文章编号:1007－6735(2024)02－0129－09 DOI: 10.13255/j.cnki.jusst.20230906002

基于多掩膜技术的无监督深度与光流估计方法

张旭东[1]， 赵柏淦[2]， 吴国庆[1]， 姚建南[2]
（1. 南通大学 信息科学技术学院，南通 226019；2. 南通大学 机械工程学院，南通 226019）

摘要： 针对自动驾驶领域现有方法在处理动态、遮挡等复杂实际场景时存在的估计不准确问题，提出了一种以多掩膜技术为基础的无监督深度与光流估计方法，通过无监督学习从单目视频序列中提取目标深度、相机运动位姿和光流信息。根据不同外点类型设计了多种特定掩膜，以有效抑制外点对光照一致性损失函数的干扰，并在位姿估计和光流估计任务中起到剔除外点的作用。引入预训练的光流估计网络，协助深度和位姿估计网络更好地利用三维场景的几何约束，从而增强联合训练性能。最后，借助训练得到的深度和位姿信息，以及计算得到的掩膜，对光流估计网络进行了优化训练。在 KITTI 数据集上的实验结果表明，该策略能够显著提升模型的性能，并优于其他同类型方法。

关键词： 无监督学习；深度估计；位姿估计；三维重建
中图分类号： TP 391 **文献标志码：** A

Unsupervised depth and optical flow estimation method based on multiple mask techniques

ZHANG Xudong[1], ZHAO Baigan[2], WU Guoqing[1], YAO Jiannan[2]
(1. *School of Information Science and Technology, Nantong University, Nantong 226019, China*; 2. *School of Mechanical Engineering, Nantong University, Nantong 226019, China*)

Abstract: In response to the challenges posed by inaccurate estimations in handling dynamic objects, occlusions, and other complex real-world scenarios within the autonomous driving domain, an unsupervised depth and optical flow estimation method based on the multi-mask technique was proposed. This approach aimed to extract target depth, camera pose, and optical flow information from monocular video sequences through unsupervised learning. The method firstly designed a variety of special masks for different outlier types which effectively suppressed the interference of outliers on the photometric consistency loss function and played a key role in outlier removal during both pose and optical flow estimation tasks. Secondly, a pre-trained optical flow estimation network was introduced to assist the depth and pose estimation networks to fully utilize the geometric constraints of the 3D scene, thus enhancing the joint training performance. Finally, the optical flow estimation network was

收稿日期: 2023−09−06
基金项目: 国家自然科学基金资助项目(51805273)；江苏省青蓝工程和江苏省高等学校重点学科建设计划资助项目
第一作者: 张旭东(1985−)，男，博士研究生．研究方向：计算机视觉、数字信号处理与应用．E-mail：zhang.xd69@ntu.edu.cn
通信作者: 赵柏淦(1990−)，男，讲师．研究方向：计算机视觉、深度学习．E-mail：zhaobaigan@ntu.edu.cn

optimally trained with the help of the depth and pose information obtained from training, as well as the computationally obtained mask. Experimental results on the KITTI dataset showed that this strategy significantly improved the performance of the model and outperforms other methods of the same type.

Keywords: *unsupervised learning*; *depth estimation*; *pose estimation*; *3D reconstruction*

从视频序列中估计场景深度是目标识别领域中的重要内容。通过融合深度信息，目标识别系统可以更好地理解目标的三维结构和姿态，从而更准确地分割和识别目标对象[1]，这为自动驾驶系统提供了关键的感知和理解能力。传统的深度估计方法依赖于图像中的几何线索进行推算，对具有挑战性的环境比较敏感，如低纹理或光照变化强烈的场景[2]。而基于深度学习的方法在应对复杂环境时具有更好的适应性，它们将图像序列作为输入，然后通过神经元的协作形成非线性映射来输出深度图，其过程类似于人眼观察后，人脑通过神经元处理成我们所需要的高维信息[3]。但这样的系统需要大量的训练才能有出色的效果。有监督的方法利用高精度的传感器如激光雷达和惯导系统等获得真值，然后通过最小化模型和真值的差额来学习神经网络的映射关系。

尽管有监督的方法取得了出色的效果[4]，但其在训练模型时依赖于大量的真值数据，而真值数据在现实中需要昂贵的设备和费时费力的人工标注才能获得，这限制了模型性能的进一步提升。与之相对应的无监督的方法在训练模型时不需要真值，这意味着基于无监督的方法可以使用更广泛的训练数据，进而使用更大量级的数据和模型来训练，从而获得更好的性能。无监督深度估计方法的通用做法是利用视图合成过程生成用于训练模型的监督信号[5-10]，具体为利用深度神经网络预测目标视图的深度图以及目标视图与多个相邻源视图之间的相机运动位姿。一旦深度和位姿被估计出来，源视图就可以利用投影关系合成一个新的目标视图，最后通过最小化合成视图与目标视图之间的光度误差来训练整个框架[11-15]。但视图合成过程要求输入图像的场景是静态的，然而，在现实世界中的场景往往包含动态物体和遮挡区域等复杂情况，这不可避免地导致了训练的不稳定。因此，一些研究工作[10-11]已经提出了多种不同的掩膜方法和损失函数设计，以减轻外点(如动态物体或遮挡区域)对视图合成过程的影响，并取得了一定的效果。但这些方法往往忽略了外点对位姿估计任务以及光流估计任务的不利影响，而这些任务由于多任务联合训练的原因也会影响到深度估计的准确性。

