Modeling and Simulation 建模与仿真, 2025, 14(2), 389-399
Published Online February 2025 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2025.142160

Hans汉斯

解耦动态区域的自监督单目深度估计模型

秦晓飞，朱勇超，侯　闯，李欣怡，诸靖宇

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2025年1月25日；录用日期：2025年2月18日；发布日期：2025年2月27日

摘　要

近年来，自监督单目深度估计因其无需深度标签的优势，在计算机视觉领域获得了广泛关注。然而，传统自监督单目深度预测方法通常基于静态场景假设，这导致在相邻帧中出现动态对象时，深度预测的精度会显著下降。为了解决这一问题，本文提出了一种多帧自监督单目深度估计模型。该模型通过分割网络预先识别图像中的运动物体，并利用多帧图像之间的光流信息来重构图像。通过将静态场景与动态物体分开处理，该方法有效提高了动态物体深度估计的准确性。此外，本文设计了动态物体重构损失(Dynamic Object Reconstruction Loss, DRL)和深度一致损失(Depth Consistency Loss, DCL)，以监督动态重构图和重构深度图的生成。实验结果表明，在三个公共数据集上，该方法优于现有的主流方法，能够在动态场景中准确预测深度图。

关键词

计算机视觉，单目深度估计，自监督学习，动态场景

Self-Supervised Monocular Depth Estimation Model with Decoupled Dynamic Regions

Xiaofei Qin, Yongchao Zhu, Chuang Hou, Xinyi Li, Jingyu Zhu

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Jan. 25th, 2025; accepted: Feb. 18th, 2025; published: Feb. 27th, 2025

Abstract

Recent years, self-supervised monocular depth estimation has garnered extensive attention in the field of computer vision due to its advantage of not requiring depth labels. However, traditional self-

文章引用: 秦晓飞, 朱勇超, 侯闯, 李欣怡, 诸靖宇. 解耦动态区域的自监督单目深度估计模型[J]. 建模与仿真, 2025, 14(2): 389-399. DOI: 10.12677/mos.2025.142160

supervised monocular depth prediction methods are typically based on the assumption of static scenes, which leads to a significant decrease in depth prediction accuracy when dynamic objects appear in consecutive frames. To address this issue, this paper proposes a multi-frame self-supervised monocular depth estimation model. The model identifies moving objects in images through a segmentation network and reconstructs images using optical flow information between multiple frames. By separating static scenes from dynamic objects, this approach effectively improves the accuracy of depth estimation for dynamic objects. Additionally, this paper have designed the Dynamic Object Reconstruction Loss (DRL) and Depth Consistency Loss (DCL) to supervise the generation of dynamic reconstruction images and reconstructed depth maps. Experimental results demonstrate that this method outperforms existing mainstream approaches on three public datasets, enabling accurate depth prediction in dynamic scenes.

Keywords

Computer Vision, Monocular Depth Estimation, Self-Supervised Learning, Dynamic Scenes

1. 引言

三维场景信息在自动驾驶[1]、机器人[2]和增强现实[3]等领域具有重要意义。然而，由于深度传感器成本较高，利用神经网络从图像中估计深度图像成为一种有效的解决方案。早期的深度估计模型需要真实的深度标签进行训练，但收集大量准确深度信息的数据既昂贵又费力。为了解决这一问题，Godard [4]等人利用立体图像对单目视频的几何信息，通过最小化源图像与目标图像之间的重投影误差来优化模型。这种方法仅需相邻帧即可进行模型训练，具有巨大的实际应用潜力。因此，本文主要研究单目视频的自监督深度估计。

在训练过程中，自监督单目深度估计模型包括深度网络和姿态网络。深度网络用于预测深度图，姿态网络用于估计相机的位姿。在估计出深度图和位姿后，利用 n 点透视法[5]重建图像并计算光度损失以更新神经网络。这些方法在室外数据集[6]-[8]上表现出色，但单目深度估计的训练面临着动态场景的挑战。

