文章编号:1674-2974(2024)08-0001-12 DOI:10.16339/j.cnki.hdxbzkb.2024273

基于语义辅助和深度时序一致性约束的自监督单目深度估计

凌传武[1],陈华[2],徐大勇[3],张小刚[1†]
(1. 湖南大学 电气与信息工程学院,湖南 长沙 410082;
2. 湖南大学 信息科学与工程学院,湖南 长沙 410082;
3. 中国烟草总公司郑州烟草研究院,河南 郑州 450000)

摘 要:通过使用相邻帧之间的光度一致性损失代替深度标签作为网络训练的监督信号,基于图像序列训练的自监督单目深度估计方法近年来受到了广泛的关注.光度一致性约束遵循了静态世界假设,而单目图像序列中存在的运动目标违反该假设,进而影响自监督训练过程中相机位姿估计精度和光度损失函数的计算精度.通过检测并移除运动目标区域,可在得到与目标运动解耦的相机位姿的同时,消除运动目标区域对光度损失计算精度的影响.为此,本文提出了一种基于语义辅助和深度时序一致性约束的自监督单目深度估计网络.首先,使用离线的实例分割网络检测可能违反静态世界假设的动态类别目标,并移除对应区域输入位姿网络从而得到与物体运动解耦的相机位姿.其次,基于语义一致性和光度一致性约束,检测动态类别目标的运动状态,使得运动区域的光度损失不影响网络参数的迭代更新.最后,在非运动区域施加深度时序一致性约束,显式对齐当前帧的估计深度值与相邻帧的投影深度值,进一步细化深度预测结果.在KITTI、DDAD以及KITTI Odometry数据集上的实验验证了所提方法与以往的自监督单目深度估计方法相比具有更出色的性能表现.

关键词:单目深度估计;自监督学习;运动目标;时序一致性
中图分类号:TP391.4 **文献标志码**:A

Self-supervised Monocular Depth Estimation Based on Semantic Assistance and Depth Temporal Consistency Constraints

LING Chuanwu[1], CHEN Hua[2], XU Dayong[3], ZHANG Xiaogang[1†]
(1. College of Electrical and Information Engineering, Hunan University, Changsha 410082, China;
2. College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China;
3. Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou 450000, China)

Abstract: Self-supervised monocular depth estimation methods trained on sequences of monocular images have received considerable attention in recent years by using the photometric consistency loss between adjacent frames

* **收稿日期**:2023-08-24
基金项目:国家自然科学基金资助项目(62171184, 62273139, 62106072), National Natural Science Foundation of China(62171184, 62273139, 62106072);国家自然科学基金区域联合重点项目(U23A20385), Joint Funds of the National Natural Science Foundation of China(U23A20385);国防预研项目(JCY2021206B015), National Defense Pre-research Foundation(JCY2021206B015)
作者简介:凌传武(1996—),男,湖南娄底人,湖南大学博士研究生
† 通信联系人,E-mail:zhangxg@hnu.edu.cn

instead of depth labels as the supervisory signal for network training. The photometric consistency constraint follows the static world assumption, but the moving objects in the monocular image sequence violate this assumption, which affects the camera pose estimation accuracy and the calculation accuracy of the photometric loss function during the self-supervised training process. By detecting and removing the moving target area, the camera pose decoupled from the target motion can be obtained, and the influence of the moving target area on the calculation accuracy of the photometric loss can be discarded. To this end, this paper proposes a self-supervised monocular depth estimation network based on semantic assistance and depth temporal consistency constraints. First, an offline instance segmentation network is used to detect dynamic category objects that may violate the static world assumption, and the corresponding region input pose network is removed to obtain a camera pose decoupled from object motion. Secondly, based on semantic consistency and photometric consistency constraints, the motion status of dynamic category targets is detected so that the photometric loss in the moving area does not affect the iterative update of network parameters. Finally, depth temporal consistency constraints are imposed in non-motion areas, and the estimated depth value of the current frame is explicitly aligned with the projected depth value of adjacent frames to further refine the depth prediction results. Experiments on the KITTI, DDAD and KITTI Odometry datasets verify that the proposed method has better performance than previous self-supervised monocular depth estimation methods.

Key words: monocular depth estimation; self supervision learning; moving object; temporal consistency

从图像中恢复深度信息是计算机视觉领域的一个基础问题,可被广泛应用于三维重建、视觉导航、虚拟现实等任务中.与基于立体匹配[1]或基于多视角的方法[2]相比,单目深度估计方法因其更少的硬件限制和高效的运行速率而具有更广泛的应用前景[3].

早期的单目深度估计大多基于传统机器学习算法,可分为参数学习和非参数学习方法.参数学习方法中需要预先设定参数模型,通过提取形状[4]、消隐点、散焦模糊[5]等线索进行推断,或构造马尔可夫模型[6]拟合图像特征与深度值之间的映射关系来计算深度.非参数学习方法在已有数据集中进行相似性检索得到相似图像集[7],通过相似图像集的深度标签推断输入图像的深度值.这些基于传统机器学习算法的深度估计算法有其应用限制:参数学习方法需要输入图像和深度值来满足设定的参数模型,对真实场景预测精度低;非参数学习方法依赖图像检索和后处理,计算量大导致难以应用于实际.随着深度学习技术的发展,许多基于深度神经网络的方法获得了出色的性能.早期的基于神经网络的研究主要采用基于监督学习的方法,采用深度卷积神经网络设计不同的网络结构[8]或损失函数[9],通过对输入图像与对应深度标签之间的映射关系进行建模来估计深度.有监督的深度估计方法在估计精度上获得了出色的性能,但因其需要收集并构建大量高质量标签数据集用于训练,方法的实用性也受到一定限制.

近期研究者们基于静态世界假设和图像几何投影模型,提出了大量的无监督/自监督单目深度估计方法.根据训练所需数据类型的差异,自监督单目深度估计方法可分为基于双目对训练和基于单目视频序列训练两类,采用双目对[10]或相邻视频帧[11]之间的光度一致性约束取代深度标签作为网络训练的隐式监督信号.基于双目对的方法建模时只需考虑校正后左右视图像素的对应匹配关系,因而建模相对较为简单且精度较高,但其需要相机标定参数和双目数据集用于训练,对数据采集系统的稳定性和标定精度有较高要求.基于视频序列训练的方法因其所需训练数据相对更易获取,可应用范围更广,目前受到研究者的广泛关注,本文亦主要研究基于视频序列训练的自监督单目深度估计方法.不同于基于双目对训练的建模,基于视频序列的方法建模时需要同时考虑场景结构和相机运动,这给单目深度估计任务带来了许多挑战.

