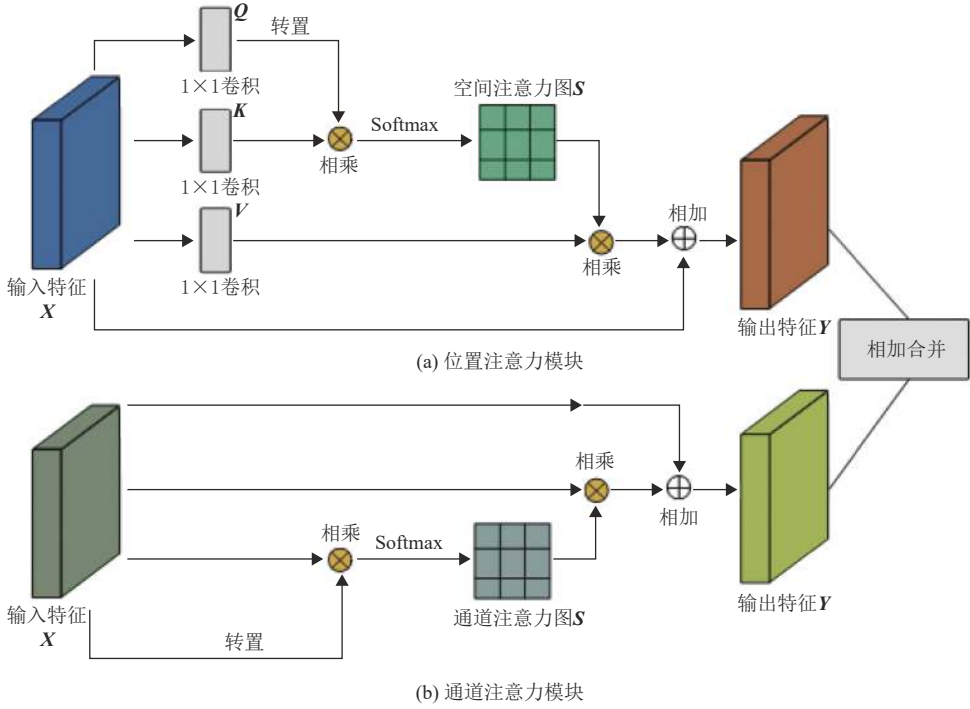


图3 双重注意力模块

Fig.3 Dual attention module

$$S_{ji} = \frac{\exp(Q_i \cdot K_j)}{\sum_{i=1}^N \exp(Q_i \cdot K_j)} \quad (2)$$

如果两个位置的特征表示越相似,它们之间的相关性就越高。同时,将输入特征 X 馈入卷积层以生成新的特征图 $V \in \mathbf{R}^{C \times H \times W}$ 。将 V 重塑为 $V \in \mathbf{R}^{C \times N}$ 然后在 V 和 S 的转置之间执行矩阵乘法。最后,为了使模块更灵活,本文将 V 和 S 相乘的结果与输入特征 X 乘以比例参数,并进行逐元素求和运算,得到最终输出 $Y \in \mathbf{R}^{C \times H \times W}$,如式(3)所示。

$$Y_j = \alpha \sum_{i=1}^N (S_{ji} V_i) + \beta X_j \quad (3)$$

其中 α 被初始化为0, β 被初始化为1,并逐渐从训练中分配两者权重。从式(3)可以得出,每个位置上的输出特征 Y 是所有位置上的特征与原始特征的加权和。因此它具有全局上下文视图,并根据空间注意力图选择性地聚合上下文。相似的对象特征实现了关联,从而提高了对象内部特征的紧凑性。

2.2.2 通道注意力模块

每个通道的高级特征图都可以视为特定于对象的响应,不同的特征图之间存在关联,这些关联与场景三维结构密切相关。通过利用通道特征图之间的相互依赖性,模型可以改进特定场景对象的特征表示。因此,本文利用通道注意力模块来显式地建模通道之间的相互依赖性,通道注意力模块的结构如