自监督深度估计模型是基于静态环境构建的，这在大多数现实世界场景中并不成立。动态物体将违反自监督深度估计模型的假设，并导致重投影不匹配问题。动态物体区域中的深度损失值无法反映真实的损失值，这将误导模型训练。静态场景的假设在 KITTI [9]等数据集中是有效的，在这些数据集中，建筑物、停放的车辆、树木等静态物体主导了场景，使神经网络能够通过静态区域学习运动物体的深度。然而，当使用具有更多运动物体的数据集时，如 nuScenes [10]和 DDAD [11]，相机的位姿估计会受到影响，导致估计的深度图精度降低。最近的工作[12] [13]试图优化动态对象区域的深度预测，并取得了显著的改进，但它们仍然存在几个缺点。它们仅在损失函数级别解决了失配问题，仍然无法通过动态物体的时间帧来推理几何约束，这限制了其准确性潜力，并且忽略了刚性物体的运动一致性。此外，静态区域应该没有运动，这种零运动约束可以促进运动物体的运动估计。

本文提出了一种适用于动态场景的自监督单目深度估计模型。为了处理动态场景，本文引入分割网络将动态物体分割出来，形成一个掩码。然后，利用光流网络估计相邻两帧图像的运动。通过结合动态物体掩码和两帧图像的运动信息，重构源图像，并用该重构源图像与深度重构图像进行相互监督。此外，为了增强光流网络对运动物体的运动估计能力，本文提出了一种动态物体重构损失。然后，重构的目标

图像通过共享权重的深度估计网络生成深度图，并用重构目标图像的深度图与目标图像的深度图进行相互监督。本文的方法重点在于损失函数的校正和相应的解耦训练过程，因此几乎可以应用于任何自监督单目深度估计模型。

2. 研究方向

2.1. 整体架构

为了解决动态物体在传统自监督单目深度估计中估计不准确的问题，本文提出了一种自监督训练模型，整体框架如图 1 所示。本文首先使用 Depth Net 和 Pose Net 得到深度和姿态先验信息，通过这两个先验信息重构目标图像。由于目标图像中存在深度估计不准确的动态物体，因此本文引入 Segmentation Net 从目标图像中分割出动态物体掩码。然后将目标图像和源图像输入 Flow Net 得到两帧之间的光流，由于光流能够直接的表示两帧之间的物体运动，因此本文利用动态物体掩码和光流还原动态物体的原始位置，还原原始位置的图像本文称其为重构源图像，本文用重构源图像监督重构目标图像以此来矫正动态物体的图像信息。最后为了深度信息的一致性，本文把重构源图像输入共享权重的 Depth Net 中得到重构源图像的深度图，并且用重构源图像的深度图和先验深度图相互监督。在本文的模型中提出了两个损失以保证监督的有效性。