基于单目视频的自监督深度估计方法训练时,使用的图像几何投影模型遵循静态世界假设,该假设下场景中的物体相邻帧之间的位移仅由相机运动

产生.但在真实场景的视频序列中,运动目标例如车辆、行人等不满足该假设.对这些运动目标区域采用光度一致性作为约束训练网络估计深度时,精度会下降.准确提取运动目标区域,消除其对优化网络参数的影响对提升深度估计的精度非常重要.为解决这一问题,现有一些研究主要通过引入光流信息[12]或语义感知信息[13]剔除动态类别目标,或改进图像投影模型来拓展处理动态类别目标.然而在一些情况下,动态类别目标可能并未发生移动,例如停放在路边的车辆.如果将这些区域直接剔除,则造成了网络可用信息的丢失;如果对这些区域扩展图像投影模型,可能会不必要地增加网络的复杂度.因此,如何准确提取运动目标区域成了一个重要挑战.

另一个挑战是由单目视觉固有尺度缺失造成的相邻视频帧深度估计结果尺度不一致问题.在视频处理领域中,视频语义分割、图像风格迁移等任务都需要考虑到相邻帧输出结果在时序上的一致性和平滑性.对于深度估计任务来说,同一个三维空间点在不同帧上的投影像素点应具有满足几何投影关系的深度值.为解决这一问题,一些工作通过引入长短期记忆网络LSTM[14]、光流网络[15]和后处理技术[16]来加强相邻帧预测结果的时序一致性,这些方法具有较高的计算需求,难以实现在线运行.

为了解决上述问题,本文提出了一种基于语义辅助和时序一致性约束的自监督单目深度估计网络.首先,利用离线的实例分割网络检测出输入图像中的动态类别实例.将移除这些区域的图像作为自监督训练框架中位姿网络的输入,消除潜在运动目标对相机位姿估计的影响.其次,提出了一种基于语义和光度一致性的实例运动状态的检测方法,准确提取运动目标区域并改进光度损失函数.再次,在非运动区域,对当前视图的预测深度和相邻视图的投影深度显式施加时序一致性约束,进一步细化深度输出结果.最后,在公开数据集上进行了相关的定量和可视化的实验验证.实验结果表明,所提方法与以往的自监督工作相比在各项评测指标上具有更优异的性能表现.

1　本文的自监督单目深度估计

本节将先介绍单目深度估计的自监督训练框架及流程;再介绍各部分网络的基本结构;最后对训练过程中的损失函数进行详细介绍.

1.1　自监督框架概览

自监督单目深度估计任务的目标是在没有深度真值用于训练的情况下为独立的目标图像帧估计对应的稠密深度图.所提自监督单目深度估计方法训练流程如图1所示,基于单目序列训练的自监督框架主要包括一个位姿估计网络和一个深度估计网络.

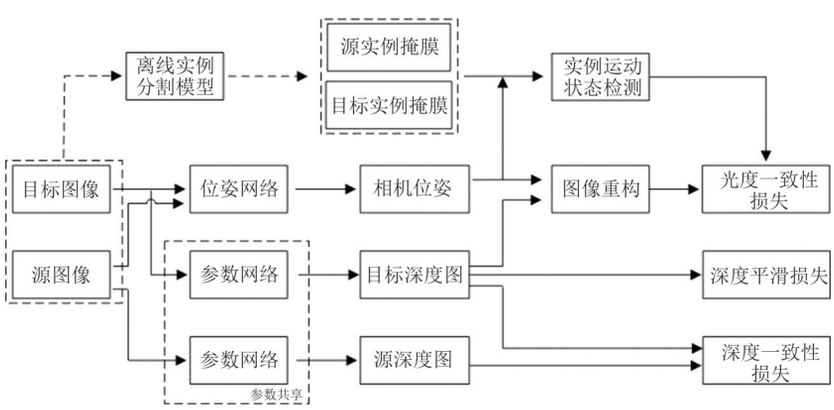


图1　训练流程图

Fig. 1　Flowchart of the training process

在训练阶段,给定单目训练序列中的目标帧图像,将相邻帧的图像作为源图像.目标图像会与源图像一同输入到位姿估计网络中从而得到输入图像间的相机位姿.此外,目标图像与源图像会分别输入到深度估计网络中从而得到目标深度图和源深度图.在得到相机位姿以及目标深度图后,使用图像重构模型从源图像中生成重构的目标图像.重构目标图像与输入目标图像之间计算光度一致性损失作为总体损失的一部分.为了克服运动目标干扰问题,使用离线的实例分割模型生成了对应源图像和目标图像

的实例掩膜.根据对应掩膜区域的语义和光度损失实现实例的运动状态检测,并在光度一致性损失中移除运动的实例、保留静止的实例.为了加强帧间深度估计的尺度一致性,在目标深度图和源深度图之间对应的非运动目标区域施加了深度时序一致性损失,进一步细化了深度估计结果.在测试阶段,只需将待测图像输入训练好的深度网络中即可得到对应的深度输出.

图2给出了所提自监督单目深度估计方法的网络框架图.给定包含n帧的连续图像序列,记为<I_1,I_2,…,I_n>.与以往的工作类似,将训练数据集中每三个连续的图像<I_{t-1},I_t,I_{t+1}>分割出来作为一个训练样本,其中I_t被定义为目标图像,其余的图像定义为源图像.为了简化描述,将I_{t-1}和I_{t+1}都记为I_s.如图2所示,目标图像和源图像被输入到深度估计网络中从而得到对应的深度图D_t和D_s.预测的目标深度图D_t可以将源视图投影到目标视图.这一投影过程遵循静态世界假设并且需要相机内参$\boldsymbol{K}$和外参$\boldsymbol{T}_{t\to s}$.在训练过程中,相机内参被设定为已知且在整个数据集上都为恒定值.位姿网络输出具有6自由度的外参$\boldsymbol{T}_{t\to s}$,即I_t和I_s之间的相对位姿.在得到深度估计图D_t和位姿$\boldsymbol{T}_{t\to s}$后,将源视图I_s投影到目标视图上从而生成重构的目标视图I_t',如下式所示:

$$P_s \sim K T_{t\to s} D_t(P_t) K^{-1} P_t \tag{1}$$

$$I_t' = \mathrm{bil}\left(I_s(P_s)\right) \tag{2}$$

式中:P_t是重构图像中像素坐标的齐次坐标.由于计算得到的P_s是连续值,而像素坐标P_t是离散值,因此使用了可微分的双线性采样bil(.)操作[17]来得到I_t'对应的像素值.自监督单目深度估计的核心是使用真实目标图像I_t和重构目标图像I_t'之间的光度损失用来训练网络.

