DOI: 10.13382/j.jemi.B2407464

基于稳定光度损失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 无锡 214122)

摘 要:在基于视频的无监督单目深度估计模型训练中,光度损失一直发挥着重要作用,但其在弱纹理区域和边缘区域等特殊区域普遍存在较大误差,导致训练网络的监督信号存在较强的不稳定性。针对这一问题,提出一种更具鲁棒性的无监督单目深度估计方法。本文方法首先结合双分支编码器和通道注意力模块来提升单帧深度网络对深度特征的提取能力,然后利用单帧深度网络结果引导进行多帧深度估计,以提高深度估计的准确性。在此基础上设计一种新型光度损失函数,通过计算图像梯度上的光度损失消除局部亮度变化引起的不合理监督,并利用连续像素之间的差异特性来定义模糊像素,最后基于二进制掩模排除由于目标帧和重构目标帧上边缘模糊像素产生的错误监督。本文方法在 KITTI 数据集的测试结果中,平均相对误差、平方相对误差、均方根误差等多项指标均有提升,平均相对误差和平方相对误差分别降低至 0.075 和 0.548。实验结果证明,与其他先进方法相比,本文方法进一步提高了现有模型的性能。

关键词: 单目深度估计;无监督学习;深度学习;光度损失;鲁棒性
中图分类号: TN911.73 **文献标识码**: A **国家标准学科分类代码**: 510.40

Unsupervised monocular depth estimation based on stable photometric loss

Qu Yi Chen Ying
(Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China)

Abstract: The photometric loss has been playing an important role in the training of video-based unsupervised monocular depth estimation models. However, it generally has large errors in special regions such as weak texture regions and edge regions, which leads to strong instability in the supervision signal of the training network. To solve the problem, a more robust unsupervised monocular depth estimation method is proposed. The method first combines the dual-branch encoder and the channel attention module to improve the extraction ability of the single-frame depth network for depth features. Then, the single-frame depth network results are used to guide the multi-frame depth estimation to improve the accuracy of depth estimation. On the basis, a new photometric loss function is designed. By calculating the photometric loss on the image gradient, the unreasonable supervision caused by local brightness changes is eliminated. At the same time, the difference between successive pixels is used to define the blurry pixels. Finally, the false supervision caused by the blurred pixels on the target frame and the reconstructed target frame is excluded based on the binary mask. In the test results of the KITTI dataset, multiple indicators such as the average relative error, the square relative error and the root mean square error have improved. The average relative error and the squared relative error are reduced to 0.075 and 0.548 respectively. The experimental result shows that the proposed method further improves the performance of existing models compared with other advanced methods.

Keywords: monocular depth estimation; unsupervised learning; deep learning; photometric loss; robustness

收稿日期:2024-04-23 Received Date: 2024-04-23
*基金项目:国家自然科学基金(62173160)资助项目

0　引　言

场景的深度估计是计算机视觉中的经典问题,在景深渲染[1]、视觉同时定位与建图(visual simultaneous localization and mapping,VSLAM)[2]、自动驾驶[3]等传统研究方向中都起着重要作用,甚至延伸至农业生产[4]、城市规划[5]等众多新兴领域。Eigen 等[6]在 2014 年首次实现将深度神经网络用于单目深度估计任务,使用包含粗尺度和细尺度两个尺度的卷积神经网络结构进行场景深度预测,自此基于深度学习的单目深度估计时代正式拉开帷幕。由于有监督模型都以深度信息作为标签进行训练,但深度信息很难获取,因此无监督模型逐渐成为单目深度估计领域的研究热点。目前比较主流的无监督单目深度估计算法基础是由 Zhou 等[7]提出的基于视频帧序列的网络架构,整个网络分为单视图深度估计网络和相机位姿估计网络两部分,通过对两个网络联合训练可以学习场景深度和相机移动轨迹,进而通过视图重构监督网络训练。Godard 等[8]提出的 Monodepth2 网络在此基础上通过全分辨率的多尺度采样方法减少视觉伪影,并通过设置掩膜来处理遮挡和运动物体问题。由于单目深度估计始终存在可解释性较差的问题,因此近些年出现了很多基于立体视觉(multi-view stereo,MVS)的方法,利用了同一相机不同时刻拍摄照片之间的几何关联,提高深度估计的准确性。其中比较有代表性的是 Wang 等[9]提出的 MOVEDepth 网络,利用单视图深度估计网络为每个像素提供一个粗糙的先验深度,之后在多帧深度估计网络中在每个像素先验深度附近进行采样和回归深度,最终结合单帧深度估计和多帧深度估计结果,精准预测场景深度。

在约束单目深度估计网络训练的过程里,光度损失一直起着重要作用。基于 MSE 的简单光度损失通常能在单目深度估计中起到很好的监督效果,且效率很高,但在边缘区域等亮度变化不稳定的特殊区域,仍存在着由亮度差而非图像重构误差引起错误监督信号的普遍问题。为了解决这一问题,Zheng 等[10]提出了一个深度图估计的概率框架,对给定的具有光度一致性的图像的像素级视图选择概率和深度图估计进行联合建模,使用归一化互相关方法(normalized cross-correlation,NCC),依靠局部统计特性来增强亮度不变性。但当亮度存在较大变化时,本文方法并不能有效消除外观偏差。Godard 等[11]提出在光度损失中引入 SSIM 项,以更好地表示真实的图片分布,但其在明显的亮度变化下鲁棒性能较差。Yang 等[12]在此基础上提出一种新的无监督单目深度估计网络,对输入图像上像素的光度不确定性进行了建模,并提供了一个可学习权重的光度误差,预测了在动态训练过程中对齐的源图像和目标图像之间的亮度变换参数,试图通过亮度变换解决亮度差异引起的稳定性问题,但需要额外的网络来计算变换参数。此外,Watson 等[13]发现在单目深度估计中使用的光度损失通常有多个局部最小值,通过这些值看似可以求出代替真实深度的估计值,但实际上会限制回归网络的学习内容,因此通过使用半全局匹配(semi-global matching,SGM)方法,从立体对图像中获得的深度提示标签,帮助网络摆脱局部最小值的局限,但其需要用到立体数据集,并使用额外的监督信号。这些方法虽然能够在一定程度上提高光度损失的稳定性,但在训练中需要引入额外的网络或数据集,这大大加重了模型训练的负担。

