基于尺度线索增强的无监督单目深度估计

曲 熠,陈 莹*
(江南大学轻工过程先进控制教育部重点实验室,江苏无锡 214122)

摘 要: 由于单目深度估计中图像与深度图存在一对多的对应关系,单目深度估计本身就存在着尺度歧义的问题. 因此,本文引入基于多视图立体匹配(Multi-View Stereo,MVS)的单目多帧深度估计方法,构造移动深度,挖掘尺度线索,将传统单目深度估计与MVS深度估计有机结合,以改善单目深度估计几何建模中固有的模糊性问题. 在此基础上,设计两个通道注意力模块,分别提高网络的场景结构感知能力和对局部信息的处理能力,从而更充分地融合不同尺度的特征,产生更精确、更清晰的深度预测. 在KITTI数据集的测试结果中,本文方法的平均相对误差和平方相对误差相较基准网络分别最高提升4.7%和8.0%,所有误差和准确率指标均超越其他主流的无监督单目深度估计方法.

关键词: 单目深度估计;无监督学习;深度学习;多尺度;通道注意力
基金项目: 国家自然科学基金(No.62173160)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)09-3217-11
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230767

Unsupervised Monocular Depth Estimation Based on Scale Clue Enhancement

QU Yi, CHEN Ying*
(*Key Laboratory of Advanced Process Control for Light Industry* (*Ministry of Education*), *Jiangnan University*, *Wuxi*, *Jiangsu* 214122, *China*)

Abstract: Due to the relationship of one-to-many between images and depth maps in monocular depth estimation, there is a problem of scale ambiguity in monocular depth estimation itself. In order to improve the inherent ambiguity problem in geometric modeling of monocular depth estimation, this paper introduces a monocular multi-frame depth estimation method based on multi-view stereo (MVS) to construct moving depth and dig the scale clues. The traditional monocular depth estimation and MVS depth estimation are organically combined to improve the inherent ambiguity problem in the geometric modeling of monocular depth estimation. On this basis, two channel attention modules are designed to improve the network's ability to perceive scene structures and process local information, so as to more fully integrate features of different scales and produce more accurate and clearer depth maps.In the test results of the KITTI dataset, the average relative error and square relative error of this paper have been improved by 4.7% and 8.0% respectively compared to the baseline network, with all error and accuracy indicators surpassing other mainstream unsupervised monocular depth estimation methods.

Key words: monocular depth estimation; unsupervised learning; deep learning; multi-scale; channel attention
Foundation Item(s): National Natural Science Foundation of China (No.62173160)

1 引言

深度估计作为计算机视觉中的基本任务,是完成目标分割[1]、自动驾驶[2]等其他视觉任务的必要条件. 虽然利用激光、结构光等三维传感器可以获得准确的深度信息,但是单目深度估计方法只需单张彩色图像就能够推断深度,无需昂贵的硬件和多传感器校准,在设备成本和场景限制上无疑更有吸引力.

相比于需要真实深度标签的有监督单目深度估计方法,近年来,泛化能力更强的无监督单目深度估计方法逐渐成为研究热点. Zhou等[3]提出通过视图重建任务来监督深度估计训练,同时从单目视频中学习深度和相机姿态变化. Godard等[4]使用全分辨率多尺度采样进行预测,通过解耦视差图像和用于计算重投影误差的彩色图像减少视觉伪影. Spencer等[5]提出一种联

收稿日期:2023-08-08;修回日期:2024-06-11;责任编辑:梅志强
*通讯作者:陈莹

合学习深度、密集特征表示和车辆自我运动的框架,在学习到的特征空间中计算特征一致性来监督训练,以提高网络在恶劣的场景条件下的鲁棒性.

由于目前的单帧深度网络没有充分利用相对大小这一判断物体深度的重要线索,导致网络预测的深度图伪影过多.针对这一问题,Johnston等[6]通引入自注意力机制来改善网络可用的上下文信息,通过使用离散的视差量来规范网络训练.叶星余等[7]则将自注意力模块与对抗生成网络(Generative Adversarial Network,GAN)结合,促使深度估计网络建立关联全局信息的长距离、多层级依赖关系.上述方法能够在一定程度上聚合上下文信息,但是对于高级和低级特征的融合都是简单地利用连接和基本卷积来实现,忽略了不同层次特征之间的语义差距.

单目深度估计方法独立地处理每帧图像,但由于单张彩色图片对应的真实场景可能有无数个,而图像中没有稳定的线索来约束这些可能性,没有充分利用连续帧中深度尺度的一致性信息,这使其精度还远不能与三维传感器相媲美.为了提高单目深度估计精度,很多研究者试图利用多帧视频序列中的空间和时间关联改善单目估计方法.Zhang等[8]利用长短期记忆(Long Short-Term Memory,LSTM)网络的优势,提出ST-CLSTM网络,该网络能够捕获视频帧之间的空间特征和时间一致性,并通过生成式对抗性学习方案,进一步加强视频帧之间的时间一致性.Wang等[9]提出了一种基于学习的多视图密集深度图和里程计估计方法,通过多视重投影和前后一致性约束充分利用过去帧的时间信息,进行当前帧深度和相机姿态估计.Wimbauer等[10]依托现有的多目立体视觉技术(Multi-View Stereo,MVS)提出一种半监督单目稠密重建架构,使用结构相似度项(Structural Similarity Index Measurement,SSIM)[11]构建成本量来对连续图像的信息进行编码.同时提出一种动态物体预测掩模以降低移动物体上的伪影对网络的干扰.

基于MVS成本量构建的方法,利用了相机不同时刻拍摄照片之间自然构成的几何关联,从不同视图中找到同一物体匹配的对应点,以同一物体在不同帧中的变化为线索,能够在连续帧中产生几何尺度一致的深度[12],从而实现对深度进行更准确的估计.但其容易受到无纹理区域、反光区域和运动物体的干扰.针对这一问题,Watson等[13]提出了使用知识蒸馏的ManyDepth网络,用单帧深度网络作为教师,多帧深度网络作为学生,利用L1损失构造一致性损失,使学生在不可信区域的输出向教师靠拢.但不可信区域的判断准则仅仅是依靠多帧深度和单帧深度的差异来计算显然不一定准确,所以Feng等[14]提出一种新的动态目标运动解纠缠模块,利用初始深度先验预测来解决最终深度预测中的目标运动失配问题.另外针对基于SSIM构建成本量的传统方法容易产生模糊和局部最小值的问题,Guizilini等[15]提出使用离散化的极线抽样来选择匹配的候选深度值,利用交叉和自注意结合来细化不同视图间的特征匹配.这些方法虽然在一定程度上改善了单目深度估计的尺度模糊问题,但网络结构和训练过程十分复杂,甚至需要额外的预训练分割网络来辅助训练,计算成本巨大.