图3(b)所示。与位置注意力模块不同,这里直接从原始特征计算出通道注意力图。具体来说,将输入特征 $X \in \mathbf{R}^{C \times H \times W}$ 重塑成 $X \in \mathbf{R}^{C \times N}$ 与其转置之间进行矩阵乘法,然后应用softmax层获得通道注意力图 $S \in \mathbf{R}^{C \times C}$,见式(4)。

$$S_{ji} = \frac{\exp(X_i \cdot X_j)}{\sum_{i=1}^N \exp(X_i \cdot X_j)} \quad (4)$$

其中 S_{ji} 衡量第 i 个通道对第 j 个通道的影响。随后,在 S 和 X 的转置之间执行矩阵乘法操作,然后将结果与输入特征 X 乘以比例参数并进行逐元素求和运算,以获得最终输出 $Y \in \mathbf{R}^{C \times H \times W}$,如式(5)所示。

$$Y_j = \lambda \sum_{i=1}^C (S_{ji} X_i) + \omega X_j \quad (5)$$

其中 λ 与 ω 分别从0和1开始逐渐学习权重。经通道注意力模块处理后,每个通道的最终特征是所有通道的特征与原始特征的加权和,该特征对特征图之间的远程依赖性进行建模,它有助于提高特征的可区分度,帮助网络表示场景的结构信息。

2.2.3 嵌入注意力模块的解码器

本文在解码器部分加入了双重注意力模块来获取上下文信息。具体来说,如图2所示,双重注意力模块被嵌入到解码器的第一和第二个上采样模块的输出处。深度估计网络采用跳跃连接方式把浅层特征

并入解码器的特征中,这种嵌入方式可以同时处理深层特征与浅层特征的全局上下文信息建模问题。中间特征经过两种注意力模块处理后通过加法运算合并在一起使特征得到加强,这样比级联更加节省运算成本。与Fu等^[9]不同的是,本文直接将原始特征输入到两种注意力模块中而没有在输入前分别作卷积处理,这样更能保持原特征的有效性。同时,对于高分辨率的特征图,双重注意力模块需要较多的运算开销,因此本文仅在解码器的前两个上采样模块中嵌入双重注意力模块。

2.3 损失函数设计

本文利用合成视图和目标视图之间的差异作为监督信号训练模型,因此图像对比损失函数的设计是重要的一环。由于相机运动中容易受光照影响,因此本文使用文献[4]中的鲁棒性较好的相似度比较函数作为模型的损失函数对视图重建好坏进行判断,即结构相似性(Structural Similarity Index, SSIM)^[11]与L1范数的组合,具体的光度损失函数为

$$L_p = \alpha \frac{1 - \text{SSIM}(\mathbf{I}_t, \hat{\mathbf{I}}_t)}{2} + (1 - \alpha) \|\mathbf{I}_t - \hat{\mathbf{I}}_t\| \quad (6)$$

其中 $\mathbf{I}$ 是真实视图, $\hat{\mathbf{I}}$ 是合成视图, α 是权重参数,这里设为0.85。在图像序列中,使用 $t-1$ 时刻和 $t+1$ 时刻的图像,分别合成 t 时刻图像,根据损失函数即可得到图像对比光度损失。为了减少遮挡和运动物体的影响,本文采用文献[12]中的取不同帧的合成损失的最小值作为最终的损失,即

$$L = \frac{1}{N} \sum_{i=0}^N \min_{t'} L_p(\mathbf{I}_t, \mathbf{I}_{t' \rightarrow t}) \quad (7)$$