基于上述分析,本文方法设计了一个具有光亮稳定性的单目深度估计模型,其核心是一种更具鲁棒性的光度损失函数,在不需要过多增加网络结构和数据集的情况下,能够简单而有效地提高现有模型的性能。本文的主要贡献如下:

1)采用传统单帧单目深度估计与基于 MVS 的多帧深度估计相结合的方式构建整体网络,同时利用双分支编码器和通道注意力模块,以深入挖掘多尺度特征并对其进行充分融合。

2)利用模糊像素的梯度特性生成二进制掩膜,筛选并排除由边缘区域的模糊像素产生的不合理损失干扰,从而进一步确保在网络训练过程中监督信号的准确性和稳定性。

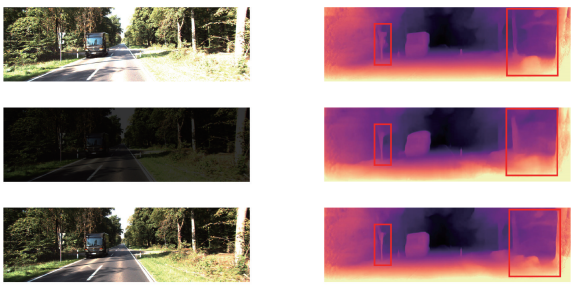
3)针对弱纹理区域的光亮变化敏感问题,利用图像的几何信息,计算基于图像梯度的光度损失,以此来抑制亮度变化引起的错误损失。

1　问题分析

从之前工作的设计思路中不难发现,基于视频的无监督单目深度估计的关键任务是利用几何投影关系重构出目标帧,这其中光度损失作为重要信号监督整个训练过程。然而当对亮度变化明显的图像进行训练时,常常会由亮度差引起错误的监督信号,而不是由图像合成误差引起正确的监督信号。以经典网络 Monodepth2[8]为例,从图 1 中不难看出,不同光亮条件对边缘区域(树木、指示牌边缘)和弱纹理区域(草丛、树林)等特殊区域的深度预测存在较大的影响。

出现在物体边缘上的预测差异实质上是由输入图像引起的[14]。具体来说,在基于视频的无监督单目深度估计任务中,首先使用二维到三维反向投影来构建三维场景中目标帧的预测深度图,然后通过三维到二维的重投影合成一个新的目标视图。在重投影的过程中,为了最小化每个像素的重投影误差,网络需要调整每个像素的

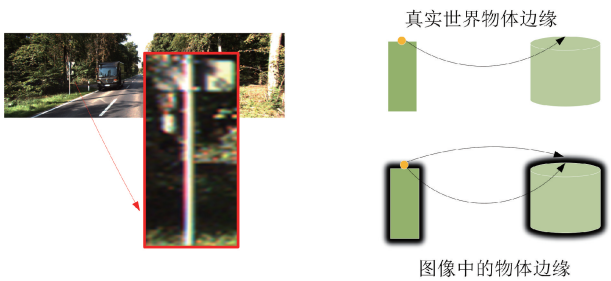
深度,使其重新投影到源视图上,最终实现每个像素光度损失的最小化。这个过程必须在每个像素都属于一个确定物体的条件下进行,也就是说每个像素都被期望附在一个实际物体上,而不能使用一个深度值来描述属于两个物体的像素。然而由于抗锯齿化处理,输入图像边缘往往呈现“混合”状态,即边缘像素为两侧像素的加权和,打破了这一训练条件,因而这些模糊像素的深度值没有物理意义。



(a) 不同光照输入　(b) 深度预测图
(a) Different light inputs　(b) Predicted depth maps

图1　不同光照下深度预测图对比
Fig. 1　Comparison of depth prediction maps under different illumination

像边缘的像素级模糊如图2(a)所示,图中指示牌边界处的像素在连接处呈渐变状,这些像素是由后方树木和指示牌的融合而成的,实际上既不是来自树木也不是来自指示牌。如图2(b)所示,模糊像素在二维空间到三维空间的反向投影中没有实际的归属对象,因此会导致模糊像素脱离主体,三维点无论是在空间上还是光度上都没有实际物理意义。在经过三维空间到二维空间的重投影后,目标图像和合成图像中的这些模糊像素没有相对应的匹配关系,从而在评估损失时会产生较大的光度损失。但是,光度损失理论上应该只存在于深度预测不正确的区域,这些模糊像素产生的损失并不代表不准确的深度预测,不管其预测的结果是什么,网络都只能从这里学习到不合理的损失。



(a) 输入图像边缘　(b) 边缘比较
(a) The edges of the input images　(b) The edges comparison

图2　图像边缘细节
Fig. 2　Image edge details

此外,传统的深度估计算法很难在亮度变化条件下对弱纹理区域进行稳定的深度估计。这是因为纹理能够反映一个图像中的同质化程度,体现了物体表面缓慢变化或者周期性变化的结构排列特点。而弱纹理区域顾名思义就是纹理特征较少的图像区域,因而具有颜色单调,点特征少,像素间可分辨性差的特点。由于这些区域具有高度一致的局部像素强度[15],缺乏明显的表面特征,所以传统的深度估计算法往往难以准确捕捉到弱纹理区域的深度信息。同时,这些区域往往更容易受到光照、阴影等因素的影响,使得表面的细节更加难以捉摸,从而在亮度变化明显的情况下增加了深度估计的难度。