本文首先设计基于通道注意力的单帧深度估计网络,从强化场景结构和突出关键细节入手,关注在以往方法中常被忽视的特征差异,更有效地融合不同尺度的特征.接着在不过多增加计算成本的前提下,以单帧深度预测结果为基础引导轻量级成本量的构建,在车身速度信息的指导下调整成本量的深度搜索范围,最后设计深度概率自适应策略融合多帧深度和单帧深度结果.通过充分挖掘图像的尺度线索,提高深度估计的准确性.本文提出的主要创新点如下:

(1)在单帧深度网络中设计了基于通道注意力机制的结构增强模块和通道校准模块,帮助网络在复杂的多尺度特征中更有效地捕获整体场景结构信息和关键通道信息,增强网络对整体特征与局部特征的分析能力.

(2)针对原有单帧深度估计方法固有的不适定问题,加入了多帧深度估计方法,利用物体在多帧图像之间的对应关系构成较明确的尺度线索,增强网络对物体尺度的感知能力,最终基于不确定性对单帧和多帧深度估计进行有效融合,得到更精确的深度估计结果.

2 架构及原理

如图1所示,本文的网络主要基于单帧深度估计和多帧深度估计两种方法构建,由单帧深度网络、相机姿态网络以及多帧深度网络三部分组成.

2.1 单帧深度的生成

单帧深度网络使用目前单目深度估计最常用的运动恢复结构(Structure From Motion,SFM)[16]网络架构,由单帧深度网络Θ_S和姿态网络Θ_p两部分组成,均采用编解码结构.输入未标记的单目视频序列中连续的两帧$\boldsymbol{I}_t, t\in\{-1,0\}$,单帧深度网络输出预测的单帧深度图$D_{\text{Mono}}$;姿态网络输出两帧间的相对位姿$[\boldsymbol{R}|\boldsymbol{T}]_{0\to-1}$.

基于移动摄像机和静态场景假设,可以利用前一帧$\boldsymbol{I}_{-1}$重构出当前帧$\boldsymbol{I}_0$,当前帧和重构帧间的差异可以作为监督网络训练的信号[3].重构当前帧的公式如下:

$$\boldsymbol{I}_{-1\to0}(D_{\text{Mono}})=\boldsymbol{I}_{-1}\{\text{proj}(D_{\text{Mono}},[\boldsymbol{R}|\boldsymbol{T}]_{0\to-1},\boldsymbol{K})\} \quad (1)$$

式中,$\boldsymbol{K}$表示已知的相机内部参数;{}表示采样操作;proj表示返回D_{Mono}的二维坐标的投影操作.

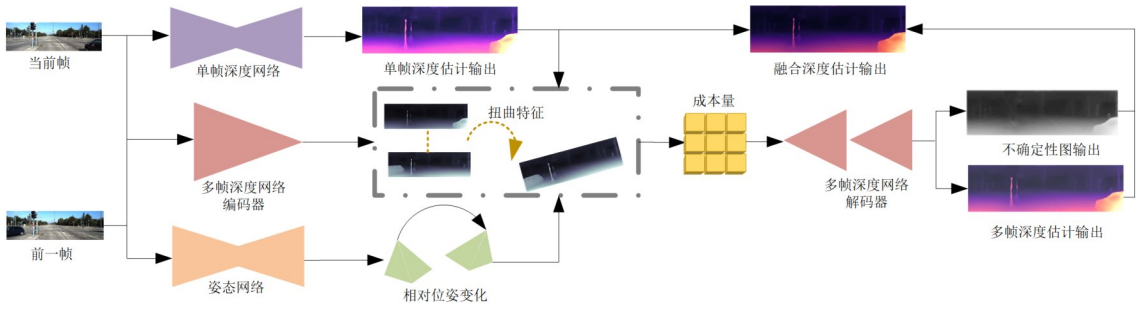


图1　总体框架图

2.2　多帧深度的生成

基于MVS的深度估计方法是根据候选深度将源图像扭曲到参考相机坐标系中构造成本量,并选择成本量的最高活性值作为最终的预测深度[17]. 候选深度需要在固定范围内选择,考虑到连续帧三角优先级[18],本文使用速度引导深度采样策略,通过预测的相机速度动态调整深度搜索范围. 具体来说,当相机移动速度较大时,前后两帧比较符合MVS的三角优先级,本文增加深度搜索范围;反之,如果相机移动速度较慢,前后两帧的几何关联较弱,本文缩小深度搜索范围. 深度图搜索范围的具体计算公式如下:

$$\begin{aligned} d_{\min(x,y)} &= (1-\gamma(\chi\|\boldsymbol{T}\|_2))D_{\text{Mono}(x,y)} \\ d_{\max(x,y)} &= (1+\gamma(\chi\|\boldsymbol{T}\|_2))D_{\text{Mono}(x,y)} \end{aligned} \tag{2}$$

其中,$\chi\|\boldsymbol{T}\|_2$表示预测的相机速度,χ表示相机帧率,$\boldsymbol{T}$表示姿态网络Θ_{P}预测的相机平移;γ表示超参数,$D_{\text{Mono}(x,y)}$表示单帧深度图第(x,y)个像素的对应值.

进而可得候选深度[19]:

$$d_i(x,y)=\frac{1}{\left(\left(\frac{1}{d_{\min}(x,y)}-\frac{1}{d_{\max}(x,y)}\right)\frac{i}{N-1}+\frac{1}{d_{\max}(x,y)}\right)} \tag{3}$$

式中,N为候选深度数,$i=0,1,\cdots,N-1$.

给定当前帧$\boldsymbol{I}_0$及其前一帧$\boldsymbol{I}_{-1}$,首先利用编码器$\Theta_{\text{M-enc}}$提取这些帧的二维特征,得到特征图$\boldsymbol{F}_n\in\mathbb{R}^{W\times H\times C}, n\in\{0,-1\}$. 基于之前的研究[20],利用相机内参$\boldsymbol{K}$和姿态网络$\Theta_{\text{P}}$估计的相对位姿$[\boldsymbol{R}|\boldsymbol{T}]_{0\to-1}$,可以将当前帧的特征图扭曲到前一帧的相机坐标系中得到扭曲特征图$\{\boldsymbol{V}_i\}_{i=0}^{N-1}\in\mathbb{R}^{W\times H\times C\times N}$:

$$\boldsymbol{v}_{t,i}=\boldsymbol{K}(\boldsymbol{R}(\boldsymbol{K}^{-1}\bullet\boldsymbol{p}_t\bullet d_{t,i})+\boldsymbol{T}) \tag{4}$$

其中,$\boldsymbol{p}_t$表示特征图$\boldsymbol{F}_0$中的第t个像素值,$d_{t,i}\in\mathbb{R}^{1\times N}$表示$\boldsymbol{p}_t$的第$i$个候选深度,$\boldsymbol{v}_{t,i}\in\mathbb{R}^{C\times N}$表示根据$d_{t,i}$计算的特征图$\boldsymbol{V}_i\in\mathbb{R}^{W\times H\times C}$中与$\boldsymbol{p}_t$对应的像素值.