这里 L_p 表示式(6)的光度损失函数, N 是总的像素数量。由于双线性插值具有次微分的特性,本文对4种尺度的输出进行损失计算,从而降低其影响。

3 实验结果与分析

本文使用公共数据集KITTI对模型进行训练并验证,同时使用Make3D数据集评估模型跨数据集的泛化能力。

3.1 实验设置

本文使用Pytorch框架对算法进行了实现,并在配备NVIDIA GTX 1080 Ti GPU的计算机上进行训练。数据的分割方面,本文使用文献[13]的数据分割方式。另外,实验遵循文献[5]中的方法预去除静态帧。输入图像的分辨率被缩放到832×256,每次训练

所用数据序列长度为3,共39 810个数据序列用于训练,4 424个用于验证。所有图像数据均使用相同的相机内参,焦距设置为KITTI数据集中所有焦距的平均值。为了增加数据的多样性,数据在训练的过程中进行了随机增强,如亮度、对比度、饱和度的调整 and 水平翻转。模型使用Adam优化算法训练模型,初始学习率设为 10^{-4} ,训练数据的批大小(batchsize)为4,共训练20个周期(epoch)。

3.2 单目深度估计评估

为验证所提方法的有效性,本文将模型在KITTI数据上的验证结果与其他方法进行比较。根据Zhou等^[5]的评估做法,深度值被限制在0.001 m到80 m之间。由于单目深度估计的尺度模糊性,本文方法不能保证产生公制尺度的结果,因此验证时需要把模型估计的深度值乘上尺度 $\delta = \text{median}(\mathbf{D}_{\text{gt}}) / \text{median}(\mathbf{D}_{\text{pred}})$ 。评价指标与文献[5]一致,包括误差指标与准确率指标。误差指标有绝对相对误差(Absolute Relative Difference, Abs Rel)、平方相对误差(Squared Relative Difference, Sq Rel)、均方根误差(Root Mean Squared Error, RMSE)和对数均方根误差(RMSE log)。准确率包含3种阈值的指标,分别为 $\delta < 1.25$ 、 $\delta < 1.25^2$ 和 $\delta < 1.25^3$ 的相应准确率。

表1展示了本文的方法与其他单目深度估计方法各项指标的对比结果。从表1中的实验数据可得出,本文提出的模型单目深度估计性能达到了较高的水平,并且比早期的监督学习方法^[13-14]更好。同时,与其他多任务模型框架^[15-16]相比,本文模型框架更简单且表现更好,只需估计深度和位姿变换,并且只使用了视图重建对比损失而不包含其他损失项。

为了理解模型的各个部分对整体表现的影响,本文进行了消融实验。具体来说,本文分析了输入分辨、编码器预训练以及注意力模块对性能表现的影响,实验结果如表1所示。可以看到,输入图像的分辨率越大,模型的表现越好,同时在ImageNet预训练过的编码器也对提高深度估计的准确度有重要作用。而表1中的实验结果也表明本文所提出的注意力模块的有效性。

为了更直观地对比本文模型与其他方法的深度估计表现,本文提供了部分深度估计的示例,如图4所示。与Zhou等^[5]和Bian等^[17]的工作相比,本文模型方法能更好地保留场景中对象的深度边界,如车、树木、行人和路灯等,灯柱等细长物体也没有出现断裂现象。

表 1 在KITTI数据集上的深度估计实验结果对比

Table 1 Comparison of experimental results of depth estimation on the KITTI dataset

方法	有无监督	误差指标				不同阈值下的准确率		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen ^[13]	有	0.203	1.548	6.307	0.282	0.702	0.890	0.890
Liu ^[14]	有	0.201	1.584	6.471	0.273	0.680	0.898	0.967
Zhou ^[5]	无	0.183	1.595	6.709	0.270	0.734	0.902	0.959
DF-Net ^[15]	无	0.150	1.124	5.507	0.223	0.806	0.933	0.973
Ranjan ^[16]	无	0.148	1.149	5.464	0.226	0.815	0.935	0.973
Bian ^[17]	无	0.128	1.047	5.234	0.208	0.846	0.947	0.976
Godard ^[12]	无	0.115	0.882	4.701	0.190	0.879	0.961	0.982
本文(416×128)	无	0.123	0.990	5.042	0.200	0.861	0.954	0.979
本文(无预训练)	无	0.128	1.047	5.123	0.206	0.855	0.951	0.978
本文(无注意力模块)	无	0.116	0.901	4.802	0.191	0.876	0.958	0.981
本文完整模型	无	0.111	0.860	4.731	0.188	0.884	0.962	0.982