针对这些问题，本文提出了一个新的无监督学习框架用于估计场景的深度和光流。首先，通过多种掩膜策略，以识别在前向传播过程场景中不同类型的外点。这些计算出的掩膜不仅用于在视图合成过程中排除外点的影响，还作为监督信号用于训练掩膜估计网络，网络经过训练后可以用于估计外点。然后，利用掩膜估计网络生成的掩膜对位姿估计网络的输入进行预处理，通过将可能包含外点的区域乘以较小的权重系数，以减轻外点对位姿估计的不利影响。此外，由于图像的光流、深度与位姿在3D几何中有着紧密的联系，引入一个经过预训练的光流估计网络，并提出了相应的3D几何约束损失，以进一步提升模型的整体性能。最后，在完成对深度和位姿估计网络的联合训练后，以无监督方式对光流估计网络进行优化训练。

1 研究方法

1.1 方法概述

单目视频序列蕴含着丰富的信息量。本文的目标在于以无监督的方式，从这些序列中学习图像的深度、相机的运动位姿，以及连续帧的运动光流。同时，通过多任务联合学习的方式，提高各任务模型在复杂场景中(如动态物体和遮挡区域)的鲁棒性。所提出的方法分为两个训练阶段：第一阶段着重训练深度估计网络、位姿估计网络以及掩膜估计网络；第二阶段是在第一阶段训练结果的基础上，对光流估计网络进行进一步优化。第一阶段的训练过程如图1所示，图中，$\boldsymbol{K}$表示相机的内参矩阵。输入为连续的两帧图像，分别称为目标视图和源视图。深度估计网络用于推算单帧图像的深度图；光流估计网络借助预训练

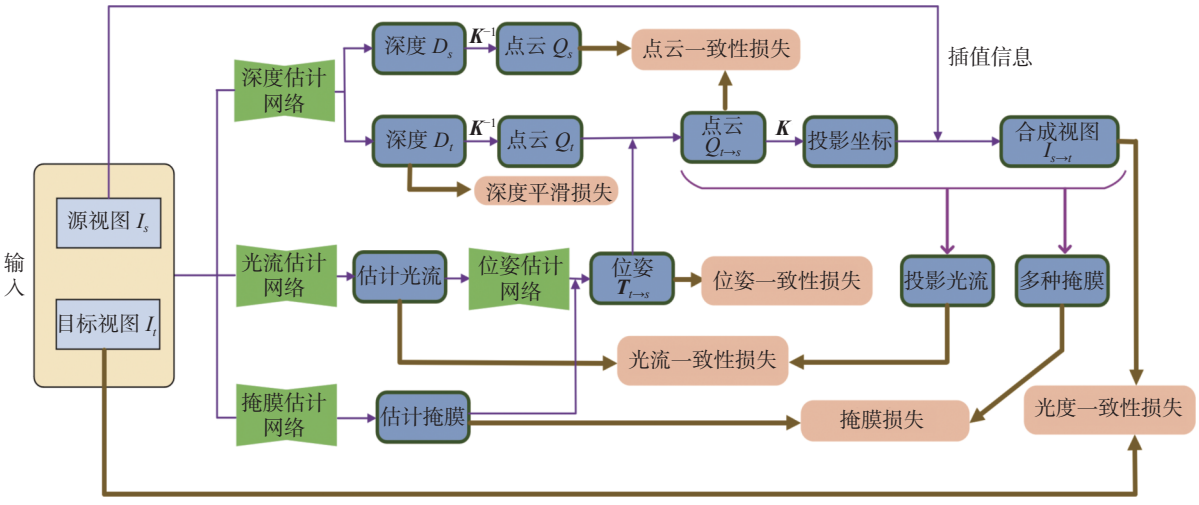

图 1　深度估计网络、位姿估计网络和掩膜估计网络联合训练示意图

Fig.1　Schematic of the joint training of depth estimation network, pose estimation network and mask estimation network.

的模型，计算目标视图和源视图之间的光流；掩膜估计网络同样以目标视图和源视图为输入，以估计场景中的外点，所得掩膜将被用于预处理光流。掩膜估计网络的监督信号主要源自视图合成过程中的掩膜计算，位姿估计网络则用于估计目标视图和源视图之间的相机运动位姿。这 3 种网络以联合训练的方式进行，所使用的损失函数为

$$\mathcal{L}_{\text{final}-l}=\sum_{l}(\mathcal{L}_{\text{pho}}+\omega_{\text{s}}\mathcal{L}_{\text{smo}}+\omega_{\text{m}}\mathcal{L}_{\text{mask}}+\omega_{\text{f}}\mathcal{L}_{\text{flo}}+\omega_{\text{i}}\mathcal{L}_{\text{icp}}+\omega_{\text{p}}\mathcal{L}_{\text{pos}}) \tag{1}$$

式中：$\mathcal{L}_{\text{pho}}$表示光度一致性损失，是训练网络主要的监督信号；$\mathcal{L}_{\text{smo}}$表示深度平滑损失，确保估计的深度值平滑；$\mathcal{L}_{\text{mask}}$表示掩膜损失，用于训练掩膜估计网络；光流一致性损失、点云一致性损失和位姿一致性损失通过 3D 场景的几何约束提升模型整体性能，分别用$\mathcal{L}_{\text{flo}},\mathcal{L}_{\text{icp}}$和$\mathcal{L}_{\text{pos}}$表示；$\omega_{\text{s}},\omega_{\text{m}},\omega_{\text{g}},\omega_{\text{f}},\omega_{\text{i}},\omega_{\text{p}}$分别表示相应损失函数所对应的权重；$l$是图像尺寸的缩放因子，与文献 [5-6] 的工作一样，深度估计网络在 4 个不同尺度上输出深度估计图，损失函数针对每个尺度分别计算。