2　改进的网络

2.1　网络架构

以MOVEDepth网络[9]为基准网络,本文方法网络构成如图3所示,总体上由深度网络和相机姿态网络组成,其中深度网络又分为单帧深度网络和多帧深度网络两个部分。

其中,相机姿态网络与文献[7]相似,以ResNet18[16]作为骨干网络,将视频中连续的两帧输入网络,通过卷积层和反卷积层构成编解码器结构,最后通过将全局平均池化层预测出具有6个自由度的相机相对位姿$[\boldsymbol{R}\mid\boldsymbol{T}]_{0\to-1}$。

帧深度网络沿用文献[9]的网络架构,依托ResNet18构建编解码器。与文献[9]不同的是,为了尽可能充分地提取不同层次的特征,设计了包含两个分支的编码器,一个分支用较小感受野编码细节信息,另一个分支通过继续增加卷积层来扩大感受野范围,从而较为完整地获取全局信息。此外,在编解码器之间设计了通道注意力模块,根据注意力图建模通道间的依赖关系,能够得到具有丰富上下文信息的聚合特征。通道注意力图的具体计算方法如式(1)所示。

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

式中:$\boldsymbol{A}_{ij}$表示注意图,$\boldsymbol{S}_{ij}=\boldsymbol{F}_i\cdot\boldsymbol{F}_j^{\mathrm{T}}$,$\boldsymbol{F}_i$,$\boldsymbol{F}_j$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,$C$表示通道总数。

最后利用残差连接保留原始的低层次特征,实现不同层次特征的充分融合,最终输出预测的单帧深度图D_{Mono}。

多帧深度网络基于MVS进行,基于文献[8]的研究基础,网络同样由编解码器组成。编码器网络以特征金字塔网络[17](feature pyramid network,FPN)为骨干网络,首先根据单帧深度网络输出以及预测的相对位姿变化调

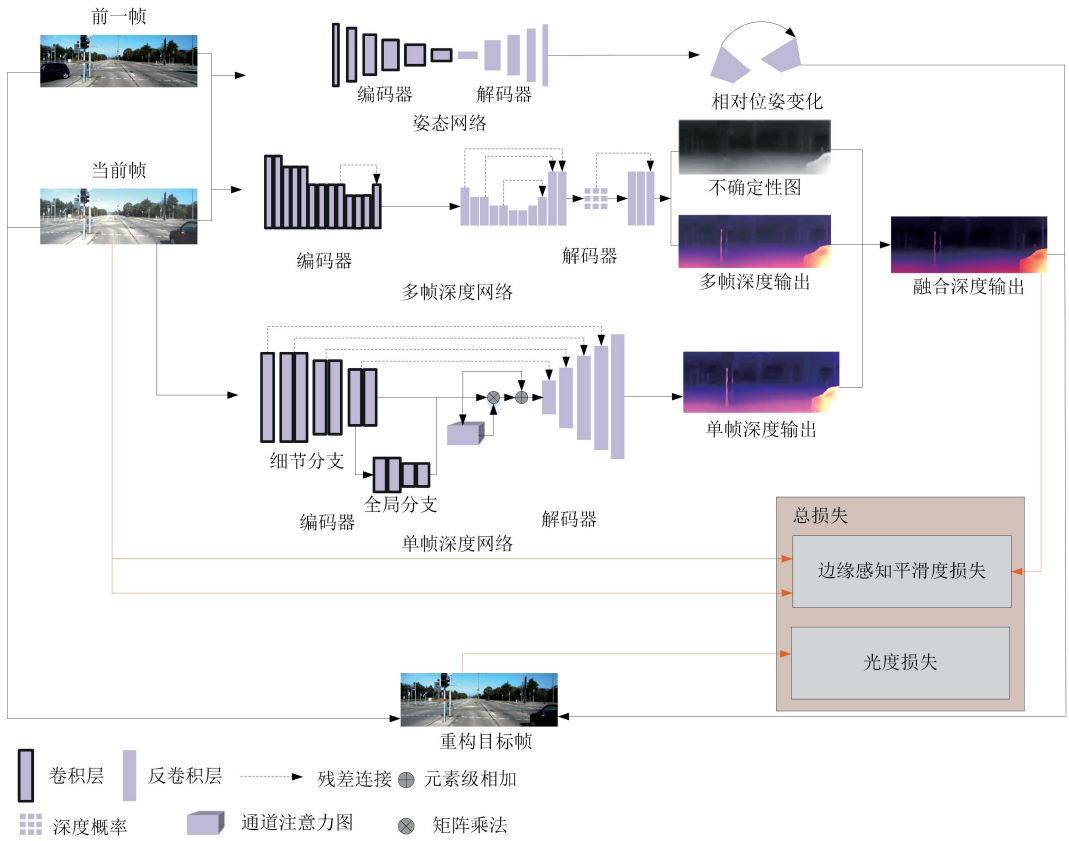


图 3　网络框架图

Fig. 3　Frame diagram of the network structure

整深度采样范围,进而得到每个像素点的候选深度值,然后利用多个三维卷积堆叠形成编码器,在不同的候选深度下构建衡量当前帧与前一帧的成本量,接着通过轻量级的解码器得到深度概率最大的深度图作为多帧深度图 D_{MVS} ,并基于深度概率生成不确定性图 U ,最终根据单帧深度 D_{Mono} 、不确定性图 U 和多帧深度 D_{MVS} 能够得到更可靠的融合深度 D_{Fuse} 为:

$$D_{\mathrm{Fuse}} = U \otimes D_{\mathrm{Mono}} + (1 - U) \otimes D_{\mathrm{MVS}} \tag{2}$$