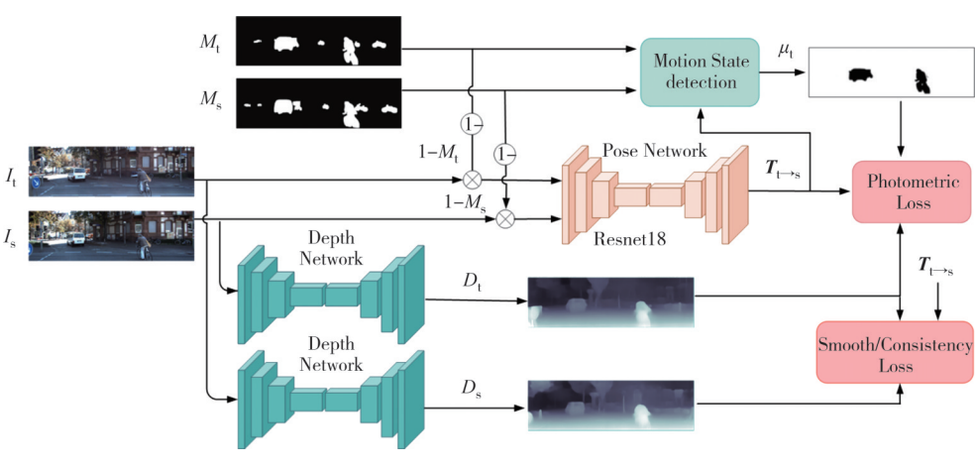


图2 网络框架图

Fig. 2 The network framework diagram

1.2 网络结构

深度估计网络将单个RGB图像作为输入,输出对应的稠密深度图.对于这一稠密预测任务,采用了全卷积架构将深度估计网络建模为类似编码-解码结构.编码器集成了移除最后一个全连接层的Resnet50网络.解码器将编码器的输入逐渐上采样到与输入图像相同的分辨率.在编码器和解码器间加入的跳过连接使得解码器可以解析更高分辨率的细节.

位姿网络按照顺序接收两个RGB图像I_t和I_s作为输入,并预测它们的相对位姿$\boldsymbol{T}_{t\to s}$.考虑到相对位姿只有6个自由度,与以往的工作类似使用了一个较浅的Resnet18模型用来提取特征,随后使用多个卷积层来回归6自由度的位姿.

为了解决运动目标问题,采用了一个离线的实例分割网络Mask RCNN[18]来检测可能违反静态世界假设的动态类别目标.本文主要关注自动驾驶场景,因此动态类别目标主要设定为人、车辆.分割网络会输出二进制语义掩膜M来指示对应位置的像素是否属于动态类别.

1.3 损失函数

1.3.1 光度一致性损失

基于目标视图深度D_t和相机位姿$\boldsymbol{T}_{t\to s}$,将源图像I_s向目标视图投影可以生成重构的目标图像I_t'.目标

图像I_t和重构的目标图像I'_t之间的光度一致性损失表示为

$$L_p^t = \frac{1}{N}\sum_{i=1}^{N}(1-\alpha)\left\|I_t(i)-I'_t(i)\right\| + \alpha\frac{1-\mathrm{SSIM}\left(I_t(i),I'_t(i)\right)}{2} \tag{3}$$

式中:N是投影之后有效像素的个数;系数α设置为0.85.

光度一致性损失包含L1损失和结构相似性损失SSIM[19],L1损失惩罚图像间的像素颜色不一致,SSIM损失归一化以单个像素为中心的图像块的像素照度,可以用于处理复杂的照明变化情况,滤波核大小设置为3×3.

上述描述的光度损失是基于从源图像到目标图像的几何变换,可以很容易地推广到相反的方向.具体来说,将I_s和I_t按顺序输入位姿网络时可以得到从源视图到目标视图的相对位姿$\boldsymbol{T}_{s\to t}$.使用预测的源图像深度D_s,使用式(1)和式(2)可以得到重构的源视图图像I'_s.真实源视图和重构源视图之间的光度一致性损失可以表示为

$$L_p^s = \frac{1}{N}\sum_{i=1}^{N}(1-\alpha)\left\|I_s(i)-I'_s(i)\right\| + \alpha\frac{1-\mathrm{SSIM}\left(I_s(i),I'_s(i)\right)}{2} \tag{4}$$

1.3.2　深度平滑损失

为了保持深度图的局部平滑性,本文应用了边缘感知的平滑损失[20]如下式所示:

$$L_{ds}^t = \frac{1}{N}\sum_{i=1}^{N}\nabla_x D_t(i)\cdot e^{-\left|\nabla_x I_t(i)\right|} + \nabla_y D_t(i)\cdot e^{-\left|\nabla_y I_t(i)\right|} \tag{5}$$

式中:∇_x和∇_y分别是沿着图像的宽度和高度方向的一阶梯度.

与1.3.1节类似,对于源图像深度图D_s,对应的深度平滑损失如下:

$$L_{ds}^s = \frac{1}{N}\sum_{i=1}^{N}\nabla_x D_s(i)\cdot e^{-\left|\nabla_x I_s(i)\right|} + \nabla_y D_s(i)\cdot e^{-\left|\nabla_y I_s(i)\right|} \tag{6}$$

1.3.3　改进的光度一致性损失

实例分割网络的输出$M_{s/t}$包含所有的动态类别目标:

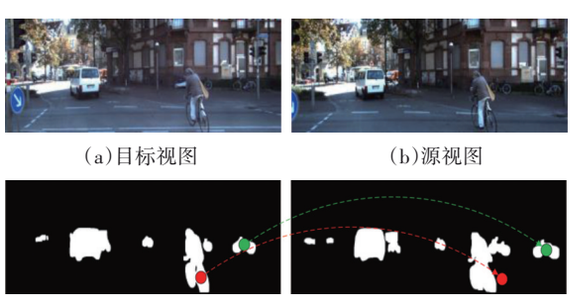
$$M_{s/t} = \bigcup_{n\in(1,2,\cdots)} M_{s/t}^{\mathrm{index}=n} \tag{7}$$

式中:n是属于动态类别的实例的个数.为了应对运动目标问题,一种简单的策略是在计算光度损失时,排除所有的动态类别目标,如下式所示:

$$L_{mp}^{s/t} = \left(1-M_{s/t}\right)\cdot L_p^{s/t} \tag{8}$$

然而,静止的动态类别目标区域是遵循光度一致假设的,例如KITTI数据集中一些训练样本中包含许多静止停放的车辆且占据图像较大区域.简单排除这些区域的光度损失使得网络丢失了监督信号,进而提高了训练过程的风险.为了解决这一问题,利用源视图和目标视图中实例掩膜的投影对齐来判断动态类别目标的运动状态,这一过程的示意图如图3所示.

在静态世界假设下,相同空间点在相邻视图下投影的像素点不仅应具有相同的颜色值,也应具有相同的语义标签.如图3所示,相邻帧之间的自行车手处于运动状态,而停靠的自行车处于静止状态.自行车手对应的语义标签在经过视图投影之后会出现偏差,表现为图中的红点.而静止的自行车投影后的语义标签会与投影之前保持一致.基于这一观察,在给定语义标签的前提下,通过源视图和目标视图之间的语义标签的投影一致性可以从语义掩膜辨别出动态类别目标的运动状态.