1.2　训练深度和位姿估计网络的主要监督信号

视图合成过程如图 2 所示，通过视图合成过程生成的光度一致性损失是训练整个框架最重要的损失函数。图 2 中的像素点p_t投影到源视图上的坐标为p_s，通过p_s最近的 4 个点的像素值（左上、右上、左下和右下），可以获得重建点的像素值。

输入连续两帧图像（目标视图I_t和源视图I_s），通过深度估计网络估计目标视图的深度图为D_t，通过位姿估计网络估计的相机运动位姿为$\boldsymbol{T}_{t\to s}$，然后通过如下公式建立目标视图和源视图之间坐标的投影关系：

$$p_s\sim \boldsymbol{K}\boldsymbol{T}_{t\to s}D_t(p_t)\boldsymbol{K}^{-1}p_t \tag{2}$$

式中：p_t和p_s分别表示目标视图和源视图中像素的坐标。

由于投影后的坐标不是整数，因此需要在源视图上对每个点进行双线性插值来获取p_s处的像素值，即利用p_s周围的 4 个像素值进行双线性插值，从而得到合成的目标视图上的像素值。当图像场景是理想状态时，合成的目标视图$I_{s\to t}$应该与目标视图I_t光度一致，利用它们之间的光度差异可

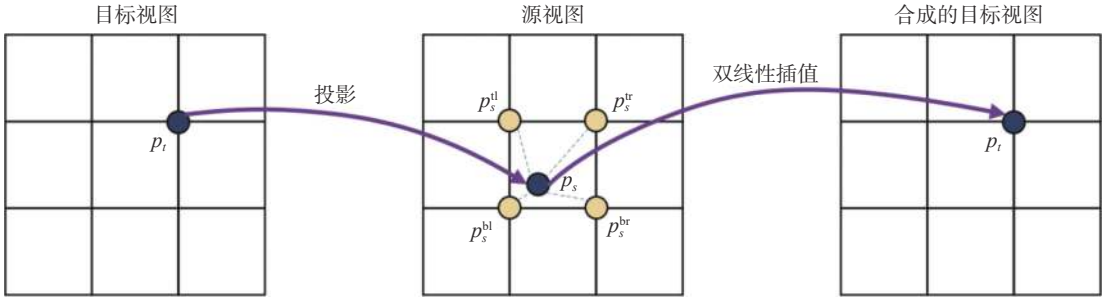

图 2　视图合成过程示意图

Fig.2　Illustration of the view synthess process

以构造训练整个框架的损失函数。与文献 [5-6] 的工作一样，使用 $L1$ 范数和结构相似性指数（structural similarity index measure, SSIM）的组合来度量合成视图和目标视图的光度差异，以 $\mathcal{L}_{\text{pe}}(I_t, I_{s\to t})$ 来表示。因此，总的光度一致性损失表达式为

$$\mathcal{L}_{\text{pho}} = \sum_s \left\{ \mathcal{L}_{\text{pe}}(I_t, I_{s\to t}) \right\} \tag{3}$$

在深度估计模型的训练中，通常使用边缘平滑损失来过滤出不正确的预测并保留清晰的细节。与文献 [6] 一样，使用的损失函数表达式为

$$\mathcal{L}_{\text{smo}} = \left|\partial_x d_t^*\right| \mathrm{e}^{-\left|\partial_x I_t\right|} + \left|\partial_y d_t^*\right| \mathrm{e}^{-\left|\partial_y I_t\right|} \tag{4}$$

式中，$d_t^* = d_t / \bar{d}_t$，d_t 表示目标图的深度值，$\bar{d}_t$ 表示深度的均值，d_t^* 表示均值归一化逆深度，可以阻止估计深度的收缩。

1.3 剔除外点的多种掩膜技术

视图合成过程不仅要求相邻帧图像上的点能够匹配，还要求场景是静态的同时相机是运动的[6]。因此，当用于训练的图像存在违反这些要求的场景时会抑制模型的训练。在使用视图合成过程计算损失函数时，需要考虑这些因素的不利影响，最直接的方法是设计掩膜来屏蔽这些外点，以防止它们参与损失函数的计算。

在视图合成过程中，当目标视图中的像素投影到源视图时，一些像素可能会投影到源视图的成像平面之外，从而无法准确地重建该点处的像素值。将所有投影到边界之外的点标记为 0，其他有效点标记为 1，可以生成一个掩膜，用边界掩膜 M_{b} 表示。通过与目标视图相乘，该掩膜可以避免一些边界点参与损失函数的计算。

场景中的遮挡区域也会影响模型的训练，当目标视图中的两个像素投影到源视图时，如果它们落在同一网格中，即它们有 4 个相同的插值点，则发生了遮挡。发生遮挡时，距离相机近的像素点会将距离远的像素点遮挡，因此，重构出的像素值是距离近的点的像素值，而距离远的则无法重构出。根据这两个像素点与相机之间的距离，将较远的点标记为 0，较近的点标记为 1，可以生成一个掩膜，用遮挡掩膜 M_{o} 表示。

由于符合视图合成过程要求的像素点，其光度误差在训练过程中将逐渐收敛到较低的值，而对于由各种不利因素导致的外点，光度误差值将始终保持在较高的值。利用这一特性，本文将那些光度误差远高于平均值的点判断为外点。具体来说，对于目标视图上的每个像素，可以根据以下表达式判断其是否为外点：

$$M_{\text{a}}(I_t, I_{s\to t}) = \begin{cases} 1, & \mathcal{L}_{\text{pho}}(I_t, I_{s\to t}) \leqslant \alpha \bar{\mathcal{L}}_{\text{pho}}(I_t, I_{s\to t}) \\ 0, & \mathcal{L}_{\text{pho}}(I_t, I_{s\to t}) > \alpha \bar{\mathcal{L}}_{\text{pho}}(I_t, I_{s\to t}) \end{cases} \tag{5}$$