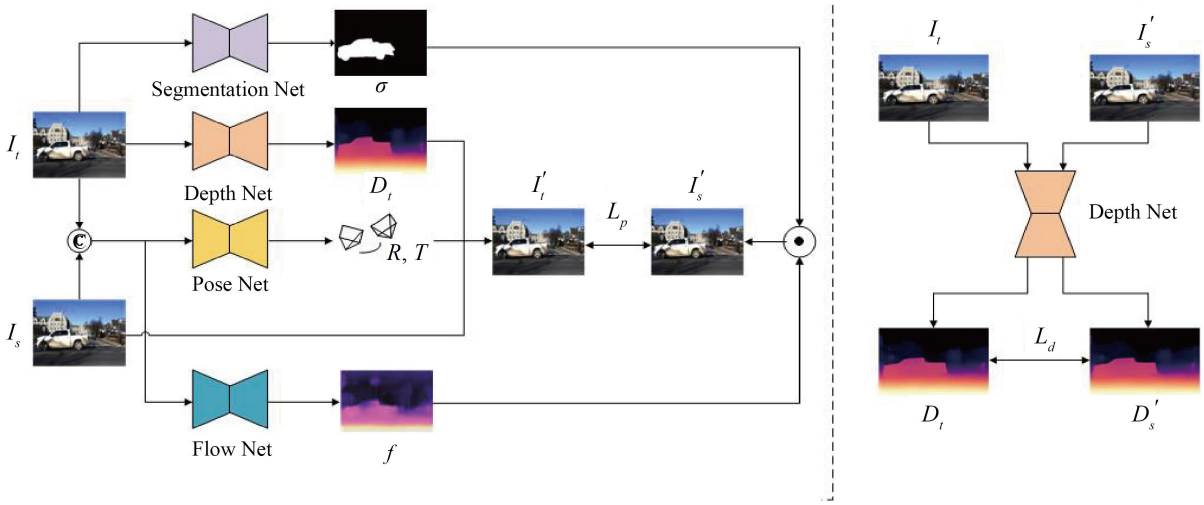

Figure 1. The overview of model framework
图 1. 模型框架总览

2.2. 先验知识

为了便于介绍，本文在本小节介绍一下自监督单目深度估计的基础知识。自监督单目深度估计是基于对极几何的原理构建的。当训练深度估计模型时，需要先找到相邻的两帧图片，然后得到两帧图片的相对位姿。有了深度信息和相对位姿之后就能够重构图像，用图像中一个点为例，源图像上的点 p_s 可由目标图像上的点 p_t 通过以下公式得到。

$$p_s \sim K\left[R_{t\to s}D_t\left(p_t\right)K^{-1}+T_{t\to s}\right] \tag{1}$$

公式中的 D_t 是 Depth Net 估计的深度， $R_{t\to s}$ 以及 $T_{t\to s}$ 分别是目标图像到源图像的旋转矩阵和平移矩阵，K 是相机内参矩阵， K^{-1} 是相机内参矩阵的逆矩阵。

从对极几何的坐标转换公式可以看出，传统的自监督单目深度估计模型只考虑了相邻两帧之间的相机变换，当图像中有动态物体时如图 2 所示。目标图像上一点经过 Depth Net 后的深度为 D_t，此深度经过对极几何得到重投影到源图像位置上，此时这一个点的深度为 D_s'，但是由于汽车时运动的，上一帧汽车的位置应该在此位置之前，即真实深度应该为 D_s，因此因为汽车的自我运动导致传统的自监督单目深度估计方法会出现 D_d 的深度误差。本文主要就是解决此问题。

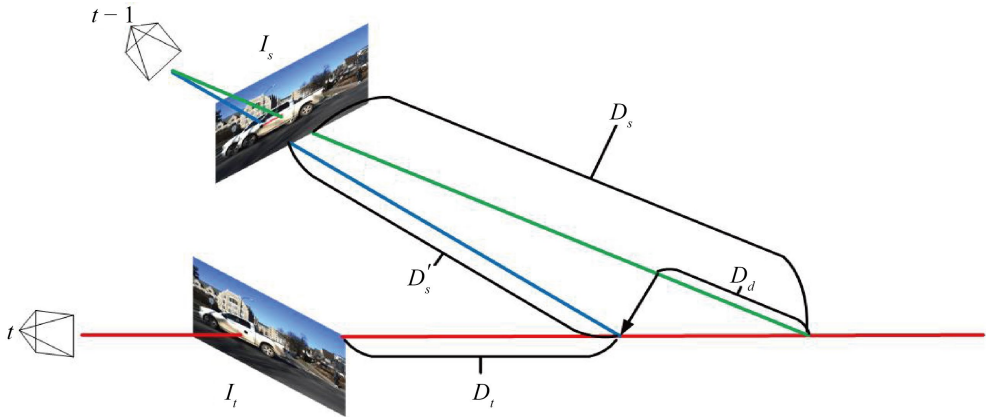

Figure 2. Schematic diagram of depth estimation error for dynamic objects
图 2. 动态物体深度错误估计原理图

2.3. 运动物体解耦

由于传统的自监督单目深度估计模型需要使用 Depth Net 得到的深度图 D_t 来重构目标图像，但是因为传统的监督方式在动态物体区域会把错误的深度也认定为正确的，为了解决动态物体导致的深度估计不准确问题，本文采用动态物体与静态场景分开处理的方法。首先利用 COCO 数据集上预训练好的分割网络把目标图像 I_t 中的动态物体分割出来得到动态物体掩码 σ。之后为了得到目标图像到源图像的自我运动，把这两帧图片输入 Flow Net 中得到目标图像到源图像的光流 $f_{t\to s}$。使用动态物体掩码和光流能够用不同于公式 1 的方法重构源图像以此解决用不准确深度图重构图像的情况。