式中:⊗表示元素级相乘。

在移动摄像机和静态场景假设下,可以找到前一帧 $\boldsymbol{I}_{-1}$ 与当前帧 $\boldsymbol{I}_0$ 相机坐标的对应关系,进而利用前一帧 $\boldsymbol{I}_{-1}$ 生成重构当前帧 $\boldsymbol{I}_{-1\to 0}$,并把重构帧和当前帧间的误差作为网络训练的监督信号。重构当前帧的公式如式(3)所示。

$$\boldsymbol{I}_{-1\to 0}(D) = \boldsymbol{I}_{-1}\{\mathrm{proj}(\mathrm{reproj}(\boldsymbol{I}_0, D, [\boldsymbol{R}\mid \boldsymbol{T}]_{0\to -1}), \boldsymbol{K})\} \tag{3}$$

其中, $\boldsymbol{K}$ 表示已知的相机内部参数; $\{\}$ 表示采样操作; reproj 表示返回 $\boldsymbol{I}_0$ 拍摄时刻的三维点云; proj 表示返回深度图 D 的二维坐标投影。

在此基础上,本文方法对损失函数进行了改进。首先在整体上仍沿用文献[9]的损失构成,由深度平滑度损失 L_{s} 和光度损失 L_{p} 共同组成损失函数。其中,用来消除低梯度区域噪声的深度平滑度损失 L_{s} 保持不变,其定义为:

$$L_{\mathrm{s}} = |\partial_x d^*| e^{-|\partial_x \boldsymbol{I}_0|} + |\partial_y d^*| e^{-|\partial_y \boldsymbol{I}_0|} \tag{4}$$

式中: d^* 为预测深度。

传统光度损失主要由 SSIM 和 L1 损失组成,用于计算原目标帧和重构帧间的相似性,其定义为:

$$L_{\mathrm{p}} = \alpha \frac{1 - \mathrm{SSIM}(\boldsymbol{I}_0, \boldsymbol{I}_{-1\to 0})}{2} + (1 - \alpha)|\boldsymbol{I}_0 - \boldsymbol{I}_{-1\to 0}| \tag{5}$$

式中: $\boldsymbol{I}_0$ 为目标帧; $\boldsymbol{I}_{-1\to 0}$ 为重构目标帧; α 为权重系数。

本文方法提出在计算光度损失之前先计算模糊掩模,提取目标图像和重建图像中有歧义的像素,从而抑制由于物体边界模糊像素造成的不合理的光度损失。同时在计算过程中引入梯度算子,以改善弱纹理区域由于局部亮度变化导致的光度损失虚高问题,从而抑制高低梯度之间剧烈的亮度变化所带来的错误监督信号。本文方法改进的光度损失定义如式(6)所示。

$$L_{\mathrm{pn}} = A \otimes [\alpha_{\mathrm{n}} \frac{1 - \mathrm{SSIM}_{\mathrm{n}}(\boldsymbol{I}_0, \boldsymbol{I}_{-1\to 0})}{2} + (1 - \alpha_{\mathrm{n}})|\boldsymbol{I}_0 - \boldsymbol{I}_{-1\to 0}|] \tag{6}$$

式中：A 表示最终的模糊掩膜；$SSIM_n$ 表示引入梯度算子后的 SSIM 损失。

2.2　模糊掩模

计算模糊掩膜首先需要通过计算空间频率确定每个像素的模糊程度。通常可以通过计算相邻像素之间的差异来表示梯度[18]，也就是空间频率：

$$\nabla_{i\pm}(x,y) = \boldsymbol{I}(x,y) - \boldsymbol{I}(x \pm 1,y) \tag{7}$$

$$\nabla_{j\pm}(x,y) = \boldsymbol{I}(x,y) - \boldsymbol{I}(x,y \pm 1) \tag{8}$$

$$\nabla(x,y) = \left\| \frac{\nabla_{i+}(x,y) - \nabla_{i-}(x,y)}{2}, \frac{\nabla_{j+}(x,y) - \nabla_{j-}(x,y)}{2} \right\|_2 \tag{9}$$

式中：(x,y) 表示任意一个像素，$\nabla_{i\pm}(x,y)$ 表示该像素沿水平方向的对应梯度，$\nabla_{j\pm}(x,y)$ 表示沿垂直方向的对应梯度，$\|\cdot\|_2$ 表示 L2 范数。

给定一个目标帧 $\boldsymbol{I}_0$，为了排除在第 2 节中描述的边缘模糊像素干扰，形成模糊映射 A_0，需要在输入图像中提取模糊性。由于相邻物体之间的色差越大，位于连接处的像素就越模糊，因此首先通过式(9)计算出频率图 f_0。根据之前的分析，位于边界处的模糊像素为了平滑物体边界处剧烈的颜色变化，其颜色呈现出两侧像素的混合形态(如图 1(a)中指示牌边缘像素逐渐从白色变为黑色)。在水平或是垂直方向，这些像素具有在相反方向的梯度具有相反符号的特点，根据这一特性，设计了一个二进制掩模 μ 来筛选这些高频像素：

$$\mu = [\nabla_{u+} \cdot \nabla_{u-} < 0 \vee \nabla_{v+} \cdot \nabla_{v-} < 0] \tag{10}$$