式中：$\bar{\mathcal{L}}_{\text{pho}}$ 是所有光度误差的平均值；M_{a} 表示均值掩膜；α 是相应的权重，反映了对异常值的容忍度。该值越大，保留的异常值越多，本文设置为 1.5。

当物体与相机移动速度相近或相机静止时，这些场景同样违反了视图合成过程要求。针对这个问题，本文采用了文献 [6] 中的自动掩膜技术，即合成目标视图的光度误差应小于直接使用源视图计算得到的光度误差。该掩膜用 M_{s} 表示，其判断表达式如下：

$$M_{\text{s}}(I_t, I_{s\to t}) = \begin{cases} 1, & \mathcal{L}_{\text{pho}}(I_t, I_{s\to t}) < \mathcal{L}_{\text{pho}}(I_t, I_s) \\ 0, & \mathcal{L}_{\text{pho}}(I_t, I_{s\to t}) \geqslant \mathcal{L}_{\text{pho}}(I_t, I_s) \end{cases} \tag{6}$$

此外，文献 [6] 还提出了最小化投影技术用于解决场景中的遮挡问题，该方法在本质上也是一种掩膜技术。该掩膜用 M_{m} 表示，其判断表达式如下：

$$M_{\text{m}}(I_t, I_{s\to t}) = \begin{cases} 1, & \mathcal{L}_{\text{pho}}(I_t, I_{s\to t}) \leqslant \min\limits_s \mathcal{L}_{\text{pho}}(I_t, I_s) \\ 0, & \mathcal{L}_{\text{pho}}(I_t, I_{s\to t}) > \min\limits_s \mathcal{L}_{\text{pho}}(I_t, I_s) \end{cases} \tag{7}$$

最后，所有掩膜通过逐元素逻辑与运算组合起来形成最终的掩膜 M_{final}，其表达式如下：

$$\begin{aligned} M_{\text{final}}(I_t, I_{s\to t}) = & M_{\text{b}}(I_t, I_{s\to t}) M_{\text{o}}(I_t, I_{s\to t}) M_{\text{a}}(I_t, I_{s\to t}) \\ & M_{\text{s}}(I_t, I_{s\to t}) M_{\text{m}}(I_t, I_{s\to t}) \end{aligned} \tag{8}$$

得到最终的掩膜后，光度一致性损失函数通过如下表达式进行更新：

$$\mathcal{L}_{\text{pho}}(I_t, I_{s\to t}) = \sum_s \left(M_{\text{final}}(I_t, I_{s\to t}) \mathcal{L}_{\text{pe}}(I_t, I_{s\to t}) \right) \tag{9}$$

将计算得到的最终掩膜用作掩膜估计网络的监督信号。掩膜估计网络将目标视图和源视图作为输入，并生成一个估计外点的掩膜。为了方便模型的训练，估计掩膜的值不是二进制，而是连续值，范围在 0~1 之间。估计的掩膜通过直接与输入的光流逐元素相乘，可以给图像上的外点区域赋予一个较小的权重，从而避免了这些外点对位姿估计产生不利影响。掩膜估计网络通过最小化估计掩膜（M_{e}）和计算得到的掩膜（M_{final}）之间的差异来进行训练，损失函数定义如下：

$$\mathcal{L}_{\text{mask}}(M_{\text{final}}, M_{\text{e}}) = -\frac{1}{n}\sum_i \left[M_{\text{final}_i}\log M_{\text{e}_i} + \left(1 - M_{\text{final}_i}\right)\log\left(1 - M_{\text{e}_i}\right)\right] \quad (10)$$

式中：n 表示掩膜 M_{e} 和 M_{final} 的元素数量；M_{final_i} 表示掩膜 M_{final} 的第 i 个元素值；M_{e_i} 表示掩膜 M_{e} 的第 i 个元素值。图 3 展示了两个通过前向计算得到的掩膜和掩膜估计网络生成的掩膜对比示例。

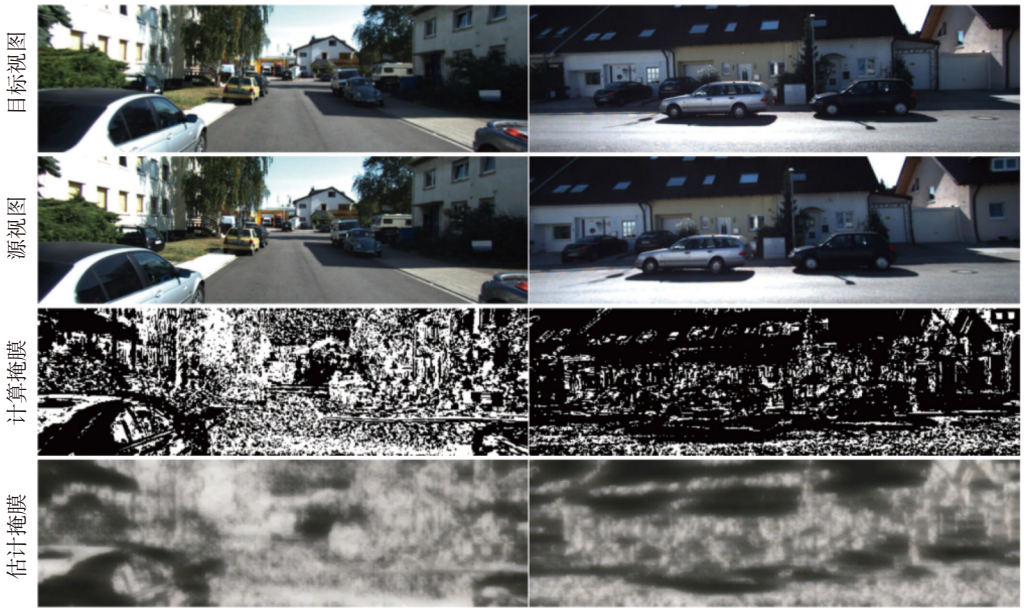

图 3　计算掩膜和估计掩膜可视化后的两个示例

Fig.3　Two examples of the visualization of calculation masks and estimation masks