(a)目标视图　(b)源视图　(c)目标视图掩膜　(d)源视图掩膜

图3　语义一致性的示意图

Fig. 3　An illustration of the semantic consistency

假设源视图语义掩膜M_s,预测位姿为$T_{t\to s}$,应用公式(1)可得到重构目标掩膜中像素坐标p_t在源视图中的匹配坐标p_s.在生成重构掩膜的过程中,采用的插值函数为最近邻采样near(.).对于每一个实例,重构掩膜计算如下:

$$\hat{M}_t^{\mathrm{index}=n} = \mathrm{near}\left(M_s^{\mathrm{index}=n}\left(p_s\right)\right), n\in(1,2,\cdots) \tag{9}$$

为了简化推导,在接下来的内容中省略实例索引index.对于一个特定的实例,如果M_t和$\hat{M}_t$有较大

的重叠范围,对应的目标就很可能是静态目标.与Dyna-DM[21]类似,使用Dice系数作为衡量两个掩膜重叠度的指标.

$$\mathrm{Dice}\left(M_{\mathrm{t}},\hat{M}_{\mathrm{t}}\right)=\frac{2\left|M_{\mathrm{t}}\cap\hat{M}_{\mathrm{t}}\right|}{\left|M_{\mathrm{t}}\right|+\left|\hat{M}_{\mathrm{t}}\right|}<\theta_1 \tag{10}$$

在上述指标的基础上,本文利用投影前后实例掩膜在对应图像区域的光度一致性构建了约束指标MPC,计算如下:

$$\mathrm{MPC}=L_{\mathrm{p}}^{\mathrm{t}}\cdot\left(M_{\mathrm{t}}\cup\hat{M}_{\mathrm{t}}\right)>\theta_2 \tag{11}$$

上式进一步约束了静态目标的掩膜重叠区域要有较小的光度损失.如果满足式(10)和式(11)中的任意一个条件,对应的实例都会被归类为运动目标,因此可以得到运动的动态类别目标掩膜λ为

$$\lambda=\begin{cases}1, \mathrm{Dice}<\theta_1 \text{ 或 } \mathrm{MPC}>\theta_2\\ 0, \text{其他}\end{cases} \tag{12}$$

因此,对应的非运动(静态)区域掩膜μ为

$$\mu(i)=1-\lambda(i) \tag{13}$$

1.3.4 时序一致性损失

为了保持相邻帧之间估计的深度图的尺度一致性,本文显式施加了深度时序一致性损失.深度时序一致性指的是目标视图的深度图与从源图像投影生成的重构目标视图的深度图应该保持一致.需要注意的是,此处重构目标视图的深度是投影深度,并不是源视图的估计深度.定义重构目标深度图中的齐次像素坐标为P_{t},使用式(1)可以得到在源视图下的三维点P_{s}.P_{s}中的z分量是通过投影目标视图中像素深度值到源视图后得到的投影深度值.P_{s}中的投影坐标即x和y分量也是连续值,因此也需要在网络估计的源视图深度图D_{s}上进行双线性插值得到连续坐标下的D'_{s}.定义投影深度图为$D_{\mathrm{s}}^{\mathrm{p}}$,该深度图应与插值后的深度图保持一致.深度一致性损失定义为:

$$L_{\mathrm{c}}^{\mathrm{s/t}}=\frac{1}{N}\sum_{i=1}^{N}\left\| D_{\mathrm{s/t}}^{\mathrm{p}}(i)-D'_{\mathrm{s/t}}(i)\right\| \tag{14}$$

光度一致性损失和深度时序一致性损失基于静态假设而设计,因此仅适用于图像中的非运动区域.平滑损失则可以在全图范围进行约束.

当相机静止或者物体与相机运动方向一致时,深度值无法根据光度损失来优化.为了应对这一问题,使用自掩膜S来区分有效像素.

$$S(i)=\begin{cases}1, \text{当 } \|I_{\mathrm{t}}(i)-I'_{\mathrm{t}}(i)\|<\|I_{\mathrm{t}}(i)-I_{\mathrm{s}}(i)\|\\ 0, \text{其他}\end{cases} \tag{15}$$

即只有在投影之后的像素误差值小于投影之前的误差值时,对应像素才被认为是有效像素.

最终的损失函数包括光度一致性损失、深度平滑损失、时序一致性损失三部分,并由静态掩膜μ和自掩膜S加权.

$$L_{\mathrm{total}}=S\cdot\mu\cdot\left[w_1\left(L_{\mathrm{p}}^{\mathrm{t}}+L_{\mathrm{p}}^{\mathrm{s}}\right)+w_2\left(L_{\mathrm{c}}^{\mathrm{t}}+L_{\mathrm{c}}^{\mathrm{s}}\right)\right]+w_3\left(L_{\mathrm{ds}}^{\mathrm{t}}+L_{\mathrm{ds}}^{\mathrm{s}}\right) \tag{16}$$

2 实验验证

2.1 数据集

KITTI[22]是一个主要关注自动驾驶场景的数据集.它包含数十小时的交通场景,使用的数据采集系统包括彩色相机、灰度相机、3D激光扫描仪和GPS定位系统.KITTI数据集的原始图像尺寸是1 242×375像素.本文使用了与Zhou等的工作[11]相同的训练集划分,并对静态帧进行了移除.本文采用了与Zhou等的工作相同的训练集,移除了静态帧,并在包含697张图片的Eigen测试集上对模型的性能进行了评估.

DDAD数据集包含不同国家的城市场景下的200段驾驶视频.与KITTI数据集相比,该数据集中包含更多的动态目标,绝大部分车辆都在路上行驶.本文采用了标准的训练集和测试集划分,其中150个场景用于训练,50个场景用于验证,测试集包含3 950张图像.

KITTI Odometry[23]数据集包含了11个具有真值轨迹的视频序列,该数据集可用来评估位姿估计的性能.本文使用了数据集中的序列00~08作为训练集,序列09~10作为测试集.