$$p_s' = p_s + f_{t\to s}\cdot\sigma \tag{2}$$

公式中 p_s 是源图像上的点，p_s' 是重构源图像上动态物体的新位置，$\cdot$ 表示逐点相乘。

由于在传统的自监督单目深度估计当中没有重构源图像这个操作，为了能够很好的利用重构源图像本文提出了额外的两个损失来促进网络的训练。

在动态区域本文使用了光流来进行重构，因为光流对于每个点的运动估计各不相同，然而在现实生活当中大多数的动态物体都是刚性的，因此在训练过程当中只需要考虑平移运动。

$$L_e = \left|\partial x\left(\overline{f_{t\to s}}\right)\cdot\sigma\right| + \left|\partial y\left(\overline{f_{t\to s}}\right)\cdot\sigma\right| \tag{3}$$

公式中 ∂x，∂y 分别为 x 方向和 y 方向的导数，$\overline{f_{t\to s}}$ 表示动态区域光流的平均值。

传统自监督单目深度模型对于静态区域的良好性能是无法否认的，因此在静态区域直接使用 L_1 损失进行监督。

$$L_s = \left|I_t' - I_s'\right|\cdot\hat{\sigma} \tag{4}$$

公式中的 $\hat{\sigma}$ 表示取动态区域相反的区域。

因此动态物体重构损失为

$$L_p = \alpha L_s + \beta L_e \tag{5}$$

公式中 α 和 β 为超参数，经过实验调整最后的值分别为 0.65 和 0.35。

2.4. 深度一致损失

如图 1 右侧所示，把目标图像 I_t 和重构源图像 I'_s 输入共享权重的 Depth Net 中分别得到了相应的深度图并且让他们相互监督，由于本文的动态物体解耦只重构了源图像的动态区域，因此本文提出了动态区域一致性损失 L_m 来增强 D_t 和 D'_s 的动态区域一致性，并且只有在不一致性大于阈值后才会激活。

$$A = \frac{\left|D_t - D'_s\right|}{\min\left\{D_t, D'_s\right\}} > 1 \tag{6}$$

$$L_m = \sum_{i\in(\sigma\cap A)} \left|D_t^i - D_s'^{,i}\right| \tag{7}$$

在运动物体解耦中，为了重构源图像 I'_s，本文直接使用光流对动态区域进行位移，这会导致在源图像 I_s 原来的动态区域出现部分丢失。为了解决这一个问题本文同时重构了目标图像 I_t 的前后帧 $\left(I_{t-1}, I_{t+1}\right)$，在前一帧丢失的部分通常在都会出现在后一帧当中，本文应用此现象提出了静态区域缺失损失 L_o。

$$L_o = \sum_{i\in\hat{\sigma}} \min\left|D_t^i - D_a^i\right| \tag{8}$$

公式中 a 表示目标图像 I_t 的前后帧 $\left(I_{t-1}, I_{t+1}\right)$。

最后深度一致性损失 L_d 是由动态区域一致性损失 L_m 和静态区域缺失损失 L_o 相加得到。

$$L_d = L_m + L_o \tag{9}$$

2.5. 模型损失

与依赖于真实标签监督的单目深度估计方法不同，自监督单目深度估计方法主要涉及图像重建[14]，其中目标图像用于监督重建目标图像。除了图像重建损失外，还需利用边缘感知平滑度损失[15]来获得深度图的合理平滑度。

图像重构损失：首先使用源图像重建目标图像。

$$I'_t = F\left(I_s, P, D_t, K\right) \tag{10}$$

公式中表示 F 重构过程，是 P 估计的相机位姿。在获得重构目标图像之后，使用结构相似性指数(SSIM)和 L_1 损失测量重构目标图像与目标图像的距离。

$$L_r\left(I_t, I'_t\right) = \alpha\frac{1 - SSIM\left(I_t, I'_t\right)}{2} + \left(1-\alpha\right)\left|I_t - I'_t\right| \tag{11}$$