其中，$[\cdot]$ 表示艾弗森括号。

然后，初始输入目标帧 $\boldsymbol{I}_0$ 的模糊映射 A_0 可计算为：

$$A_0 = \mu \otimes f_0 \tag{11}$$

其中，$\otimes$ 表示元素级相乘，同理可得源帧的模糊映射 A_{-1}。

然而由于光度损失基于目标帧 $\boldsymbol{I}_0$ 和重构目标帧 $\boldsymbol{I}_{-1\to0}$ 进行计算，因而这两个图像都会导致损失的不可信，所以在掩膜计算的过程中同样应该考虑重构目标帧 $\boldsymbol{I}_{-1\to0}$ 的影响。将式(3)中 $\boldsymbol{I}_0$ 的重投影过程定义为 $\Re$：

$$\Re = (\mathrm{reproj}(\boldsymbol{I}_0, D_{\mathrm{Mono}}, [\boldsymbol{R} \mid \boldsymbol{T}]_{0\to-1}) \tag{12}$$

$\Re$ 表示了 $\boldsymbol{I}_0$ 中的每个像素在源帧 $\boldsymbol{I}_{-1}$ 中的对应位置。但 $\Re$ 中不仅仅包含了生成的重构帧与目标帧之间的像素对应关系，还包含了源帧 $\boldsymbol{I}_{-1}$ 中的模糊像素对重构帧 $\boldsymbol{I}_{-1\to0}$ 的干扰信息。因此可以对源帧的模糊映射 A_{-1} 进行双线性采样来得到投影模糊映射 $A_{-1\to0}$，以确定来自源帧的模糊像素在重构帧中的对应像素。

最后基于逻辑上的或操作取投影模糊映射 $A_{-1\to0}$ 和目标帧模糊映射 A_0 中的像素级最大值：

$$A = [max(A_0, A_{-1\to0}) < \delta] \tag{13}$$

本文方法 δ 取 0.3。

2.3　梯度光度损失

基于视频的无监督单目深度估计方法需要在亮度恒定假设下进行。然而，在相机自动曝光和照明自然变化的影响下，这一假设经常不能成立。很多时候，由于受到亮度变化的影响，即使重构帧与目标帧能实现比较精准的对齐，由光亮引起的较大光度损失也会误导网络训练最终导致估计错误。传统的损失函数虽然结合了基于图像的 L1 和 SSIM 误差，能够在处理光亮问题时留有余地，在一定程度上适应亮度变化，但在面对明显亮度变化时仍显得力不从心。在此背景下，本文方法提出了一种简单有效的光度损失计算方案，在保证监督力度的同时，能够稳定应对亮度变化。这一方案将梯度图纳入考量，用以衡量监督重构帧和目标帧间的结构相似度，从而满足上述两大要求。

对于给定的目标帧 $\boldsymbol{I}_0$ 和通过式(3)计算的重构帧 $\boldsymbol{I}_{-1\to0}$，在计算光度损失之前可按照下列方式计算它们的梯度：

$$\boldsymbol{G}_0 = \frac{1}{2}(\partial_x \boldsymbol{I}_0 + \partial_y \boldsymbol{I}_0) \tag{14}$$

$$\boldsymbol{G}_{-1\to0} = \frac{1}{2}(\partial_x \boldsymbol{I}_{-1\to0} + \partial_y \boldsymbol{I}_{-1\to0}) \tag{15}$$

然后，利用梯度图来计算结构相似度，同时保留原本利用彩色图像计算的 SSIM 和 L1 损失，式(16)具体可表示为：

$$L_{\mathrm{pn}} = A \otimes [\alpha_1 L_1 + \alpha_2 L_2 + \alpha_3 L_3] \tag{16}$$

其中，$\alpha_3 = 1 - \alpha_1 - \alpha_2$，$L_1 = \dfrac{1 - \mathrm{SSIM}(\boldsymbol{I}_0, \boldsymbol{I}_{-1\to0})}{2}$，$L_2 = \dfrac{1 - \mathrm{SSIM}(\boldsymbol{G}_0, \boldsymbol{G}_{-1\to0})}{2}$，$L_3 = |\boldsymbol{I}_0 - \boldsymbol{I}_{-1\to0}|$。

为更清晰地阐明梯度光度损失对亮度变化区域的有效性，可以为亮度转换建立一个线性模型[19]：

$$\boldsymbol{b} = m\boldsymbol{p} + n \tag{17}$$

其中，$\boldsymbol{p}$ 和 $\boldsymbol{b}$ 分别表示给定图像像素及其对应的亮度。在此公式下，像素的亮度差异主要由线性亮度参数 m 和 n 决定。对于传统方法，有亮度变化和没有亮度变化的像素差值为 $(m-1)\boldsymbol{p} + n$，对于梯度方法，值为 $(m-1)\nabla\boldsymbol{p}$，其中 ∇ 表示像素梯度。这表明梯度方法能通过消除亮度附加因子 n 来有效减轻亮度变化的影响。由于光亮敏感的无纹理区域上像素的梯度值非常小，基于梯度图计算的光度损失相比传统损失受亮度影响会大大减小，因而稳定性更强。

3　实验结果与分析

本节验证了本文方法所提出的模型能够在多样化光照环境中稳定生成高质量的深度图，并进一步设计消融

实验验证了本文方法改进工作所带来的具体成效。此外,将本文方法与近年来涌现的若干前沿深度估计方法进行对比,以评估本文方法的先进性。

3.1　数据集

1) KITTI 数据集是目前深度估计领域的主流评估基准数据集,具有多重任务属性,目标检测、目标分割、深度估计等多个领域都有广泛应用。KITTI 数据集包含源自农村地区、城市环境及高速公路等不同场景下的 394 个道路案例,不仅提供灰色及彩色图像,还附带地面真实深度图像,以满足多样化的研究与评估需求。

2) Make3D 数据集是单目深度估计的开拓性数据集,其包含了来自多个城市及自然区域的 534 组图像对。每组图像对由一张 RGB 图像及其对应的深度图像组成,旨在为深度估计任务提供丰富的视觉与深度信息。这些数据被划分为两个子集:包含 400 组图像的训练集,用于模型训练与参数优化;包含 134 组图像的测试集,用于评估模型的泛化能力与预测精度。目前,Make3D 数据集主要被用于各类泛化性实验,以验证不同单目深度估计方法的可靠性。