接着按通道将特征图$\boldsymbol{V}_{i,i\in[0,N-1]}$、$\boldsymbol{F}_{-1}$均分成$G$组,利用组相关性构建成本量$\boldsymbol{S}$[21],衡量当前帧与前一帧之间的视觉相似性:

$$\boldsymbol{s}_t^g=\frac{1}{G}<\boldsymbol{v}_{t,i}^g,\boldsymbol{f}_t^g> \tag{5}$$

其中,$\boldsymbol{v}_{t,i}^g\in\mathbb{R}^{\frac{C}{G}\times N}$表示根据第$i$个候选深度计算的特征图$\boldsymbol{V}_i$中第$t$个像素的第$g$组特征值;$\boldsymbol{f}_t^g\in\mathbb{R}^{\frac{C}{G}\times 1}$表示特征图$\boldsymbol{F}_{-1}$中第$t$个像素的第$g$组特征值;<,>表示内积;$\boldsymbol{s}_t^g\in\mathbb{R}^{1\times N}$表示最终成本量$\boldsymbol{S}\in\mathbb{R}^{W\times H\times C\times N}$中的第$t$个像素的第$g$组特征值;$G$表示总组数,本文取值为16.

然后通过解码器$\Theta_{\text{M-dec}}$生成不确定性图$\boldsymbol{U}$和多帧深度图D_{MVS}:

$$\boldsymbol{U},D_{\text{MVS}}=\Theta_{\text{M-dec}}(\boldsymbol{S}) \tag{6}$$

并根据单帧深度D_{Mono}、不确定性图$\boldsymbol{U}$和多帧深度D_{MVS}得到融合深度D_{Fuse}:

$$D_{\text{Fuse}}=\boldsymbol{U}\otimes D_{\text{Mono}}+(1-\boldsymbol{U})\otimes D_{\text{MVS}} \tag{7}$$

其中,$\otimes$表示元素级相乘.

3　网络细节

针对多尺度特征融合不充分的问题,本节设计了一种基于通道注意力的单帧深度估计网络,增强网络对于场景结构和关键通道特征的捕获能力,从而使网络得到更准确的相对尺度线索. 同时利用深度概率,将单帧深度估计结果与多帧深度估计结果进行有机结合,保证网络在不同区域预测性能的稳定性.

3.1　基于通道注意力的单帧深度估计网络

如图2所示,本文的单帧深度网络编码器以ResNet[22]作为骨干网络提取语义特征,接着将这些特征输入结构增强模块. 常见的使用通道注意力机制的方法[23,24]通常将其设置在解码器内部,与这些方法不同,本文在编解码器之间加入了通道注意力模块,以捕获更丰富的几何上下文信息,缩小编解码器间的语义差距. 在解码器阶段对编码器输出的特征进行上采样逐步恢复空间分辨率,在此过程中利用多种尺度的通道校准模块来重新调整网络对于不同通道特征的关注度,突出包含物体边缘等关键细节的局部特征. 结构增强模块和通道校准模块结合,使网络更有针对性地关注特征图的通道特征,提取更准确的整体和局部信息,从而提高网络对局部物体在整体场景中相对大小的感知能力. 最后在

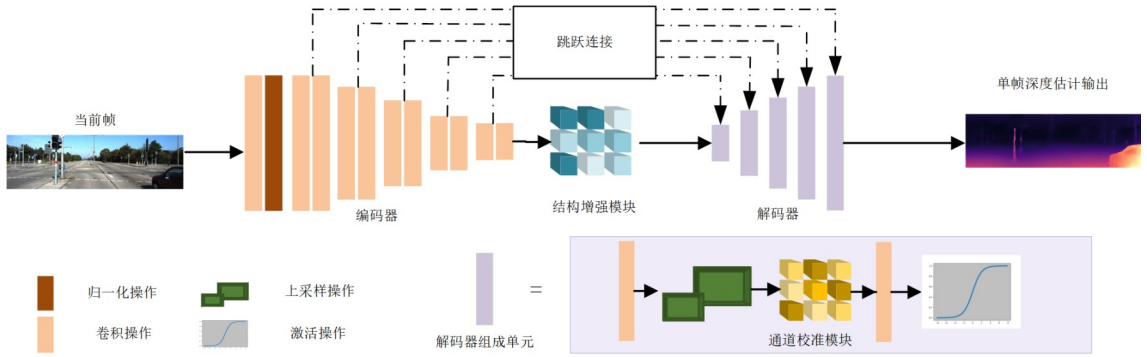


图2 单帧深度网络结构图

多个尺度上使用最近邻插值,对预测的深度图进行上采样到原始输入分辨率,并在该分辨率下计算训练损失.

3.1.1 结构增强模块

在深度估计中,每个高级特征图都可以看作某一个区域的特异性响应,不同的区域响应之间有所关联.每个通道特征图能从其他通道特征图中获得关联响应,捕获更多来自遥远区域的相对深度信息,从而增强对场景结构的感知力[25].因此,本文提出了结构增强模块来感知通道间的依赖关系,聚合不同区域的上下文信息.

首先生成一个注意矩阵,用来建模任意两个通道图间的关系.如图3所示,利用编码器输出的特征图$\boldsymbol{F}$,计算特征相似性图$\boldsymbol{M}$:

$$\boldsymbol{M}_{ij}=\boldsymbol{F}_i\cdot\boldsymbol{F}_j^{\mathrm{T}} \tag{8}$$

其中,$\boldsymbol{F}_i$,$\boldsymbol{F}_j$表示两个不同的通道特征图;$\boldsymbol{F}_j^{\mathrm{T}}$表示的$\boldsymbol{F}_j$转置矩阵.

通道图间的相似性表明区域响应关系,若两个特征图有较高的相似性,说明其对同一区域有较强响应.因此本文通过softmax操作,将相似度$\boldsymbol{M}$转换为注意力图$\boldsymbol{A}$:

$$\boldsymbol{A}_{ij}=\frac{\exp(\max_i(\boldsymbol{M})-\boldsymbol{M}_{ij})}{\sum_{j=1}^{C}\exp(\max_i(\boldsymbol{M})-\boldsymbol{M}_{ij})} \tag{9}$$

其中,C表示通道总数.