1.4　3D 几何约束损失

针对光照变化显著的场景中光度一致性损失不成立的问题，本文还提出了光流一致性损失、点云一致性损失和位姿一致性损失。这些损失函数利用了不受光照变化影响的 3D 场景中固有的几何约束关系，作为光度一致性损失补充的监督信号。

为了在估计相机位姿时更好地提取图像的几何特征，本文使用光流估计网络输出的光流作为位姿估计网络的输入，估计到的光流（F_{e}）还可以为框架提供额外的监督信号。具体来说，利用估计的深度和估计的相机位姿，可以通过以下表达式计算投影光流（F_{pro}）：

$$F_{\text{pro}}(p_t) = \boldsymbol{K}\boldsymbol{T}_{t\to s}D_t(p_t)\boldsymbol{K}^{-1}p_t - p_t \quad (11)$$

计算得到的投影光流（F_{pro}）在理论上应与估计光流（F_{e}）在场景中静态区域内的大小一致。通过 $L1$ 范数表示的光流一致性损失如下所示：

$$\mathcal{L}_{\text{flo}} = \sum_{p_t\in V}\left\|F_{\text{pro}}(p_t) - F_{\text{e}}(p_t)\right\|_1 \quad (12)$$

式中，V 表示剔除外点后的有效区域。

通过投影过程，可以确定目标视图的深度图与源视图的深度图之间存在坐标对应关系。与其他方法[15]通过直接求差来比较差异不同的是，本文构造了点云一致性损失来惩罚不一致的深度估计。通过估计位姿和估计深度将目标视图的点云 Q_t 转换为源视图视角下的点云 $Q_{t\to s}$，然后通过 ICP（iterative closest point）算法计算点云 $Q_{t\to s}$ 和点云 Q_s 中对应点之间距离最小化的变换矩阵，其表达式如下：

$$\text{argmin}_{\boldsymbol{T}'}\left\{\frac{1}{2}\sum_i\left\|\boldsymbol{T}'Q^i_{t\to s} - Q_s^{c(i)}\right\|^2\right\} \quad (13)$$

式中：$c(i)$ 表示由 ICP 找到的点到点对应关系；$\boldsymbol{T}'$ 表示计算出的最佳变换矩阵。

ICP 的另一个输出是残差，其表达式如下：

$$r^i = Q^i_{t\to s} - \boldsymbol{T}'^{-1}Q_s^{c(i)} \quad (14)$$

残差 r^i 表示的是在应用变换矩阵 $\boldsymbol{T}'$ 之后，两个点云中对应点之间的残余距离。当估计的位姿和深度都完美时，点云 $Q_{t\to s}$ 和点云 Q_s 可以完美对齐。否则，利用 ICP 算法生成的变换矩阵和残差来调整位姿和深度，进行更好的对齐。具体来说，利用计算出的变换矩阵 $\boldsymbol{T}'$ 来近似估计位姿 $\boldsymbol{T}_{t\to s}$ 损失的负梯度，并利用计算出的残差 r^i 来近似估计深度 D_t 损失的负梯度。因此，点云一致性损失表示如下：

$$\mathcal{L}_{\text{icp}} = \left\|\boldsymbol{T}' - \boldsymbol{I}\right\|_1 + \left\|r^i\right\|_1 \quad (15)$$

式中：$\boldsymbol{I}$表示单位矩阵。

位姿一致性损失通过确保连续三帧序列中的位姿变换矩阵密切耦合来获得。具体来说，借助位姿估计网络，可以估计三帧图像两两之间的位姿信息，分别表示为$\boldsymbol{T}_{t-1\to t}$，$\boldsymbol{T}_{t\to t+1}$和$\boldsymbol{T}_{t-1\to t+1}$。利用它们之间的变换关系，可以得到位姿一致性损失，表达式如下：

$$\mathcal{L}_{\text{pos}}=\boldsymbol{T}_{t-1\to t}\boldsymbol{T}_{t\to t+1}-\boldsymbol{T}_{t-1\to t+1} \tag{16}$$

1.5 光流估计网络的优化训练

第二阶段的训练是在第一阶段训练结果的基础上，对光流估计网络进行进一步优化。首先，利用估计的光流和源视图，通过视图合成的过程来重建目标视图($I_{s\to t}^{\text{flow}}$)。将重建的目标视图与实际目标视图之间的差异作为光度一致性损失。此外，根据式(8)计算的外点掩膜同样适用于构建此处的重建损失。基于估计光流的光度一致性损失如下所示：

$$\mathcal{L}_{\text{flo-pho}}=\sum_{s}\left\{M_{\text{final}}(I_t,I_{s\to t})\mathcal{L}_{\text{pe}}\left(I_t,I_{s\to t}^{\text{flow}}\right)\right\} \tag{17}$$