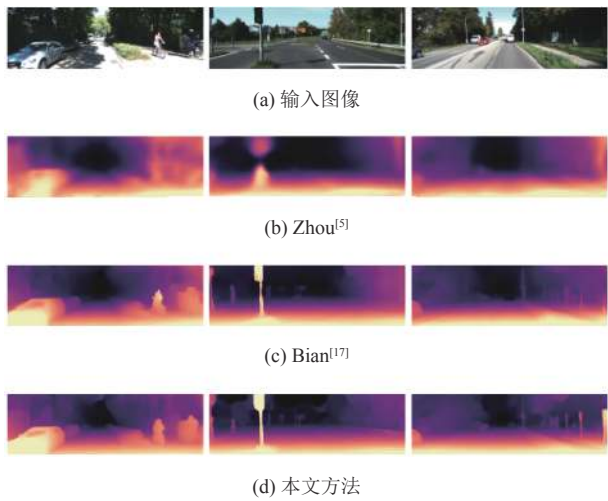


图 4 KITTI数据集上深度估计效果图

Fig.4 Depth estimation results on the KITTI dataset

3.3 模型泛化能力评估

出于验证模型的泛化能力的目的,本文在不对模型参数进行任何调整的情况下用Make3D数据集进行了测试,实验结果如表2所示。同时本文对部分数据进行可视化以便更直观地进行对比,效果图如图5所示。从表2数据以及图5可得出本文模型有较好的跨数据泛化能力,能较好地推断出Make3D数据集的三维场景,但Sq Rel和RMSE两个指标比DF-Net^[15]略差,主要原因是本文只在KITTI数据集上训练,而DF-Net^[15]在Cityscapes和KITTI两个数据集上都进行了训练,本文模型对远处的物体的深度估计偏近所造成的。从可视化的部分实例中同样可以看出本文的模型比Zhou^[5]的泛化能力更强,估计的深度细节更加清晰,但是在远景的部分泛化能力不够好,估计的深度比实际要近。

表 2 在Make3D数据集的泛化性能实验结果

Table 2 Results of generalization performance experiments in the Make3D data set

方法	误差指标			
	Abs Rel	Sq Rel	RMSE	RMSE log
Godard ^[4] (双目)	0.544	10.94	11.760	0.193
Zhou ^[5]	0.383	5.321	10.470	0.478
DF-Net ^[15]	0.331	2.698	6.89	0.416
本文	0.314	3.160	7.397	0.164