2.2 评测指标

本文使用了Eigen等[8]提出的评估指标来评测所提方法在深度估计上的精度.评估指标包括4个误差指标(越小越好)和3个精度指标(越大越好).这些指标如下所示.

$$\text{Abs Rel: } \frac{1}{N}\sum_{i=1}^{N}\left\| D_{\mathrm{gt}}^{i}-\hat{D}^{i}\right\|/D_{\mathrm{gt}}^{i};$$

$$\text{Sq Rel: } \frac{1}{N}\sum_{i=1}^{N}\left\| D_{\mathrm{gt}}^{i}-\hat{D}^{i}\right\|^2/D_{\mathrm{gt}}^{i};$$

$$\text{RMSE: } \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\| D_{\text{gt}}^{i} - \hat{D}^{i} \right\|^{2}};$$

$$\text{RMSE log: } \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\| \log D_{\text{gt}}^{i} - \log \hat{D}^{i} \right\|^{2}};$$

$$\delta = \max\left(\frac{D_{\text{gt}}^{i}}{\hat{D}^{i}}, \frac{\hat{D}^{i}}{D_{\text{gt}}^{i}}\right) < t,\ t \in \left[1.25, 1.25^{2}, 1.25^{3}\right]$$

式中:N是在真值深度图中有效像素的数量.对于有效像素i,D_{gt}^{i}和$\hat{D}^{i}$分别是真值深度和预测深度.对于位姿评测,使用标准评测指标包括平移误差t_{err},旋转误差r_{err}以及绝对轨迹误差ATE.

2.3 实现细节

本文使用了PyTorch框架来实现所提出的方法.深度估计网络的编码器部分基于Resnet50结构,而位姿估计网络使用的是一个更轻量级的Resnet18结构.深度估计网络的大部分卷积层后会使用ELU作为非线性激活函数,而深度输出的卷积层使用Sigmoid激活函数.位姿估计网络会输出6自由度的相机位姿,包括3个欧拉角和3个平移分量.解码器部分使用RELU非线性激活函数.离线的实例分割网络采用的是Mask RCNN架构.由于KITTI数据集和DDAD数据集没有完整的语义标签,使用了在COCO数据集上的Mask-RCNN R50-FPN3x模型权重.

对于KITTI数据集,输入图像的分辨率调整为832像素×256像素用于训练.此外,在训练阶段会应用数据增强来提高网络的鲁棒性.数据增强策略包括随机水平翻转和缩放,增广概率均为0.5.训练样本长度设置为3,即当前帧、当前帧的前一帧和后一帧.使用Adam优化器训练网络,批处理大小和学习率分别设置为8和1e-4,学习总轮次为50.损失函数中的超参数设置为w_1=1,w_2=0.5,w_3=0.1.深度评测范围为0~80 m.

对于DDAD数据集,输入图像的分辨率会调整为640像素×384像素.批处理大小和学习率分别设置为4和1e-4.深度评测范围为0~200 m.其余参数与KITTI训练参数相同.

2.4 单目深度估计

2.4.1 KITTI数据集

在KITTI Eigen测试数据集上评估了本文模型在单目深度估计任务上的性能.为了与以往的方法进行对比,使用2.2节中评测指标在80 m范围内评估网络的性能,定量对比结果如表1所示.在对比实验的监督信号列中,S和M分别代表使用双目对训练和使用单目视频序列训练的自监督单目深度估计方法.各项评价指标的最优值进行加粗显示.与所对比的自监督方法相比,所提方法在大多数评价指标上实现了最佳的性能.

表1 KITTI Eigen测试集上的定量比较结果

Tab.1 Quantitative evaluation results on KITTI Eigen split

方法	监督信号	误差指标				精度指标		
		Abs Rel	Sq Rel	RMSE	RMSE log	δ< 1.25	δ< 1.25^2	δ< 1.25^3
Garg等[10]	*S*	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Chen等[24]	*S*	0.118	0.905	5.096	0.211	0.839	0.945	0.977
Pillai等[25]	*S*	0.112	0.875	4.958	0.207	0.852	0.947	0.977
Godard等[20]	*S*	0.107	0.849	**4.764**	0.201	0.874	0.953	0.977
周大可等[26]	*S*	**0.104**	**0.820**	4.785	**0.200**	**0.874**	**0.953**	**0.978**
Zhou等[11]	*M*	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Mahjourian等[27]	*M*	0.163	1.768	6.856	0.283	0.678	0.885	0.957
Wang等[28]	*M*	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Godard等[20]	*M*	0.115	0.882	4.701	0.190	0.879	0.961	0.982
Chen等[29]	*M*	0.118	0.909	4.816	0.195	0.876	0.955	0.980
Klingner等[13]	*M*	0.113	0.880	4.695	0.192	**0.884**	0.961	0.981
Bian等[30]	*M*	0.114	0.813	4.706	0.191	0.873	0.960	0.982
Zhang等[31]	*M*	**0.112**	0.856	4.778	0.190	0.880	0.961	0.982
Saunders等[21]	*M*	0.115	**0.785**	4.698	0.192	0.871	0.959	0.982
本文	*M*	0.116	0.807	**4.618**	**0.187**	0.875	**0.963**	**0.983**

可视化的结果比较如图4所示.早期工作中,Zhou等[11]的方法在捕获图像的细节方面表现欠佳.Godard等[20]所提方法对天空区域处理不佳.Bian等[30]通过帧间深度一致性约束提高了模型对全局结构信息的处理能力,改善了天空区域的深度估计结果,但仍有不足.综合来看,通过准确移除光度损失中运动目标,并且在静态区域施加深度时序一致性约束,本文方法对场景中较远的背景如天空区域处理得更好,对于近处的目标如路桩、指示牌和绿化柱的深度估计结果具有更清晰的边界.