公式中为超参数，设置为 0.85。

边沿平滑损失：为了防止估计的深度图过于平滑和粗糙，本文采用了边沿平滑损失。

$$L_{\text{smooth}} = \left|\partial_x d_t^*\right| e^{-\left|\partial_x I_t\right|} + \left|\partial_y d_t^*\right| e^{-\left|\partial_y I_t\right|} \tag{12}$$

公式中 ∂_x，∂_y 分别是 x 和 y 方向上的导数，$d_t^* = d_t / d'_t$ 表示平均归一化逆深度。

总损失：本文的模型是一个多尺度编码器－解码器结构，从三个尺度共同计算损失。

$$L_{total} = \frac{1}{3}\sum_{s\in\left(1,\frac{1}{2},\frac{1}{4}\right)}\left(L_r + L_d + L_p + \psi L_{\text{smooth}}\right) \tag{13}$$

公式中 s 表示图像的尺度， ψ 是超参数由实验最终设置为 $1/10^3$。

3. 实验结果与分析

3.1. 数据集与评价指标

nuScenes：来自 Motional 的 nuScenes [10]数据集是一个大规模的自动驾驶数据集，提供来自 6 个摄像头、1 个激光雷达、5 个雷达、GPS 和 IMU 的全面传感器数据。它包含 140 万张图像、390K 激光雷达扫描、140 万次雷达扫描和 140 万个物体边界框。提供了 23 个类别的 2Hz 精确 3D 注释以及对象属性。这些数据涵盖了波士顿和新加坡的不同地点，跨越了不同的时间和天气条件。

KITTI：KITTI 和 TTIC 创建的 KITTI [9]数据集是自动驾驶的基准。它以现实世界的城市、农村和高速公路场景为特色，每张图像最多有 15 辆汽车和 30 名行人。该数据集包括 389 个立体和光流对、39.2 公里的视觉里程计、200K+3D 注释图像(10 赫兹)、全景图像、激光雷达数据和导航信息。深度预测通常在[0, 80]米以内。

DDAD：TRI 提供的 DDAD [11]数据集支持自动驾驶的密集深度估计研究。它包含来自跨大陆自动驾驶车队的单目视频和激光雷达生成的深度数据。场景覆盖美国和日本城市，支持长距离(高达 250 米)深度估计。它有助于自动驾驶的精确障碍物检测和场景理解。

评价指标：为了定量地评价深度估计模型，本文使用以下常用的度量作为评价指标：绝对相对误差(Absolute Relative Error, Abs Rel)、平方相对误差(Square Relative Error, Sq Rel)、均方根误差(Root Mean Square Error, RMSE)、均方根对数误差(Root Mean Square Logarithmic Error, RMSElog)、$\xi<1.25$，$\xi<1.25^2$，$\xi<1.25^3$。

$$\mathrm{AbsRel}=\frac{1}{|N|}\sum_{i\in N}\frac{\left|D_i-D_i^*\right|}{D_i^*} \tag{14}$$

$$\mathrm{SqRel}=\frac{1}{|N|}\sum_{i\in N}\frac{\left\|D_i-D_i^*\right\|}{D_i^*} \tag{15}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{|N|}\sum_{i\in N}\left\|D_i-D_i^*\right\|^2} \tag{16}$$

$$\mathrm{RMSE\ log}=\sqrt{\frac{1}{|N|}\sum_{i\in N}\left\|\log\left(D_i\right)-\log\left(D_i^*\right)\right\|^2} \tag{17}$$

$$\max\left(\frac{D_i}{D_i^*},\frac{D_i^*}{D_i}\right)=\xi<\mathrm{thr} \tag{18}$$

这里 D_i 是像素点 i 的深度， D_i^* 是真实标签， N 表示真实标签点的数量， thr 表示阈值。

3.2. 实验细节

该方法在 PyTorch 中实现，并在两块 NVIDIA 3090 GPU 上进行训练。Depth Net 初始学习率为 10^{-4}，在 2 个迭代周期后降至 5×10^{-5}，而其他部分的初始学习率都设置为 5×10^{-5}。本文使用带有线性调度器的 AdamW [16]优化器来衰减学习率，模型总共训练了 25 个 epoch。采用图像翻转和抖动的数据增强方法来增强模型的泛化能力。本模型的基础 Depth Net 使用的是深度估计性能优秀且网络轻量的 Lite-Mono [6]，Flow Net 使用的是 RAFT [17]，Pose Net 则是用了基本的 Resnet18。