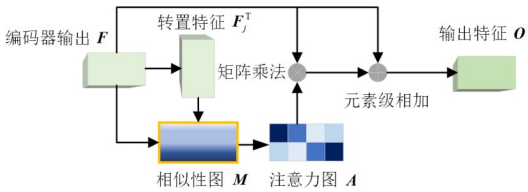


图3 结构增强模块

最后,在特征图$\boldsymbol{F}$和注意力图$\boldsymbol{A}$之间进行元素级的和积运算,得到最终输出$\boldsymbol{O}$:

$$\boldsymbol{O}_i=\sum_{j=1}^{C}(\boldsymbol{A}_{ij}\boldsymbol{F}_j)+\boldsymbol{F}_i \tag{10}$$

其中,$\boldsymbol{O}_i$表示输出的第i个通道图.

与文献[25]直接将编码器最后一层特征输入到结构感知模块的做法不同,本文使用连接策略融合来自所有中间阶段的输出,以此作为结构增强模块的输入特征,在不改变尺度的情况下,本文方法能使更丰富的语义信息参与结构感知.

3.1.2 通道校准模块

解码器在恢复分辨率的过程中,通过跳跃连接将空间细节丰富的低层次特征与全局信息丰富的高层次特征相融合.简单的求和或连接等操作,忽视了不同尺度特征的差异性,将来自不同尺度特征的深度值直接融合,未对不同特征通道进行区分,不能准确表达出原始特征图中的细节信息,妨碍了网络对局部相关性的捕捉,最终导致深度图的细节层次模糊.因此,本文提出了一个通道校准模块,通过一维卷积让高低层次所有特征通道共享权重信息,在避免降维的同时高效捕获局部跨通道信息交互[26],根据不同通道自身的层次特点对其特征的关注重心进行调整,以强调包含更准确深度细节的通道特征,实现对多尺度特征更有效地融合.

如图4所示,先将低级特征$\boldsymbol{F}_{\mathrm{l}}$和高级特征$\boldsymbol{F}_{\mathrm{h}}$通过concat操作相连,然后利用二维卷积层、归一化以及激活操作得到聚合特征$\boldsymbol{F}_{\mathrm{c}}$:

$$\boldsymbol{F}_{\mathrm{c}}=\mathrm{ReLU}(\mathrm{BN}(\mathrm{Conv2}(\mathrm{Con}(\boldsymbol{F}_{\mathrm{l}},\boldsymbol{F}_{\mathrm{h}})))) \tag{11}$$

式中,Con()表示concat操作,Conv2()表示二维卷积,BN()表示批归一化,ReLU()表示ReLU激活操作.

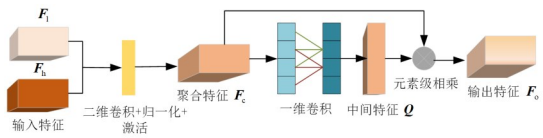


图4 通道校准模块

接着通过全局平均池化压缩$\boldsymbol{F}_{\mathrm{c}}$,并且使用一维卷积来共享相同的学习参数,实现通道之间的信息交互:

$$\boldsymbol{Q}=\mathrm{Sigmoid}(\mathrm{Conv1}(\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{F}_{\mathrm{c}(i,j)})) \tag{12}$$

其中,Sigmoid()表示Sigmoid激活操作,Conv1()表示一维卷积,H,W表示特征图$\boldsymbol{F}_{\mathrm{c}}$的长度和宽度.

$\boldsymbol{Q}$中的权重表明了对应通道的重要性,因此可以利用$\boldsymbol{Q}$自适应地在多个尺度上强调包含关键细节的通道:

$$\boldsymbol{F}_{\mathrm{o}}=\boldsymbol{Q}\otimes\boldsymbol{F}_{\mathrm{c}}+\boldsymbol{F}_{\mathrm{c}} \tag{13}$$

其中,$\otimes$表示元素级相乘,$\boldsymbol{F}_{\mathrm{o}}$表示最终输出特征.

3.2　基于深度概率的多帧深度估计网络

针对基于MVS的多帧深度估计在无纹理区域、反光区域和动态区域精度较差的问题,本文引入基于不确定性的融合方法,通过在融合深度的过程中中分配不同的权重,来得到更可靠的最终深度预测结果.

多帧深度的具体生成过程如图5所示,首先将连续的两帧输入到以特征金字塔网络(Feature Pyramid Network,FPN)[27]为骨干网络的编码器$\Theta_{\text{M-enc}}$中,根据式(4)、式(5)计算出成本量.然后将成本量输入到轻量级成本量解码器$\Theta_{\text{M-enc}}$中得到深度概率$\boldsymbol{P}\in\mathbb{R}^{W\times H\times N}$,接着通过localmax操作[28]得到多帧深度:

$$D_{\text{MVS}}(\boldsymbol{p}_i)=[(\sum_{j=m-r}^{m+r}\boldsymbol{p}_{i,j})^{-1}\sum_{j=m-r}^{m+r}\frac{1}{d_j}\boldsymbol{p}_{i,j}]^{-1} \tag{14}$$

式中,$\boldsymbol{p}_i\in\mathbb{R}^N$表示深度概率$\boldsymbol{P}$的第$i$个像素值;$m=\text{argmax}(\boldsymbol{p}_i)$表示$\boldsymbol{p}_i$取最大值时对应的候选深度索引值;$r$表示半径参数,本文取1.

因为深度概率分布的随机性与MVS的深度不确定性呈正相关[29],所以本文利用深度概率解码器$\Theta_{\text{M-dec-u}}$从深度概率$\boldsymbol{P}$的熵中得到不确定性图$\boldsymbol{U}$:

$$\boldsymbol{U}(\boldsymbol{p}_i)=\Theta_{\text{M-dec-u}}(\sum_{j=0}^{N-1}-\boldsymbol{p}_{i,j}\log\boldsymbol{p}_{i,j}) \tag{15}$$

最终如式(7)所示,将不确定性图$\boldsymbol{U}$作为最终融合单帧深度和多帧深度结果的指导依据.

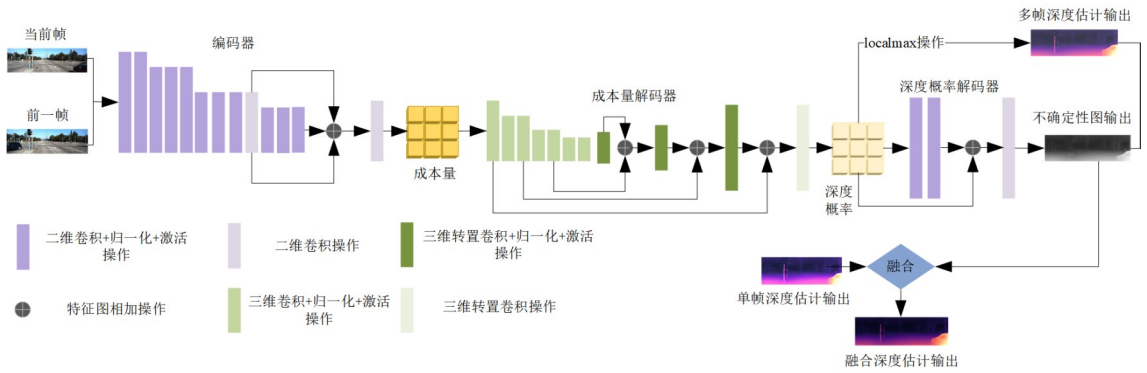


图5　多帧深度网络架构图

3.3　损失函数

本文使用深度平滑损失L_{S}和重投影损失L_{P}对单帧深度网络、姿态网络以及多帧深度网络的训练进行监督.深度平滑度损失L_{S}则用来消除低梯度区域的噪声,其定义为:

$$L_{\mathrm{S}}=|\partial_x d^*|\mathrm{e}^{-|\partial_x \boldsymbol{I}_0|}+|\partial_y d^*|\mathrm{e}^{-|\partial_y \boldsymbol{I}_0|} \tag{16}$$

式中,d^*为预测的深度.