3.2　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。将式(16)中的参数 α_1、α_2、α_3 分别设为 0.75、0.10 和 0.15。

3.3　评估指标

基于之前的研究基础,本文方法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:

$$\frac{1}{|T|}\frac{\sum_{d}|d-d^*|}{d^*} \tag{18}$$

平方相对误差 Sq Rel:

$$\frac{1}{|T|}\frac{\sum_{d}\|d-d^*\|^2}{d^*} \tag{19}$$

均方根误差 RMSE:

$$\sqrt{\frac{1}{|T|}\sum_{d}\|d-d^*\|^2} \tag{20}$$

对数均方根误差 RMSE log:

$$\sqrt{\frac{1}{|T|}\sum_{d}\|\log d-\log d^*\|^2} \tag{21}$$

阈值精度满足条件:

$$\delta=\max\left(\frac{d_i}{d_i^*},\frac{d_i^*}{d_i}\right)<H_{thr} \tag{22}$$

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;设置阈值 $H_{thr}=1.25,1.25^2,1.25^3$。

3.4　实验结果分析

本小节中将对本文方法所提出的模型进行对比评估,为此本文方法挑选了一系列具有代表性的单目深度估计方法,在 KITTI 数据集与 Make3D 数据集上开展实验,通过分析实验结果展现本文方法在单目深度估计领域中取得的性能优化成果。表中 D 表示有监督方法,M 表示无监督方法;粗体标注每一项指标的最优结果。

根据表 1 所示,本文方法所构建的模型在 Make3D 数据集上各项指标均优于其他先进的无监督单目深度估计模型,甚至某些有监督模型。

表 1　在 Make3D 数据集上的测试结果对比

Table 1　Comparison of experimental results on the Make3D dataset

方法	监督方法	Abs Rel	Sq Rel	RMSE	RMSE ln
Laina[20]	D	**0.204**	**1.840**	**5.683**	**0.084**
Monodepth[8]	M	0.544	10.940	11.760	0.193
Zhou[7]	M	0.383	5.321	10.470	0.478
Monodepth2[9]	M	0.322	3.589	7.417	0.163
Li 等[15]	M	0.327	3.580	7.363	0.164
本文方法	M	0.264	2.251	6.595	0.132

从表 2 的数据分析可得,在相似分辨率下,本文方法所提出的模型在 KITTI 数据集上各项指标均优于其他无监督单目深度估计模型。即便与采用更高分辨率图像或依赖有监督学习的方法相比,本模型在多数评价指标上依然取得出了最优结果。Li 等[15]提出的模型和本文方法改进的模型均考虑在梯度图上进行光度损失以改善模型的光亮稳定性,本文方法提出的模型在误差指标和准确率指标上均优于前者,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降低了 33.6%、35.5%和 18.6%。

实验结果证明了本文方法提出的方案精确度更高,能够有效改善深度估计水平;同时本文方法模型展现出良好的泛化能力,能在不同数据集的测试中稳定保持优良发挥。

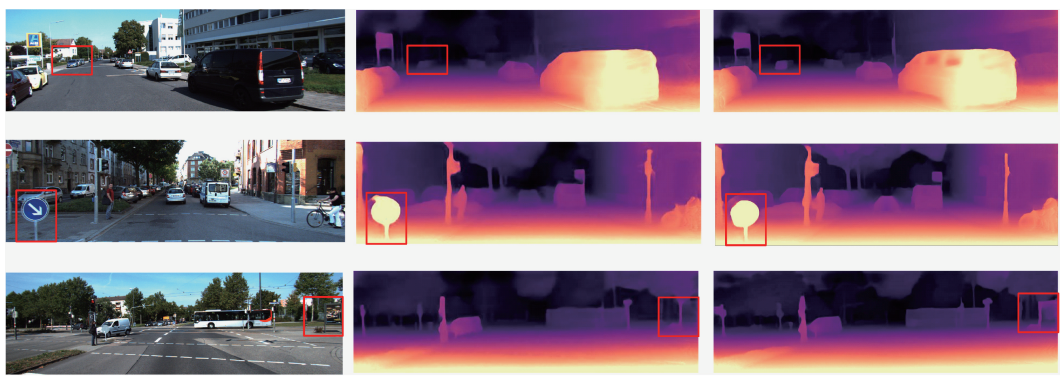
图 4 和 5 展示了本文方法在 KITTI 数据集与 Make3D 数据集上的深度可视化结果,以便直观地反映改进效果。图中深度值的大小通过亮度的变化来体现,具体而言,亮度由低到高对应着深度值由小到大的变化。

表 2　在 KITTI 数据集上的测试结果对比

Table 2　Comparison of experimental results on the KITTI dataset

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	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
Li 等[21]	620×188	D	0.113	—	4.687	—	0.856	0.962	**0.988**
Akada 等[22]	960×288	D	0.168	1.288	5.498	0.235	0.771	0.921	0.973
Monodepth2[8]	640×192	M	0.115	0.903	4.863	0.193	0.877	0.959	0.981
Li 等[15]	640×192	M	0.113	0.850	4.768	0.190	0.877	0.960	0.982
DynaDepth[23]	640×192	M	0.109	0.787	4.705	0.195	0.869	0.958	0.981
DIFFNet[24]	640×192	M	0.102	0.764	4.483	0.180	0.896	0.965	0.983
ManyDepth 等[25]	640×192	M	0.098	0.770	4.459	0.176	0.900	0.965	0.983
DynamicDepth[26]	640×192	M	0.096	0.720	4.458	0.175	0.897	0.964	0.984
DepthFormer[27]	640×192	M	0.090	0.661	4.149	0.175	0.905	0.967	0.984
MOVEDepth[9]	640×192	M	0.085	0.670	4.266	0.165	0.910	0.967	0.984
Chen 等[28]	1024×320	M	0.094	0.681	4.392	0.185	0.892	0.962	0.981
SGDepth[29]	1280×384	D+M	0.107	0.768	4.468	0.186	0.891	0.963	0.892
本文方法	640×192	M	**0.075**	**0.548**	**3.878**	**0.153**	**0.925**	**0.972**	0.986