Table 1. Experimental results on the nuScenes dataset. “OM” denotes whether moving objects are considered, and “-” indicates that the data is not provided in the paper
表 1. nuScenes 数据集上的实验结果。“OM”表示是否考虑运动物体，“-”表示文章中没有提供该数据

方法	OM	Abs Rel	Sq Rel	RMSE	RMSElog	$\xi < 1.25\uparrow$	$\xi < 1.25^2\uparrow$	$\xi < 1.25^3\uparrow$
Monodepth2 [18]		0.425	16.592	10.040	0.402	0.723	0.827	0.887
Packnet [19]		0.309	2.891	7.994	-	0.547	0.796	0.899
FSM [20]		0.334	2.845	7.786	-	0.580	0.761	0.894
VoluFusion [21]		0.271	-	7.391	-	0.726	-	-
zeroDepth [11]		0.236	-	7.054	-	0.747	-	-
Lite-Mono [6]		0.491	15.578	9.807	0.449	0.720	0.831	0.879
SurroundDepth [7]		0.245	3.030	6.835	-	0.719	0.878	0.935
ZoeDepth [22]		0.504	-	7.717	-	0.255	-	-
WSGD [23]		0.176	1.603	6.036	0.245	0.750	0.912	0.963
MonoProb [24]		0.219	-	9.559	-	0.684	-	-
Dynamo (MD2) [8]	√	0.193	2.285	7.357	0.287	0.765	0.885	0.935
Dynamo (LM) [8]	√	0.179	2.118	7.050	0.271	0.787	0.896	0.940
Ours	√	0.152	1.627	7.083	0.252	0.792	0.916	0.965

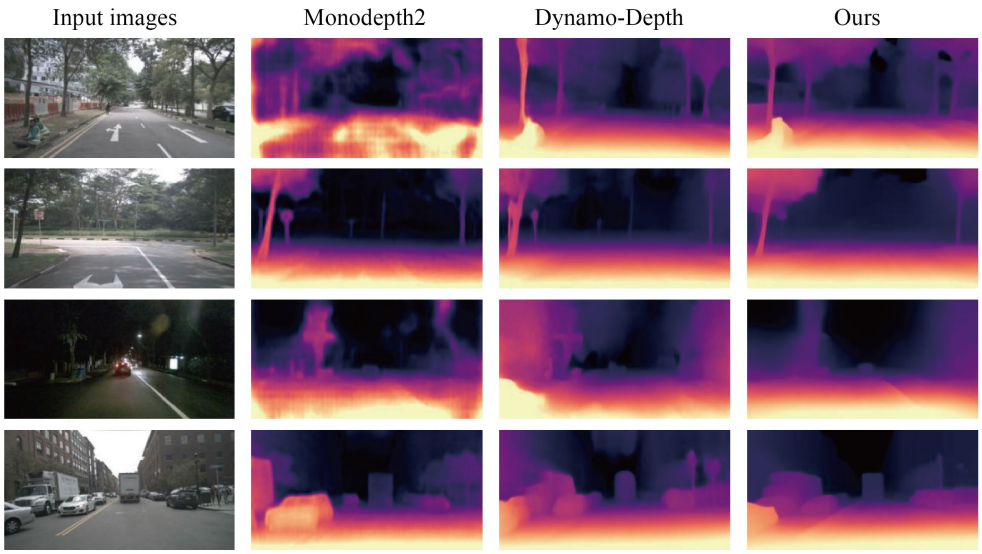

Figure 3. Comparison of visualization results on the nuScenes dataset
图 3. nuScenes 数据集上可视化比较结果

3.3. 对比与分析

为了验证模型的先进性，本文在两个公共数据集上训练和测试方法，并且在 DDAD 数据集上零样本验证模型的泛化性。

nuScenes： nuScenes 是一个具有挑战性的数据集，其中包含了大量的动态对象，并且还跨越了不同的天气和时间。表 1 和图 3 分别显示了模型在 nuSense 数据集上定量和定性结果。表中可以发现模型在