通过使用第一阶段所估计的深度和位姿信息，以式(12)计算的刚性光流作为监督信号，用以指导光流估计网络的训练，其不一致性表示为$\mathcal{L}_{\text{flo-rig}}$。此外，本文将文献[9-11]中用于感知遮挡区域的光流前后向一致性约束引入到总的损失函数中，以对场景中的有效区域进行约束，其表达式如下所示：

$$\mathcal{L}_{\text{flo-con}}=\sum_{p_t\in V}\|F_{t\to s}(p_t)-F_{s\to t}(p_t+F_{t\to s}(p_t))\|_1 \tag{18}$$

最后，为保证估计光流的平滑性，本文参考文献[9-11]中使用光流的平滑损失，表示为$\mathcal{L}_{\text{flo-smo}}$，其表达式与深度平滑损失相同，都将图像的梯度作为先验约束添加到光流的梯度中。因此，第二阶段训练的总的损失函数表达如下：

$$\mathcal{L}_{\text{final-2}}=\mathcal{L}_{\text{flo-pho}}+\omega_{\text{rig}}\mathcal{L}_{\text{flo-rig}}+\omega_{\text{con}}\mathcal{L}_{\text{flo-con}}+\omega_{\text{smo}}\mathcal{L}_{\text{flo-smo}} \tag{19}$$

2 实验和结果

2.1 实验设置

本文框架包含4个子网络。对于深度估计网络，采用的是带有跳跃连接的编码解码结构，它以单个图像作为输入并输出相应的深度图。编码器采用ResNet18网络[16]，主要由级联的残差卷积神经网络组成，包含1 100万个可训练参数。解码器则由级联的反卷积层组成。对于位姿估计网络，同样采用的是ResNet18网络作为特征提取器，并在最终预测之前加入全局平均池化层以获得6个自由度的相机位姿(3个平移和3个旋转欧拉角)。由于使用的是预处理后的光流作为输入，因此网络的输入通道数设置为2。掩膜估计网络与深度估计网络具有相同的网络结构，仅输入部分不同，其输入为两帧图像在RGB通道上的堆叠，输出为每个像素在场景中是有效点的概率。光流估计网络采用PWC-Net模型[17]，该模型是一种紧凑但有效的光流模型，通过多尺度特征的串联实现光流估计的由粗到精，并在KITTI数据集上有着优秀性能。第一阶段的训练通过预先计算好光流可以加速框架的训练过程。

整体框架在PyTorch平台运行，所有实验都使用2个NVIDIA显卡(RTX 2080 Ti)和Intel Core i7 3.6 GHz CPU进行，并使用Adam进行优化。为了与其他工作[5-6]进行公平比较，图像分辨率裁剪为640×192，并且也使用了数据增强技术，如裁剪、随机缩放和水平翻转。对于所有基于ResNet18的模型均使用了ImageNet上预训练权重[18]来初始化编码器部分。第一阶段和第二阶段的实验参数经调试后的设置如表1所示。第一阶段的训练时间约为27 h，网络训练了22个epoch；第二阶段的训练时间约为32 h，网络训练了30个epoch。

表1 网络模型参数设置

Tab.1 Parameter setting for network model

参数名称	第一阶段的参数值	第二阶段的参数值
损失函数超参数	$\omega_{\text{s}}=0.002,\omega_{\text{m}}=0.2,\omega_{\text{f}}=0.35,$ $\omega_{\text{i}}=0.15,\omega_{\text{p}}=0.35$。	$\omega_{\text{rig}}=0.2,\omega_{\text{con}}=0.2,$ $\omega_{\text{smo}}=0.002$
批处理个数	8	8
初始学习率	0.000 1	0.000 1
学习率衰减	每5个epoch后乘以0.6	每5个epoch后乘以0.6

2.2 数据集和评估指标

本文在KITTI基准数据集[19]上进行了训练和测试，该数据集是自动驾驶领域中最受欢迎的数据集。它提供了56个驾驶场景，每秒10帧，图像分辨率约为1 226×370，涵盖城市、住宅区和高速公路等多种具有挑战性的复杂现实场景。此外，该数据集通过高精度激光雷达、惯导系统和差分GPS获得相机运动位姿和图像深度的真值。

对于深度估计评估，为了与其他无监督深度估计工作保持一致[5-9]，从而方便比较，本文采用了 Eigen 等[4] 分割 KITTI 数据集的方法对数据集进行了分割，然后以连续 3 帧为一对，设置的训练集包含 39 810 对图像，验证集包含 4 424 对图像，测试集则包含 697 个代表性帧。对于位姿估计评估，也采用了大部分研究的做法，即使用 KITTI Odometry 数据集中的 Sequences 00 至 08 进行训练，使用 Sequences 09 和 10 进行测试。对于光流估计评估，采用了 KITTI flow 数据集中带标签的 200 个训练图像作为测试集，将剔除掉与这 200 个图像相关场景后剩下的图像全都用作训练集。

与先前的研究一样[5-9]，位姿估计使用绝对轨迹误差（absolute trajectory error, ATE）[5] 进行评估。深度估计使用的指标包括绝对相对误差 Abs Rel、平方相对误差 Sq Rel、均方根误差 RMSE、对数均方根误差 RMSE log，以及预测精度 δ[6]。其中，误差指标值（Abs Rel，Sq Rel，RMSE，RMSE log）越低、精度指标值（δ）越高，则表示模型的性能越好。评估光流的指标为平均端点误差（endpoint error, EPE），计算的是估计光流与光流真值的欧式距离，评估光流的另外一个指标为 $F1$ 得分，表示的是在光流预测中，有多少比例的像素其端点误差（EPE）大于等于真实光流值的 3 像素，并且大于等于真实光流值的 5%[9]。EPE 和 $F1$ 的值越低，表示模型的性能越好。