重投影损失L_{P}主要由SSIM和L1损失两部分组成,其作用是计算原当前帧和重构当前帧间的相似度,其定义为:

$$L_{\mathrm{P}}=\alpha\frac{1-\text{SSIM}(\boldsymbol{I}_0,\boldsymbol{I}_{t'\to 0})}{2}+(1-\alpha)|\boldsymbol{I}_0-\boldsymbol{I}_{t'\to 0}| \tag{17}$$

式中,$\boldsymbol{I}_0$为目标帧;$\boldsymbol{I}_{t'\to 0}$为重构目标帧;α为权重系数.

在结合L_{P}和L_{S}时,沿用Monodepth2[4]中提出的最小光度误差策略:

$$L(D)=\min(L_{\mathrm{P}})+\beta L_{\mathrm{S}} \tag{18}$$

式中,β表示L_{P}和L_{S}之间的权重系数.

最终的损失函数由三部分组成:

$$L_{\text{Final}}=L(D_{\text{Mono}})+L(D_{\text{MVS}})+L(D_{\text{Fuse}}) \tag{19}$$

式中,D_{Mono}表示单帧深度估计结果,D_{MVS}表示多帧深度估计结果,D_{Fuse}表示融合的深度估计结果.

4　网络细节实验结果与分析

本节验证了本文提出的模型能够输出尺度准确、细节丰富的深度图,并通过消融实验验证了本文各项改进工作的效果.同时与近年来一些先进的深度估计方法[19,30~36]进行对比实验,具体如下:

Li等[30]和Akada等[31]:均采用有监督方式进行训练.其中Li等[30]将单目深度估计作为一个多类别密集标记任务,通过分层特征融合的方式来实现尺度感知;Akada等[31]提出一种基于无监督域自适应(Unsupervised Domain Adaptation,UDA)的深度估计网络,通过学习域不变性特征,借助合成图像完成深度估计任务.

MOVEDepth[19]:提出了一种改进的多帧深度估计框架,以单帧深度先验以及预测的相机速度为基础,构建轻量级成本量并对其进行回归计算得到深度,减少了多帧匹配的几何模糊性,本文以此作为基准网络.

DIFFNet[32]:设计了一种新的内部特征融合机制,

在上下采样过程中学习语义特征和空间特征的高分辨率表示,以弥合编码器和解码器输出的特征差距.

SGDepth[33]:提出采用预训练的语义分割网络来指导深度估计过程,将无监督方法和有监督方法相结合以高效完成跨域训练.

Chen等[34]:提出了一种基于遮挡感知的蒸馏策略,将从立体模型中学习到的深度知识转移到单目模型中,充分利用立体深度估计算法和单目深度估计算法的互补性提高预测精度.

DynaDepth[35]:提出一种集成视觉信息和惯性测量单元(inertial measurement unit,IMU)信息的尺度感知框架,通过惯性测量单元光度损失和跨传感器光度一致性损失提供密集监督以改善单目深度估计中固有的尺度问题.

SC-DepthV3[36]:引入了一个有监督的单目深度估计预训练模型生成先验深度,同时通过计算自监督训练中的前后向深度不一致性来生成区分动态和静态区域的掩膜,以规范动态区域的深度估计.

4.1 数据集

(1) KITTI数据集[37]是目前深度估计领域的主流评测数据集,包含来自农村、城市、高速公路等多个场景的真实图像数据. 本文使用Eigen等[38]提出的方法,在训练前使用预处理去除静态帧,并将数据集拆分成39 810张训练集,4 424张验证集以及697张测试集.

(2) DDAD[39]数据集是一种新的深度估计基准数据集,相比KITTI数据集,它能够在具有多样化的城市条件下进行更长距离和更密集的深度评估,本文在此数据集上进行泛化性实验以验证所提出模型推广到更多场景的可能性.

4.2 实施细节

本网络在单张RTX3090Ti显卡上进行单帧深度网络模型、姿态网络模型以及多帧深度网络模型的训练,使用Adam优化器[40]进行优化,共训练20个epochs,批量大小为8,输入分辨率为640×192. 前15个epochs的初始学习率设置为10^{-4},剩余5个epochs衰减为初始学习率的0.1倍. 遵循Godard等[4]的思路构建姿态网络模型,将SSIM权值α设为0.85,边缘感知平滑权值β设为10^{-3}. 对于MVS成本量构建,沿用MOVEDepth[19]的设置,将式(2)超参数γ设为0.15,式(3)候选深度数D设为16.

4.3 评估指标

基于之前的研究基础,本文采用如下指标对提出的网络进行量化度量.

平均相对误差Abs Rel:

$$\frac{1}{|T|}\frac{\sum|d-d^*|}{d^*} \tag{20}$$

平方相对误差Sq Rel:

$$\frac{1}{|T|}\frac{\sum\|d-d^*\|^2}{d^*} \tag{21}$$

均方根误差RMSE:

$$\sqrt{\frac{1}{|T|}\sum\|d-d^*\|^2} \tag{22}$$

对数均方根误差RMSE log:

$$\sqrt{\frac{1}{|T|}\sum\|\ln d-\ln d^*\|^2} \tag{23}$$

域值精度满足条件:

$$\delta=\max\left(\frac{d_i}{d_i^*},\frac{d_i^*}{d_i}\right)<\mathrm{thr} \tag{24}$$

式中,d表示真实深度图中某一像素值;d^*表示预测深度图中该像素对应的像素值;T表示像素总数;常用阈值$\mathrm{thr}=1.25,1.25^2,1.25^3$.

4.4 实验结果分析

为对本文提出的模型的进行定量评估,本文在KITTI数据集和DDAD数据集上将测试指标与近来一些先进的单目深度估计方法进行比较. 表中D代表有监督方法,M代表无监督方法;粗体标注每一项指标的最优结果.