在图 4 的 3 幅图中,本文方法展现出了显著的优势,不仅能够更为精确地刻画物体的边缘深度,让边缘更为分明,还在弱纹理区域取得更优效果。例如在第 1 张和第 2 张图中,本文方法预测的深度图中远处的车辆和路边的指示牌边缘分界清楚,形状轮廓更加清晰,细节更加准确。针对场景中的弱纹理区域,如第 3 张图中的宣传栏,使用本文方法在预测的深度图中明显能看到宣传栏的位置和大小,这在文献[9]方法中都不能实现。



(a) 原图
(a) Original image

(b)文献[9]方法
(b) Reference [9] method

(c) 本文方法
(c) The proposed method

图 4　在 KITTI 数据集上的可视化对比

Fig. 4　Visual comparison on KITTI dataset

如图 5 所示,与文献[9]方法相比不难发现,本文方法能够能够捕捉到更为丰富和细腻的细节信息,边缘深度把握更完整准确,且弱纹理区域的估计也更为准确。例如在第 1 张图中,对比本文方法与文献[9]方法的预测深度图,本文方法输出的深度图指示牌的形状更为准确完整。针对第 2 张图和第 3 张图来说,对于弱纹理区域的估计,如第 2 张图中的房屋和第 3 张图中的树冠,本文方法在整体上都比文献[9]方法更精准,细节捕捉更全面准确,甚至对于树叶缝隙间的深度也尝试进行估计。

3.5　消融实验

为了分别验证模糊掩膜和光度梯度在网络中的影响,本节进行了如表 3 的消融实验。从实验结果中可知,使用模糊掩膜能够过滤掉产生光度损失虚假信号的模糊像素,平方相对误差相对基准网络降低 3%;同时使用光度梯度损失能够有效改善弱纹理区域中亮度变化带来的光度损失不准确的问题,平均相对误差、平方相对误差和均方根误差等指标相对基准网络均有小幅降低。两者单独使用均能改善实验结果,但两者结合能相互补充,效果更佳,这一点在实验结果中能够得到印证,同时使用这两个组分使误差指标和准确率指标结果均有提高,其中平均相对误差和平方相对误差相对基准网络分别降低 3.8%和 3.3%。

前沿用的光度损失函数如式(5)所示,式中的 α 取值为 0.85。本文方法对光度损失进行了改进,因而根据

(a) 原图 (a) Original image　(b)文献[9]方法 (b) Reference [9] method　(c) 本文方法 (c) The proposed method

图 5　在 Make3D 数据集上的可视化对比

Fig. 5　Visual comparison on Make3D dataset

之前的经验将 α_1、α_2、α_3 以 0.85、0、0.15 为起始值进行变换，实验结果如表 4 所示。实验结果表明当 α_1、α_2、α_3 分别取 0.75、0.10、0.15 时实验取得最优结果，该设置更有利于网络预测精度的提升。

4　结　论

本文方法设计了一种能够提高深度估计框架鲁棒性的无监督单目深度估计网络，基于单帧深度估计结果进行多帧深度估计，并通过调整感受野和通道权重实现对图像深度特征的充分利用。在此基础上通过基于梯度图的光度损失来缓解弱纹理区域对光亮变化的敏感程度，并利用模糊掩膜减少输入图像中的模糊像素在边缘区域的光度损失下产生的不合理监督。实验结果表明，本文方法在不同的公开数据集上均能表现出优越性能，对弱纹理区域和边缘区域的深度估计也更加准确。本文方法仅针对无监督单目深度估计通常假设的静态场景，未来的研究需要更加关注运动和遮挡问题，设计一种适应性更强的网络。

表 3　验证各组分性能的消融实验结果

Table 3　The results of ablation experiments to verify the properties of each component

方法	网络改进	模糊掩膜	光度梯度损失	误差指标(越小越好)				预测准确率(越大越好)		
				Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
基准网络	–	–	–	0.085	0.670	4.266	0.165	0.910	0.967	0.984
本文方法	√	–	–	0.078	0.566	3.942	0.155	0.921	**0.972**	**0.986**
	√	√	–	0.077	0.549	3.897	0.155	0.923	**0.972**	**0.986**
	√	–	√	0.078	0.546	3.907	0.155	0.921	0.971	**0.986**
	√	√	√	**0.075**	**0.548**	**3.878**	**0.153**	**0.925**	**0.972**	**0.986**

表 4　验证光亮损失参数取值的消融实验结果

Table 4　The results of ablation experiments to verify the photometric loss parameter value

α_1	α_2	α_3	误差指标(越小越好)				预测准确率(越大越好)		
			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$
0.85	0	0.15	0.078	0.566	3.942	0.155	0.921	**0.972**	**0.986**
0.75	0.10	0.15	**0.075**	**0.548**	**3.878**	**0.153**	**0.925**	**0.972**	**0.986**
0.65	0.20	0.15	0.076	0.579	3.905	**0.153**	**0.925**	**0.972**	**0.986**
0.75	0.15	0.10	0.077	0.568	3.913	**0.153**	0.923	**0.972**	**0.986**
0.75	0.05	0.20	0.076	0.553	3.892	0.154	0.923	**0.972**	**0.986**

参考文献

[1] PENG J W, CAO ZH G, LUO X R, et al. Bokehme: When neural rendering meets classical rendering[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022: 16283-16292.