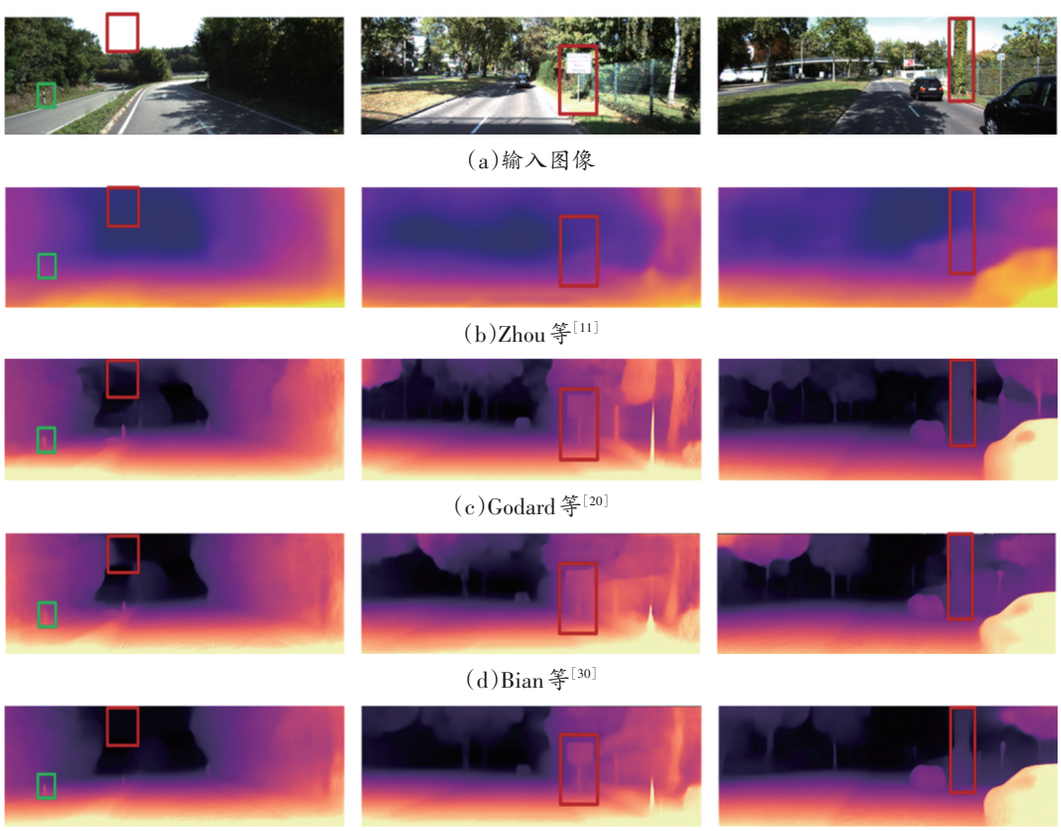
2.4.2 DDAD数据集

为了进一步验证所提方法的性能,在相比KITTI具有更多运动目标的DDAD数据集上进行了相关的实验验证.定量对比结果如表2所示.与所对比法相比,所提方法在大多数评价指标上实现了最佳的性能,仅在RMSE和δ<1.25指标上略低于Bian等[30]的方法.

DDAD数据集上的可视化对比结果如图5所示.在该数据集上,本文方法同样在远处背景和近处的目标的深度估计上都表现得更为出色.本文方法在第一幅图像的深度估计结果中对树林和天空区域的区分更为明显,在第二幅图像的交通灯区域具有更锐利的深度边界,在第三幅图像的运动的车辆部分的深度估计更为平滑且完备.

2.5 位姿估计

位姿网络是基于视频序列训练的自监督单目深



(a)输入图像

(b)Zhou等[11]

(c)Godard等[20]

(d)Bian等[30]

(e) 本文

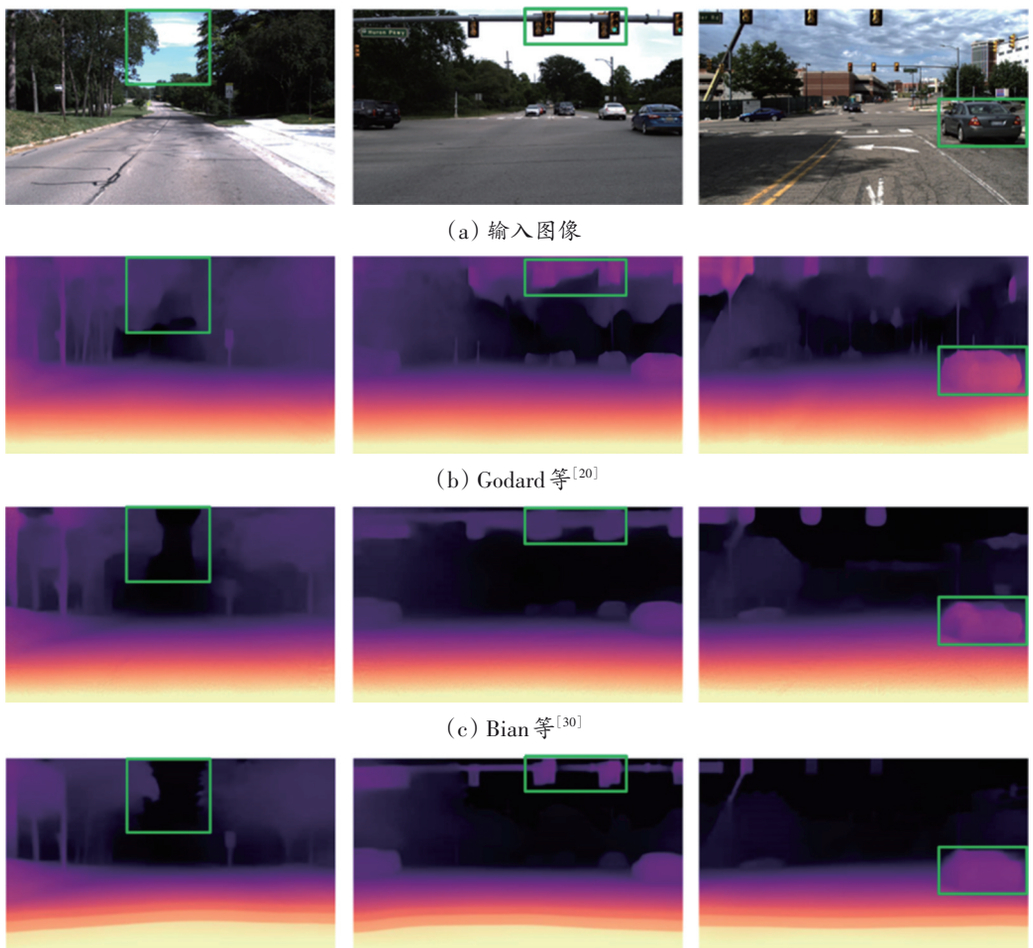
图4 KITTI Eigen测试集上的可视化结果

Fig. 4 Qualitative results on the KITTI Eigen test set

表2 DDAD测试集上的定量比较结果

Tab.2 Quantitative evaluation results on DDAD test set

方法	监督信号	误差指标				精度指标		
		Abs Rel	Sq Rel	RMSE	RMSE log	δ<1.25	δ<1.25^2	δ<1.25^3
Godard等[20]	*M*	0.239	12.547	18.392	0.316	0.752	0.899	0.949
Klingner等[13]	*M*	0.200	7.944	17.149	0.289	0.769	0.911	0.957
Bian等[30]	*M*	0.169	3.877	**16.290**	0.280	0.773	0.905	0.951
本文	*M*	**0.160**	**3.297**	16.922	**0.266**	**0.768**	**0.912**	**0.956**



(a) 输入图像

(b) Godard 等[20]

(c) Bian 等[30]

(d) 本文

图 5　DDAD 测试集上的可视化结果

Fig. 5　Qualitative results on the DDAD test set

度估计方法所特有的网络分支.准确的位姿估计对视图的准确重构起着重要的作用.

为了评估位姿网络的性能,将所提方法与以往的方法在 KITTI Odometry 数据集上进行了对比.由于位姿网络预测的是相对位姿,在测试阶段预测位姿会与真实位姿的尺度进行对齐[30].定量的对比结果如表 3 所示.t_{err}(%)和 r_{err}($100^{-1}\cdot m^{-1}$)分别代表在测试轨迹上的平均平移 RMSE 漂移和平均旋转 RMSE 漂移;ATE 表示绝对轨迹误差.评测指标综合考虑了模型在整个测试序列上的性能.结果表明本文的方法在测试序列 09 上优于其他方法.在测试序列 10 上,本文的方法仅次于 Bian 等[30].