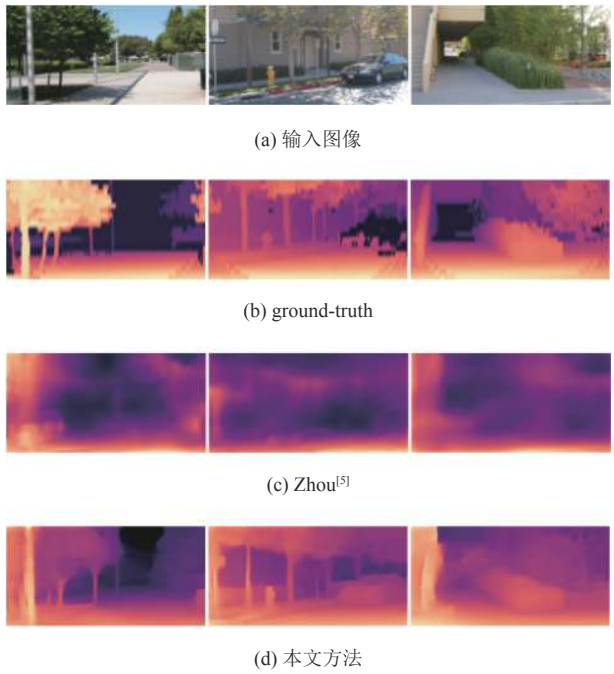


图 5 Make3D数据集深度估计效果图

Fig.5 Depth estimation results on the Make3D dataset

4 结论

本文提出了一种基于双重注意力模块的单目无

监督深度估计方法。方法的核心是基于自注意力的双重注意力模块,这种模块能有效表示远程的上下文信息,帮助网络更容易估计出细节更好的深度信息。本文所提方法模型在KITTI数据集上深度估计的性能达到了较好的效果,在Make3D数据集上的实验也表明模型具有较好的泛化能力。本文方法相比当前其他工作也较为简单,没有复杂的损失函数设计和多任务训练的架构即可达到当前最先进的性能,然而本文所提方法在位姿变换估计方面性能依旧不够好,下一步工作从视频序列信息方面对位姿估计网络进行优化从而改善位姿估计的精度,并且后续工作将引入双目线索增强模型的深度估计精度,增强模型实用性和通用性。

参考文献:

- [1] 朱福利, 曾碧, 曹军. 基于粒子滤波的SLAM算法并行优化与实现[J]. 广东工业大学学报, 2017, 34(2): 92-96.
ZHU F L, ZENG B, CAO J. Parallel optimization and implementation of SLAM algorithm based on particle filter [J]. Journal of Guangdong University of Technology, 2017, 34(2): 92-96.
- [2] 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. Amsterdam: Springer, 2016: 842-857.
- [3] GARG R, BG V K, CARNEIRO G, *et al.* Unsupervised CNN for single view depth estimation: Geometry to the Rescue[C]//European Conference on Computer Vision. Amsterdam: Springer, 2016: 740-756.
- [4] GODARD C, AODHA O M, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]//IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6602-6611.
- [5] 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. Honolulu: IEEE, 2017: 6612-6619.
- [6] MAHJOURIAN R, WICKE M, ANGELOVA A, *et al.* Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5667-5675.
- [7] YIN Z C, SHI J P. GeoNet: unsupervised learning of dense depth, optical flow and camera pose[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1983-1992.
- [8] HUANG J, LEE A B, Mumford D. Statistics of range images[C]//Proceedings IEEE Conference on Computer Vision and Pattern Recognition. Hilton Head Island: IEEE, 2000: 324-331.
- [9] FU J, LIU J, TIAN H, *et al.* Dual attention network for scene segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3146-3154.
- [10] HE K, ZHANG X, REN S, *et al.* Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 770-778.
- [11] 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.
- [12] GODARD C, AODHA O M, BROSTOW G J, *et al.* Digging into self-supervised monocular depth estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 3828-3838.
- [13] EIGEN D, PUHRSCHE C, FERGUS R. Depth map prediction from a single image using a multiscale deep network[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT press, 2014: 2366-2374.
- [14] LIU F, SHEN C, LIN G, *et al.* Learning depth from single monocular images using deep convolutional neural fields [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 38(10): 2024-2039.
- [15] ZOU Y, LUO Z, HUANG J, *et al.* DF-Net: unsupervised joint learning of depth and flow using cross-task consistency[C]//European Conference on Computer Vision. Munich: Springer International Publishing, 2018: 38-55.
- [16] RANJAN A, JAMPANI V, BALLES L, *et al.* Adversarial collaboration: joint unsupervised learning of depth, camera motion, optical flow and motion segmentation[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 12240-12249.
- [17] BIAN J, LI Z, WANG N, *et al.* Unsupervised scale-consistent depth and ego-motion learning from monocular video[C]//Proceedings of the 32th International Conference on Neural Information Processing Systems. Vancouver: MIT press, 2019: 35-45.

(责任编辑: 赵少飞)