从表1中可以看出,在640×192分辨率下,本文模型的所有指标均优于其他近似分辨率甚至更高分辨率的无监督单目深度估计模型,即使与有监督方法相比,本文方法在绝大部分指标上也都能取得最优表现. DepthFormer[15]的网络模型和本文改进的模型均考虑将MVS和注意力机制相结合,同在640×192分辨率下,本文提出的模型在所有指标上均优于前者,其中平均相对误差、平方相对误差和对数均方根误差相较于前者分别提高了10.0%、6.8%和8.5%. 实验结果证明了本文提出的方法误差小,精度高,能够有效提高深度估计质量. 从表2中可以看出,本文模型所有指标仍优于其他先进的无监督单目深度估计模型,绝大多数指标优于使用了有监督预训练的深度估计模型,均方根误差相对基准网络提高了9.8%. 实验结果表明,本文模型能够推广到更具挑战性的场景.

4.5 消融实验

本小节通过消融实验分析了基于不确定性的深度融合方法对最终深度预测结果的影响,并验证本文提出的通道注意力模块对提升预测深度精度的有效性.

4.5.1 深度融合

基于不确定性的深度融合策略可以改善单帧深度估计固有的几何问题,同时弥补多帧深度估计在无纹理区域、反光区域和动态物体区域的不足. 如图6所示,移动的汽车和反光墙面均被视为多帧深度估计高

表1　在KITTI数据集上的测试结果对比

方法	分辨率	监督方式	误差指标(越小越好)				预测准确率(越大越好)		
			Abs Rel	Sq Rel	RMSE	RMSE log	δ<1.25	δ<1.25^2	δ<1.25^3
Li等[30]	620×188	D	0.113	—	4.687	—	0.856	0.962	**0.988**
Akada等[31]	960×288	D	0.168	1.288	5.498	0.235	0.771	0.921	0.973
Monodepth2[4]	640×192	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
DynaDepth[35]	640×192	M	0.109	0.787	4.705	0.195	0.869	0.958	0.981
DIFFNet[32]	640×192	M	0.102	0.764	4.483	0.180	0.896	0.965	0.983
ManyDepth等[13]	640×192	M	0.098	0.770	4.459	0.176	0.900	0.965	0.983
DynamicDepth[14]	640×192	M	0.096	0.720	4.458	0.175	0.897	0.964	0.984
DepthFormer[15]	640×192	M	0.090	0.661	4.149	0.175	0.905	0.967	0.984
MOVEDepth[19]	640×192	M	0.085	0.670	4.266	0.165	0.910	0.967	0.984
Chen等[34]	1 024×320	M	0.094	0.681	4.392	0.185	0.892	0.962	0.981
SGDepth[33]	1 280×384	D+M	0.107	0.768	4.468	0.186	0.891	0.963	0.892
ours	640×192	M	**0.081**	**0.616**	**4.100**	**0.160**	**0.917**	**0.969**	0.985

表2　在DDAD数据集上的测试结果对比

方法	监督方式	误差指标(越小越好)				预测准确率(越大越好)		
		Abs Rel	Sq Rel	RMSE	RMSE log	δ<1.25	δ<1.25^2	δ<1.25^3
Monodepth2[4]	M	0.239	12.547	18.392	0.316	0.752	0.899	0.949
PackNet[39]	M	0.182	7.945	15.021	0.259	0.828	0.925	0.961
MOVEDepth[19]	M	0.136	3.027	12.478	—	0.835	—	—
SC-DepthV3[36]	D+M	0.142	3.031	15.868	0.248	0.813	**0.922**	**0.963**
ours	M	0.134	**2.900**	**11.273**	**0.187**	**0.853**	0.919	0.945

(*a*) 原图

(*b*) 不确定性图

图6　不确定性图可视化结果

度不确定的区域,即单帧深度估计相对确定的区域. 在深度融合过程中,利用不确定性构造深度融合掩膜,可以实现用更可靠的单帧深度估计或多帧深度估计结果补偿另一方估计不准确的部分.

为进一步验证基于不确定性的深度融合策略的先进性,设计如表3所示的消融实验. 其中加权平均融合法采用将单帧深度结果和多帧深度结果直接进行线性加权平均的融合方式,得到融合深度估计结果:

$$D_{\text{Fuse}}=\omega D_{\text{Mono}}+(1-\omega)D_{\text{MVS}} \tag{25}$$

式中,ω取0.5.

从实验结果中可以看出,相对于单独使用单帧深度估计、多帧深度估计以及基于加权平均的融合方法进行预测,基于不确定性图对上述两种估计方法进行融合,平均相对误差最高提升26.3%,平方相对误差最高提升26.5%,对数均方根误差最高提升16.8%. 实验表明,学习深度不确定性并以此为基础进行深度融合,有助于充分发挥单帧深度估计和多帧深度估计的互补性,提高所有深度指标的精度.

4.5.2　通道注意力模块

结构增强模块通过关注通道间的区域响应,在通

表3 验证深度融合作用的消融实验结果

方法	误差指标(越小越好)				预测准确率(越大越好)		
	Abs Rel	Sq Rel	RMSE	RMSE log	δ<1.25	δ<1.25^2	δ<1.25^3
单帧深度估计	0.110	0.839	4.677	0.187	0.886	0.962	0.982
多帧深度估计	0.096	0.718	4.438	0.176	0.899	0.963	0.983
基于加权平均的融合深度估计	0.099	0.748	4.485	0.177	0.899	0.964	0.983
基于不确定性的融合深度估计	0.081	0.616	4.100	0.160	0.917	0.969	0.985

道维度上聚合上下文特征,使每个通道图从非相邻区域得到更多的场景几何信息,从而更好地实现对场景结构的理解. 通道校准模块自适应地选择突出含有丰富细节特征的通道,使解码器在逐步恢复空间分辨率的过程中尽可能保留细节,帮助网络进行更清晰的深度预测.

从表4中可以看出,与基准网络方法相比,使用结构增强模块或通道校准模块,平均相对误差最高提高2.3%,平方相对误差最高提高6.8%,且这两个模块带来的参数量和计算量的增幅很小,对推理时间的影响也微乎其微,以较小的代价实现了精度提升;同时使用结构增强模块和通道校准模块比单独使用两者之一,平均相对误差最高提高2.4%,平方相对误差最高提高2.1%. 实验结果证明,单独使用结构增强模块或通道校准模块均能提升预测精度,且两者结合效果更佳.