表 3　KITTI Odometry 数据集上的视觉里程计结果

Tab.3　Visual odometry results on KITTI Odometry dataset

方法	监督信号	序列 09			序列 10		
		t_{err}/%	r_{err}/[°·(100 m)$^{-1}$]	ATE/m	t_{err}/%	r_{err}/[°·(100 m)$^{-1}$]	ATE/m
Zhan 等[32]	*S*	11.89	3.60	52.12	12.82	3.41	24.70
Li 等[33]	*S*	7.01	3.60	—	10.63	4.60	—
Yin 等[34]	*M*	28.72	9.80	158.45	23.90	9.00	43.04
Shen 等[35]	*M*	9.91	3.80	27.08	12.18	5.90	24.44
Godard 等[20]	*M*	17.17	3.85	76.22	11.68	5.31	20.35
Bian 等[30]	*M*	7.31	3.05	23.56	7.79	4.90	12.00
本文	*M*	6.44	0.55	17.86	9.85	1.06	13.48

可视化结果如图6所示,本文位姿网络估计的轨迹与其他方法相比与真值拟合程度更高.

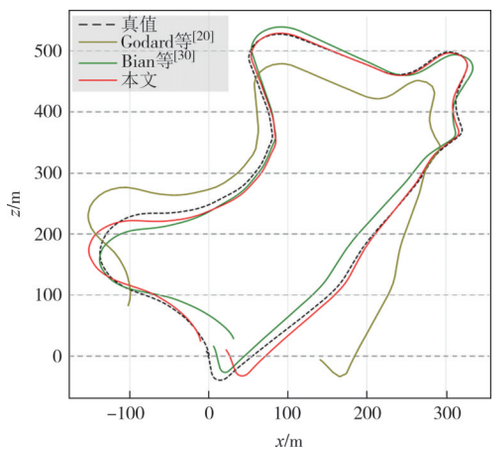


图6 KITTI Odometry 数据集中序列09上的估计轨迹

Fig. 6 Estimated trajectory on sequence 09 of KITTI Odometry dataset

2.6 消融实验

2.6.1 动态掩膜和深度时序一致性

为了验证所提语义掩膜策略和深度一致性损失的有效性,在KITTI Eigen测试集和DDAD测试集上进行了消融实验.消融实验从基线模型开始,然后添加改进策略直到完整的模型.其中动态掩膜用于在光度损失中移除运动目标,时序一致性即在非运动区域应用的帧间深度一致性约束.定量结果如表4所示,结果表明这两个改进策略在两个数据集上都能有效提升基线模型性能,综合两者可以得到更优的结果.

此外,由于DDAD数据集具有更多的运动目标,在基线模型上引入动态掩膜策略相比KITTI上的实验具有更显著的性能改进效果,这也进一步验证了所提语义掩膜策略的有效性.

表4 消融实验结果

Tab.4 Ablation experimental results

数据集	方法	误差指标				精度指标		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
KITTI	基线方法	0.143	1.026	5.058	0.211	0.820	0.945	0.980
	基线+动态掩膜	0.136	0.964	4.964	0.204	0.833	0.949	0.981
	基线+时序一致性	0.122	0.846	4.796	0.194	0.857	0.958	0.983
	本文	**0.116**	**0.807**	**4.618**	**0.187**	**0.875**	**0.963**	**0.983**
DDAD	基线方法	0.191	4.110	17.174	0.300	0.737	0.886	0.941
	基线+动态掩膜	0.164	3.488	17.629	0.278	0.762	0.904	0.951
	基线+时序一致性	0.174	3.974	**16.909**	0.288	0.765	0.896	0.948
	本文	**0.160**	**3.297**	16.922	**0.266**	**0.768**	**0.912**	**0.956**

2.6.2 MPC指标阈值的确定

基于估计的位姿和目标深度图,可以将源视图图像/掩膜投影生成重构的目标图像/掩膜.图7展示单个训练样本中实例掩膜的投影过程.图中第一列到第三列分别为源图像/实例、目标图像/实例以及重构的目标图像/实例.目标掩膜与重构的目标掩膜之间的重叠程度可以用来指示对应的目标的运动状态.如第1.3.3节所述,使用Dice系数作为重叠程度的指标,并计算在两个实例掩膜的并集处的光度损失MPC,定量结果如表5所示.

由于语义标签投影过程的潜在要求是投影公式中预测的位姿和深度值都是完美的,而这在实际操作过程中显然无法满足.与文献[21]类似,Dice中的阈值θ_1设置为0.8,允许投影后的掩膜与目标掩膜有一定的错位.

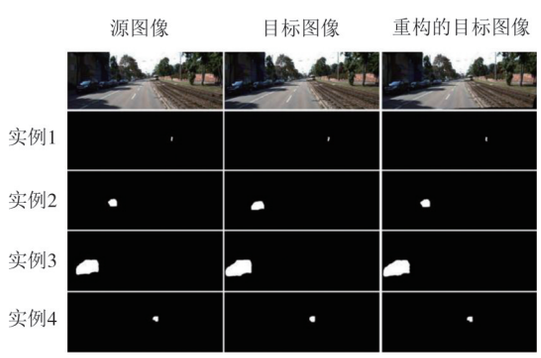


图7 实例掩膜投影示例

Fig.7 An example of the instance mask projection

如表5所示,图7中的第1个和第4个实例属于运动目标,但是Dice指标只能区分出第1个实例.出现这种情况的一个可能原因是相邻帧之间的时间间隔较小,无法用Dice指标区分物体的运动.然而,运动的实例区域通常会有更大的光度损失,从实验结

表5　图7中不同实例的Dice和MPC值

Tab.5　Dice and MPC value for different instances in Fig. 7

实例	运动状态	Dice	MPC
1	运动	0.674	0.276
2	静止	0.894	0.074
3	静止	0.938	0.081
4	运动	0.872	0.202

果中也可以看出运动实例通常有更大的MPC值.

因此,MPC指标提供了一个额外的判断标准.为了确定合适的MPC阈值θ_2,本文在0.1~0.2的范围内,以0.02为间隔进行了相关实验,定量对比结果如表6所示.从实验结果看出,θ_2取值为0.14时取得了相对最佳的评估结果.