2.3　在 KITTI 数据集上的性能评估

本文对单目深度估计模型和位姿估计模型在 KITTI 数据集上进行了评估。对于深度估计，本文使用 Eigen 等[4] 分割的 KITTI 数据集进行训练和测试，并与文献 [6] 一样，将每个预测的深度图与相应的真值进行对齐，深度上限设为 80 m。定性比较结果如图 4 所示，本文的模型与其他方法相比能更清晰地预测各种对象的边界，尤其是动态对象，如第一列和第二列中的自行车和汽车。这主要得益于在训练模型时考虑了场景中外点对模型训练的负面影响，并通过多种掩膜技术予以消除。在第三和第四列中，场景的光照变化明显，本文方法仍能很好地预测出物体的轮廓。这主要得益于在训练过程中，本文除了使用光照一致性损失外，还使用了其他不受光照变化影响的监督信号，从而对光照变化有了更好的鲁棒性。定量比较结果如表 2 所示，与其他同类型方法相比，本文的模型表现最好，尤其与文献 [6] 相比，在网络模型相同的前提下，本文的模型通过多种优化策略使预测精度得到了显著改进。

对于位姿估计，本文在 KITTI Odometry 数据集上进行了实验。为了公平比较，输入序列从 3 帧修改为 5 帧，与其他方法[5-6] 保持一致。定量比较结果如表 3 所示，本文方法取得令人满意的改进。一方面，其他方法在位姿估计中使用的是原始的 RGB 图像，其中包含了冗余信息，这些信息并不利于模型学习相机的运动位姿；另一方面，场景中存在许多干扰因素，比如遮挡和动态物体，这对位姿估计模型也产生了不利影响。而本文方法使用光流作为输入，其中包含相机的运

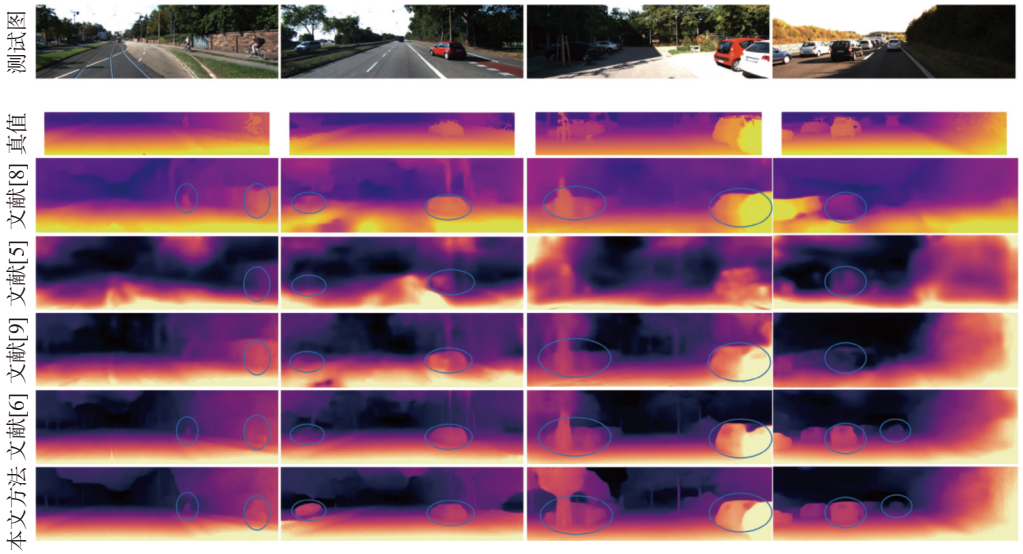

图 4　深度估计的定性比较结果

Fig.4　Comparison of the qualitative results for depth estimation

表 2 深度估计的定量比较结果

Tab.2 Comparison of the quantitative results for depth estimation

方法	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
文献[5]	0.208	1.768	6.856	0.283	0.678	0.885	0.957
文献[6]	0.154	1.218	5.699	0.231	0.798	0.932	0.973
文献[7]	0.144	1.391	5.869	0.241	0.803	0.928	0.969
文献[8]	0.163	1.240	6.220	0.250	0.762	0.916	0.968
文献[9]	0.155	1.296	5.857	0.233	0.793	0.931	0.973
文献[15]	0.137	1.089	5.439	0.217	0.830	0.942	0.975
本文方法	0.110	0.801	4.538	0.188	0.887	0.964	0.984

表 3 位姿估计的定量比较结果

Tab.3 Comparison of the quantitative results for pose estimation

方法	序列09的ATE	序列10的ATE
文献[2]	0.014 ± 0.008	0.012 ± 0.011
文献[5]	0.016 ± 0.009	0.013 ± 0.009
文献[8]	0.013 ± 0.010	0.012 ± 0.011
文献[9]	0.012 ± 0.007	0.012 ± 0.009
文献[15]	0.016 ± 0.007	0.015 ± 0.015
本文方法	0.007 ± 0.005	0.007 ± 0.005

动信息，更容易被模型所学习。此外，通过估计得到的掩膜对输入进行预处理可以有效避免外点的干扰，从而显著提高了位姿估计的准确性。

在光流估计方面，本文在 KITTI flow 数据集上进行了定量和定性实验，结果如表 4 和图 5 所示。与其他经典的光流估计网络相比，本文的方法表现出色。尤其是与文献 [17] 的方法相比，本文在基于其网络架构和预训练模型的基础上，通过增加深度信息、位姿信息和外点信息，设计了 4 种不同的损失函数，以进一步优化预训练的光流估计网络。定性和定量的结果均展示了优化策略的显著效益。