表4 验证通道注意力模块各组分性能的消融实验结果

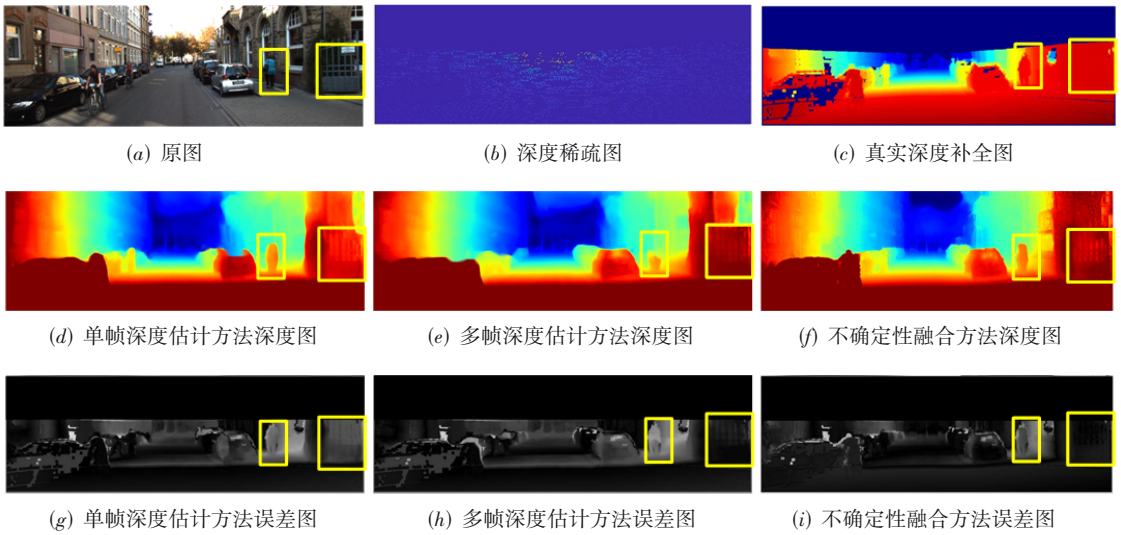
方法	结构增强	通道校准	误差指标(越小越好)				预测准确率(越大越好)			单帧深度网络复杂度指标		
			Abs Rel	Sq Rel	RMSE	RMSE log	δ<1.25	δ<1.25^2	δ<1.25^3	计算量/GB	参数量/MB	推理时间/Fps
基准网络	不包含	不包含	0.085	0.670	4.266	0.165	0.910	0.967	0.984	8.031	14.330	18.991
本文方法	包含	不包含	0.083	0.638	4.180	0.162	0.914	0.968	0.985	8.031	14.330	18.845
	不包含	包含	0.083	0.624	4.117	0.161	0.914	0.968	0.985	9.832	16.177	17.342
	包含	包含	0.081	0.616	4.100	0.160	0.917	0.969	0.985	9.832	16.177	17.113

为更直观地证明通道注意力模块的有效性,将通道注意力模块输入输出特征的通道图映射到RGB颜色空间,并最终投影到原始的RGB图像中生成热力图. 从图7中可以看出,编码器输出特征的注意力主要集中在图片中心的小范围区域(如道路中心、路边的某辆汽车),通过结构增强模块捕获远程依赖关系来聚合场景特征,以及细节强调模块对不同层次特征的关键局部信息的进一步强调,经过通道注意力模块的特征能扩大网络的关注范围,增强对场景整体结构的理解能力,从广阔区域(如路边树木、路边的多辆汽车)获得更多额外的深度感知. 此外,如图7(*c*)所示,通道注意力模块强调了天空这一消失点区域,这是理解场景相对尺度的有力线索,同时网络更专注于关键的前景对象,如汽车等,这有助于其获得更多的深度细节信息.

(*a*) 原图 (*b*) 编码器输出特征 (*c*) 通道注意力模块输出特征

图7 通道注意力模块可视化结果

图8(*f*)、图8(*g*)、图8(*h*)为预测深度图与真实深度补全图的可视化误差,颜色由黑至白表示误差不断增大. 从中可以明显看出,多帧深度估计方法对于栅栏等静态物体的深度预测较为准确,对于运动中的行人的预测准确性较差;而单帧深度估计方法与之相反. 经过不确定性融合策略的深度估计结合了两种方法的长处,无论在静态物体还是动态物体上都取得了较为准确的预测结果.



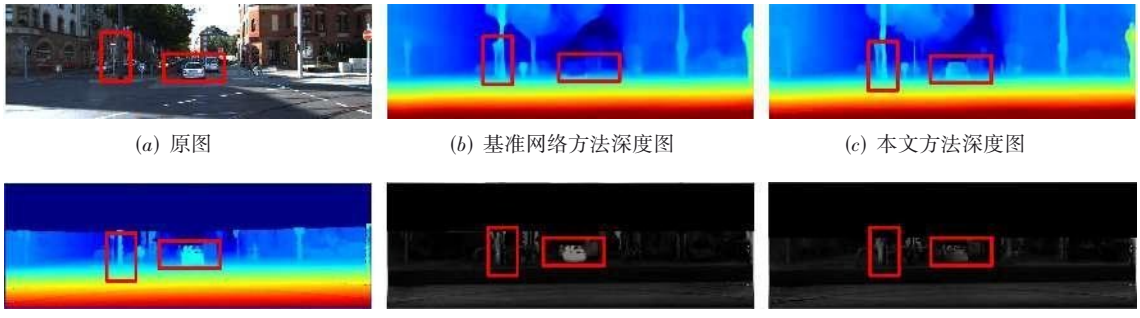
(*a*) 原图　(*b*) 深度稀疏图　(*c*) 真实深度补全图

(*d*) 单帧深度估计方法深度图　(*e*) 多帧深度估计方法深度图　(*f*) 不确定性融合方法深度图

(*g*) 单帧深度估计方法误差图　(*h*) 多帧深度估计方法误差图　(*i*) 不确定性融合方法误差图

图8　不同深度估计方法可视化对比结果

图9(*b*)、图9(*c*)对比可以看出,本文提出的方法对物体深度细节的把握更完整. 从图9(*e*)、图9(*f*)的对比中不难看出,本文方法输出的深度图中能精准地估计出较远距离的汽车深度,而在基准网络方法输出的深度图中该车几乎未被体现,同时本文方法对于距离相近的电线杆与指示牌能够进行更准确的深度估计与区分.



(*a*) 原图　(*b*) 基准网络方法深度图　(*c*) 本文方法深度图

(*d*) 真实深度补全图　(*e*) 基准网络方法误差图　(*f*) 本文方法误差图

图9　本文方法与基准网络可视化对比结果

5　结论

为改善目前无监督单目深度估计方法中存在的尺度模糊问题,本文提出了一种基于尺度线索增强的无监督单目深度估计网络. 该网络以单帧深度预测结果为中心,利用预测的车身速度自适应调整构建成本量的深度范围,最终通过深度概率将单帧深度预测和多帧深度预测进行结合. 同时在网络中插入基于通道注意力的结构增强模块和细节强调模块,通过捕获远程依赖关系来聚合上下文特征,并强调关键的通道特征,实现多尺度特征的充分融合. 在KITTI数据集上的实验证明了与之前的方法相比,本文提出的方法能有效提高输出深度图的准确性,产生更稳健的深度预测结果.