表6　θ_2不同取值下的实验结果

Tab.6　Experimental results of parameter $\theta_2$under different values

θ_2	Abs Rel	Sq Rel	RMSE	δ< 1.25
0.10	0.121	0.898	4.729	0.869
0.12	0.118	0.880	4.751	0.871
0.14	0.116	**0.807**	**4.618**	0.875
0.16	**0.115**	0.832	4.741	0.874
0.18	0.117	0.809	4.629	0.872
0.20	0.117	0.846	4.631	**0.876**

3　结 论

本文提出了一种有效的基于图像序列训练的自监督单目深度估计方法.为了防止与相机运动耦合的目标运动干扰相机位姿估计的精度,在预计算的语义掩膜的辅助下对这些区域进行屏蔽.通过结合源帧与目标帧实例掩膜语义时间一致性和光度一致性约束,提出了一种判别动态类别目标中的静态目标的检测方法,确保这些区域的光度损失能够有效用于更新网络权重.此外,对深度图引入了时序一致性约束,以稳定相邻帧的深度估计.实验结果表明所提方法在KITTI和DDAD数据集上都实现了高质量的表现.

本文证明了基于语义掩膜投影检测动态类别目标的方法的有效性,但也存在一定的局限性.例如阈值指标在不同的数据集上可能需要重新确定,对于更复杂的运动场景还需要更细化的处理策略等.此外,引入其他的骨干网络例如Visual Transformer可能进一步提升任务的预测精度.未来计划研究场景理解的多任务学习,包括深度估计、语义分割以及光流估计等.

参考文献

[1] LAGA H, JOSPIN L V, BOUSSAID F, et al. A survey on deep learning techniques for stereo-based depth estimation[J]. IEEE Trans Pattern Anal Mach Intell, 2022, 44(4): 1738-1764.

[2] FURUKAWA Y, HERNÁNDEZ C. Multi-view stereo: a tutorial [J]. Foundations and Trends in Computer Graphics and Vision, 2015, 9(1/2): 1-148.

[3] MING Y, MENG X Y, FAN C X, et al. Deep learning for monocular depth estimation: a review [J]. Neurocomputing, 2021, 438: 14-33.

[4] ZHANG R, TSAI P S, CRYER J E, et al. Shape-from-shading: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(8): 690-706.

[5] TAO M W, HADAP S, MALIK J, et al. Depth from combining defocus and correspondence using light-field cameras[C]//2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia. IEEE, 2013: 673-680.

[6] SAXENA A, CHUNG S H, NG A Y. Learning depth from single monocular images [C]//Proceedings of the 18th International Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada. ACM, 2005: 1161-1168.

[7] KARSCH K, LIU C, KANG S B. Depth transfer: depth extraction from video using non-parametric sampling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(11): 2144-2158.

[8] 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-Volume 2. Montreal, Canada. ACM, 2014: 2366-2374.

[9] FU H, GONG M M, WANG C H, et al. Deep ordinal regression network for monocular depth estimation [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE, 2018: 2002-2011.

[10] GARG R, VIJAY KUMAR B G, CARNEIRO G, et al. Unsupervised CNN for single view depth estimation: geometry to the rescue [M]//LEIBE B, MATAS J, SEBE N, et al, eds. Computer Vision - ECCV 2016. Cham: Springer International Publishing, 2016: 740-756.

[11] 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). Honolulu, HI, USA. IEEE, 2017: 6612-6619.

[12] LIU L, ZHAI G Y, YE W L, et al. Unsupervised learning of scene flow estimation fusing with local rigidity[C]//IJCAI, 2019: 876-882.

[13] KLINGNER M, TERMÖHLEN J A, MIKOLAJCZYK J, et al. Self-supervised monocular depth estimation: solving the dynamic

object problem by semantic guidance[C]//European Conference on Computer Vision. Cham:Springer,2020:582-600.

[14] ZHANG H K, LI Y, CAO Y, et al. Exploiting temporal consistency for real-time video depth estimation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul,Korea (South). IEEE,2019:1725-1734.

[15] LI S Y,LUO Y,ZHU Y, et al. Enforcing temporal consistency in video depth estimation [C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Montreal, BC,Canada. IEEE,2021:1145-1154.

[16] KOPF J,RONG X J,HUANG J B. Robust consistent video depth estimation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville,TN,USA. IEEE, 2021:1611-1621.

[17] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[EB/OL]. https://arxiv.org/abs/1506.02025.

[18] HE K M,GKIOXARI G,DOLLÁR P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020,42(2):386-397.

[19] 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.

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

[21] SAUNDERS K,VOGIATZIS G,MANSO L J. Dyna-DM:dynamic object-aware self-supervised monocular depth maps [C]//2023 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC). Tomar,Portugal. IEEE,2023:10-16.

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

[23] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite [C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence,RI,USA. IEEE, 2012: 3354-3361.

[24] CHEN P Y, LIU A H, LIU Y C, et al. Towards scene understanding: unsupervised monocular depth estimation with semantic-aware representation [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach,CA,USA. IEEE,2019:2619-2627.

[25] PILLAI S,AMBRUŞ R,GAIDON A. SuperDepth:self-supervised, super-resolved monocular depth estimation [C]//2019 International Conference on Robotics and Automation (ICRA). Montreal,QC,Canada. IEEE, 2019: 9250-9256.

[26] 周大可,田径,杨欣. 结合局部平面参数预测的无监督单目图像深度估计[J]. 中国图像图形学报,2021,26(1):165-175.
ZHOU D K,TIAN J,YANG X. Unsupervised monocular image depth estimation based on the prediction of local plane parameters [J]. Journal of Image and Graphics,2021,26(1):165-175.(in Chinese)

[27] MAHJOURIAN R, WICKE M, ANGELOVA A. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE,2018:5667-5675.

[28 WANG C Y, BUENAPOSADA J M, ZHU R, et al. Learning depth from monocular videos using direct methods [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City,UT,USA. IEEE,2018:2022-2030.

[29] CHEN S,PU Z D,FAN X, et al. Fixing defect of photometric loss for self-supervised monocular depth estimation [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022,32(3):1328-1338.

[30] 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.

[31] ZHANG Y R, GONG M G, LI J Z, et al. Self-supervised monocular depth estimation with multiscale perception[J]. IEEE Transactions on Image Processing: a Publication of the IEEE Signal Processing Society,2022,31:3251-3266.

[32] ZHAN H Y, GARG R, WEERASEKERA C S, et al. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City,UT,USA. IEEE,2018:340-349.

[33] LI R H,WANG S,LONG Z Q, et al. UnDeepVO: monocular visual odometry through unsupervised deep learning [C]//2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane,QLD,Australia. IEEE,2018:7286-7291.

[34] YIN Z C,SHI J P. GeoNet:unsupervised learning of dense depth, optical flow and camera pose[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. IEEE,2018:1983-1992.

[35] SHEN T W,LUO Z X,ZHOU L, et al. Beyond photometric loss for self-supervised ego-motion estimation[C]//2019 International Conference on Robotics and Automation (ICRA). Montreal,QC, Canada. IEEE,2019:6359-6365.