[2] 张耀, 吴一全, 陈慧娴. 基于深度学习的视觉同时定位与建图研究进展[J]. 仪器仪表学报, 2023, 44(7): 214-241.
ZHANG Y, WU Y Q, CHEN H X. Research progress of visual simultaneous localization and mapping based on deep learning [J]. Chinese Journal of Scientific Instrument, 2023, 44(7): 214-241.

[3] 郑自立, 徐健, 刘秀平, 等. 联合多注意力和 C-ASPP 的单目 3D 目标检测[J]电子测量与仪器学报, 2023, 37(8): 241-248.
ZHENG Z L, XU J, LIU X P, et al. Combined multi-attention and C-ASPP network for monocular 3D object detection[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(8): 241-248.

[4] 龙燕, 高研, 张广犇. 基于改进 HRNet 的单幅图像苹

果果树深度估计方法[J]. 农业工程学报, 2022, 38(23): 122-129.
LONG Y, GAO Y, ZHANG G B. The depth estimation method of single apple tree images based on improved HRNet [J]. Transactions of the Chinese Society of Agricultural Engineering, 2022, 38(23): 122-129.

[5] 廖钊宏, 张依晨, 杨飚, 等. 基于 Swin Transformer-CNN 的单目遥感影像高程估计方法及其在公路建设场景中的应用[J]. 测绘学报, 2024, 53(2): 344-352.
LIAO ZH H, ZHANG Y CH, YANG B, et al. Monocular remote sensing image elevation estimation method based on swin transformer-CNN and its application in highway construction scenarios[J]. Acta Geodaetica et Cartographcia Sinica, 2024 53(2): 344-352.

[6] 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, 27: 2366-2374.

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

[8] GODARD C, AODHA O M, FIRMAN M, et al. Digging into self-supervised monocular depth estimation [C]. Proceedings of the IEEE International Conference on Computer Vision, 2019: 3827-3837.

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

[10] ZHENG E, DUNN E, JOJIC V, et al. Patchmatch based joint view selection and depthmap estimation [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1510-1517.

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

[12] YANG N, STUMBERG L, WANG R, et al. D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 1281-1292.

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

[14] CHEN X Y, LI T H, ZHANG R N, et al. Frequency-aware self-supervised monocular depth estimation [C]. Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2023: 5808-5817.

[15] LI R, HE X, ZHU Y, et al. Enhancing self-supervised monocular depth estimation via incorporating robust constraints [C]. Proceedings of the 28th ACM International Conference on Multimedia, 2020: 3108-3117.

[16] HE K M, ZHANG X Y, REN SH Q, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[17] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117- 2125.

[18] HEISE P, KLOSE S, JENSEN B, et al. Pm-huber: Patchmatch with huber regularization for stereo matching[C]. Proceedings of the IEEE International Conference on Computer Vision, 2013: 2360-2367.

[19] WANG R, SCHWORER M, CREMERS D. Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 3903-3911.

[20] LAINA I, RUPPRECHT C, BELAGIANNIS V, et al. Deeper depth prediction with fully convolutional residual networks[C]. Proceedings of the Fourth International Conference on 3D Vision, 2016: 239-248.

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

[22] AKADA H, BHAT S F, ALHASHIM I, et al. Self-supervised learning of domain invariant features for depth estimation [C]. Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 2022: 3377-3387.

[23] ZHANG S, ZHANG J, TAO D CH. Towards scale- aware, robust, and generalizable unsupervised monocular depth estimation by integrating IMU motion dynamics [C]. Proceedings of the European Conference on Computer Vision, 2022: 143- 160.

[24] ZHOU H, GREENWOOD D, TAYLOR S. Self-supervised monocular depth estimation with internal feature fusion [C]. Proceedings of the 32nd British

Machine Vision Conference, 2021: 378-391.

[25] WATSON J, MAC A O, PRISACARIU V, et al. The temporal opportunist: Self-supervised multi-frame monocular depth [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021: 1164- 1174.

[26] FENG Z Y, YANGH L, JING L L, et al. Disentangling object motion and occlusion for unsupervised multi-frame monocular depth [C]. Proceedings of the European Conference on Computer Vision, 2022: 228-244.

[27] GUIZILINI V, AMBRUS R, CHEN D, et al. Multi-frame self-supervised depth with trans-formers [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022: 160-170.

[28] CHEN ZH, YE X Q, YANG W, et al. Revealing the reciprocal relations between self-supervised stereo and monocular depth estimation [C]. Proceedings of the IEEE International Conference on Computer Vision, 2021: 15529- 15538.

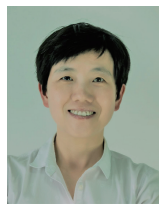
[29] KLINGNER M, TERMOHLEN J A, MIKOLAJCZYK J, et al. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance [C]. Proceedings of the European Conference on Computer Vision, 2020: 582-600.

作者简介

曲熠,江南大学硕士研究生,2021 年获江南大学学士学位,主要研究方向为计算机视觉与模式识别。

E-mail: 6211905004@ stu. jiangnan. edu. cn

Qu Yi is a M. Sc. candidate at Jiangnan University. She received her B. Sc. degree from Jiangnan University in 2021. Her main research interests include computer vision and pattern recognition.



陈莹(通信作者),江南大学物联网工程学院教授,2005 年于西安交通大学获得博士学位,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

Chen Ying (Corresponding author), professor in the college of Internet of Things Engineering, Jiangnan University. She received her Ph. D. degree from Xi'an Jiaotong University in 2005. Her main research interests include machine vision, information fusion, pattern recognition.