参考文献

[1] 苏天康, 宋慧慧, 樊佳庆, 等. 深度信号引导学习混合变换器的高性能无监督视频目标分割[J]. 电子学报, 2023, 51(5): 1388-1395.
SU T K, SONG H H, FAN J Q, et al. Learning depth signal guided mixed transformer for high-performance unsupervised video object segmentation[J]. Acta Electronica Sinica, 2023, 51(5): 1388-1395. (in Chinese)

[2] KIRAN B R, SOBH I, TALPAERT V, et al. Deep reinforcement learning for autonomous driving: A survey[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(6): 4909-4926.

[3] ZHOU T H, BROWN M, SNAVELY N, et al. Unsupervised learning of depth and ego-motion from video[C]//

2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 6612-6619.

[4] GODARD C, AODHA O MAC, FIRMAN M, et al. Digging into self-supervised monocular depth estimation[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 3827-3837.

[5] SPENCER J, BOWDEN R, HADFIELD S. DeFeat-net: General monocular depth via simultaneous unsupervised representation learning[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 14390-14401.

[6] JOHNSTON A, CARNEIRO G. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 4756-4765.

[7] 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 X Y, HE Y L, RU S N. Unsupervised monocular depth estimation and visual odometry based on generative adversarial network and self-attention mechanism[J]. Robot, 2021, 43(2): 203-213. (in Chinese)

[8] ZHANG H K, SHEN C H, LIi Y, et al. Exploiting temporal consistency for real-time video depth estimation[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 1725-1734.

[9] WANG R, PIZER S M, FRAHM J M. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2019: 5555-5564.

[10] WIMBAUER F, YANG N, VON STUMBERG L, et al. MonoRec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 6112-6122.

[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: A Publication of the IEEE Signal Processing Society, 2004, 13(4): 600-612.

[12] 周晓清, 王翔, 郑锦, 等. 基于自适应空间稀疏化的高效多视图立体匹配[J]. 电子学报, 2023, 51(11): 3079-3091.
ZHOU X Q, WANG X, ZHENG J, et al. Adaptive spatial sparsification for efficient multi-view stereo matching[J]. Acta Electronica Sinica, 2023, 51(11): 3079-3091. (in Chinese)

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

[14] FENG Z Y, YANG L, JING L L, et al. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 228-244.

[15] GUIZILINI V, AMBRUS R, CHEN D, et al. Multi-frame self-supervised depth with transformers[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 160-170.

[16] ULLMAN S. The interpretation of structure from motion [J]. Proceedings of the Royal Society of London. Series B, Biological Sciences, 1979, 203(1153): 405-426.

[17] YAO Y, LUO Z X, LI S W, et al. MVSNet: Depth inference for unstructured multi-view stereo[C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 785-801.

[18] SCHÖNBERGER J L, ZHENG E L, FRAHM J M, et al. Pixelwise view selection for unstructured multi-view stereo[C]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 501-518.

[19] WANG X F, ZHU Z, HUANG G, et al. Crafting monocular cues and velocity guidance for self-supervised multi-frame depth learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(3): 2689-2697.

[20] GU X D, FAN Z W, ZHU S Y, et al. Cascade cost volume for high-resolution multi-view stereo and stereo matching[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 2495-2504.

[21] WANG F, GALLIANI S, VOGEL C, et al. PatchmatchNet: learned multi-view patchmatch stereo[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 14194-14203.

[22] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[23] HWANG S J, PARK S J, BAEK J H, et al. Self-super-

vised monocular depth estimation using hybrid transformer encoder[J]. IEEE Sensors Journal, 2022, 22(19): 18762-18770.

[24] ZHAO C Q, ZHANG Y M, POGGI M, et al. MonoViT: self-supervised monocular depth estimation with a vision transformer[C]//2022 International Conference on 3D Vision (3DV). Piscataway: IEEE, 2022: 668-678.

[25] YAN J X, ZHAO H, BU P H, et al. Channel-wise attention-based network for self-supervised monocular depth estimation[C]//2021 International Conference on 3D Vision (3DV). Piscataway: IEEE, 2021: 464-473.

[26] WANG Q L, WU B G, ZHU P F, et al. ECA-net: Efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 11534-11542.

[27] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 2117-2125.

[28] WANG F, GALLIANI S, VOGEL C, et al. IterMVS: Iterative probability estimation for efficient multi-view stereo [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 8606-8615.

[29] ZHANG J Y, LI S W, LUO Z X, et al. Vis-MVSNet: Visibility-aware multi-view stereo network[J]. International Journal of Computer Vision, 2023, 131(1): 199-214.

[30] LI B, DAI Y C, HE M Y. Monocular depth estimation with hierarchical fusion of dilated CNNs and soft-weighted-sum inference[J]. Pattern Recognition, 2018, 83: 328-339.

[31] AKADA H, BHAT S F, ALHASHIM I, et al. Self-supervised learning of domain invariant features for depth estimation[C]//2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2022: 3377-3387.

[32] ZHOU H , GREENWOOD D , TAYLOR S. Self-supervised monocular depth estimation with internal feature fusion[C]// The 32nd British Machine Vision Conference. Durham: BMVA, 2021: 378-391.

[33] KLINGNER M, TERMÖHLEN J A, MIKOLAJCZYK J, et al. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance [C]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 582-600.

[34] CHEN Z, YE X Q, YANG W, et al. Revealing the reciprocal relations between self-supervised stereo and monocular depth estimation[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2021: 15529-15538.

[35] ZHANG S, ZHANG J, TAO D C. Towards scale-aware, robust, and generalizable unsupervised monocular depth estimation by integrating imu motion dynamics[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2022: 143-160.

[36] SUN L B, BIAN J W, ZHAN H Y, et al. SC-DepthV3: Robust self-supervised monocular depth estimation for dynamic scenes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(1): 497-508.

[37] 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.

[38] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network [J]. Advances in Neural Information Processing Systems, 2014, 3(January): 2366-2374.

[39] GUIZILINI V, AMBRUS R, PILLAI S, et al. 3D packing for self-supervised monocular depth estimation[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 2485-2494.

[40] KINGMA D P, BA J. Adam: A method for stochastic optimization[EB/OL]. (2017-01-30) [2023-06-08]. https://arxiv.org/abs/1412.6980.

作者简介

曲　熠　女,1999年6月出生于山东省烟台市.2021年获江南大学学士学位,现为江南大学控制科学与工程专业硕士研究生.主要研究方向为计算机视觉与模式识别.
E-mail: 6211905004@stu.jiangnan.edu.cn

陈　莹　女,1976年11月出生于浙江省丽水市.2005年于西安交通大学获得博士学位,现为江南大学物联网工程学院教授.主要研究方向为机器视觉、信息融合、模式识别.
E-mail: chenying@jiangnan.edu.cn