表 4 光流估计的定量比较结果

Tab.4 Comparison of the quantitative results for optical flow estimation

方法	训练集		测试集
	EPE	*F*1	*F*1
文献[9]	10.81	—	—
文献[10]	8.80	28.94%	29.46%
文献[11]	8.98	26.01%	27.70%
文献[17]	8.15	24.35%	—
本文方法	7.84	23.67%	23.51%

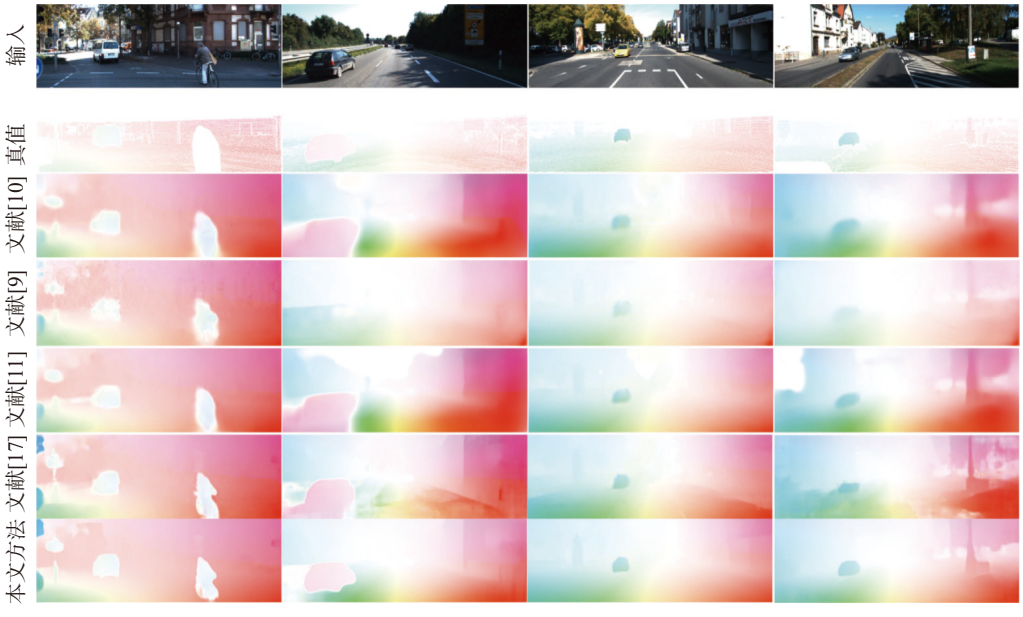

图 5 光流估计的定性比较结果

Fig.5 Comparison of the qualitative results for optical flow estimation

3　结束语

提出了一种基于多掩膜技术的无监督深度和光流估计方法。通过多种掩膜技术，有效降低了现实复杂场景中大量外点对视图合成、位姿估计和光流估计的负面影响。同时，利用深度、位姿和光流 3 个任务之间的相关性，先是通过预训练的光流模型辅助深度和位姿的训练，再利用训练后的深度和位姿对光流模型进行优化训练，从而促进三者性能的提升。在 KITTI 数据集上的实验结果表明，该方法表现出良好的效果，并在与同类型方法的比较中展现出一定的先进性。未来的研究可考虑引入 Transformer 等大容量模型进行框架设计，并利用更多数据进行训练以进一步提升模型性能。

参考文献：

[1] 仇旭阳, 黄影平, 郭志阳, 等. 基于深度学习的障碍物检测与深度估计 [J]. 上海理工大学学报, 2020, 42(6): 558–565.

[2] MUR-ARTAL R, MONTIEL J M M, TARDÓS J D. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147–1163.

[3] 徐天意, 夏明, 李峰, 等. 基于深度学习的多模态气管插管智能目标检测 [J]. 上海理工大学学报, 2021, 43(5): 436–442,483.

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

[5] ZHOU T 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. Honolulu, 2017: 6612–6619.

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

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

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

[9] YIN Z C, SHI J P. GeoNet: unsupervised learning of dense depth, optical flow and camera pose[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 1983–1992.

[10] MEISTER S, HUR J, ROTH S. UnFlow: unsupervised learning of optical flow with a bidirectional census loss[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, 2018: 888.

[11] ZOU Y L, LUO Z L, HUANG J B. DF-Net: unsupervised joint learning of depth and flow using cross-task consistency[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, 2018: 38–55.

[12] 李文举, 李梦颖, 崔柳, 等. 基于金字塔分割注意力网络的单目深度估计方法 [J]. 计算机应用, 2023, 43(6): 1736–1742.

[13] 吴俊贤, 何元烈. 基于通道注意力的自监督深度估计方法 [J]. 广东工业大学学报, 2023, 40(2): 22–29.

[14] 蒲正东, 陈姝, 邹北骥, 等. 基于高分辨率网络的自监督单目深度估计方法 [J]. 计算机辅助设计与图形学学报, 2023, 35(1): 118–127.

[15] BIAN J W, LI Z C, WANG N Y, et al. Unsupervised scale-consistent depth and ego-motion learning from monocular video[C]//Proceedings of the 33rd Annual International Conference on Neural Information Processing Systems. Vancouver, 2019: 4.

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

[17] SUN D Q, YANG X D, LIU M Y, et al. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 8934–8943.

[18] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211–252.

[19] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Providence, 2012: 3354–3361.

（编辑：丁红艺）