性能上远远优于其他现有的方法。并且因为上述的动态物体深度估计不准确的问题，Monodepth2 在性能上遭到灾难性的打击。他们只采用传统自监督模型来监督自己的网络。然而这种方法实际上并没有解决动态区域估计不准确的问题。相比之下，通过本文的方法网络性能在所有指标上都有明显的提升。

KITTI：本文也在 KITTI 这个大家广泛使用的公开数据集中进行了训练和测试。实验结果如表 2 和图 4 所示。实验表明本模型难以在 KITTI 数据集的动态区域受益，这是因为 KITTI 数据集中的大多数动态物体在前后两帧中没有移动。然而动态物体解耦会默认人，车这一类常见的动态物体时刻都在运动，当数据集中大量出现这一类物体时，会给模型带来一定的噪声进而影响模型的最终性能。虽然运动物体解耦模块会对静态区域造成一定的影响。但是对于模型几乎是可以接受的。

Table 2. Experimental results on the KITTI dataset. "OM" denotes whether moving objects are considered, and "-" indicates that the data is not provided in the paper
表 2. KITTI 数据集上的实验结果。"OM"表示是否考虑运动物体，"-"表示文章中没有提供该数据

方法	OM	Abs Rel	Sq Rel	RMSE	RMSElog	$\xi < 1.25\uparrow$	$\xi < 1.25^2\uparrow$	$\xi < 1.25^3\uparrow$
Monodepth2 [18]		0.115	0.903	4.863	0.190	0.877	0.959	0.981
Packnet [19]		0.111	0.785	4.601	0.189	0.878	0.960	0.982
Lite-mono [6]		0.103	0.798	4.514	0.179	0.897	0.964	0.983
Geonet [25]	√	0.155	1.296	5.857	0.233	0.793	0.931	0.973
TrianFlow [26]	√	0.113	0.704	4.581	0.184	0.871	0.961	0.984
Dynamic [27]	√	0.130	0.950	5.138	0.209	0.843	0.948	0.978
Insta-DM [28]	√	0.112	0.777	4.772	0.191	0.872	0.959	0.982
Dynamo (MD2) [8]	√	0.120	0.864	4.850	0.195	0.858	0.956	0.982
Dynamo (LM) [8]	√	0.112	0.785	4.505	0.166	0.873	0.959	0.984
Ours	√	0.109	0.795	4.605	0.186	0.883	0.960	0.983

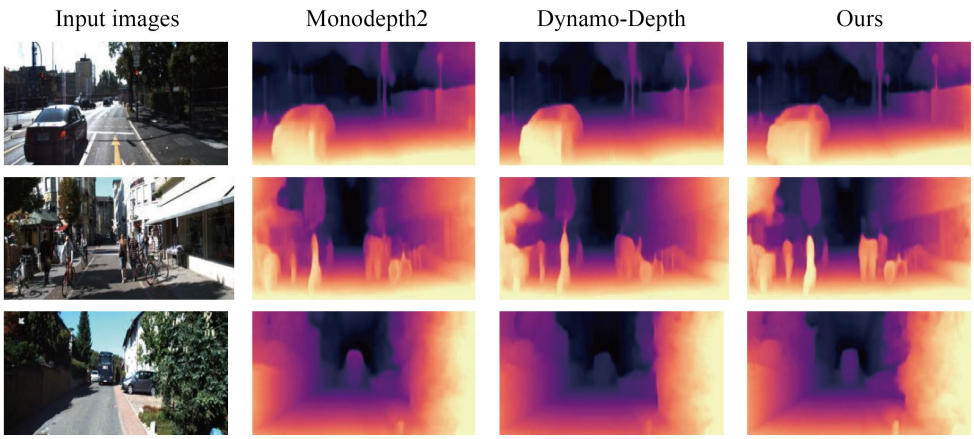

Figure 4. Comparison of visualization results on the KITTI dataset
图 4. KITTI 数据集上可视化比较结果

DDAD：为了验证模型的泛化能力，本文拿 nuScenes 数据集上训练的网络直接在 DDAD 数据集上测试。实验结果如表 3 和图 5 所示。尽管模型是零样本在 DDAD 数据集上测试，但模型性能上也优于大多数非零样本模型。这归功于本文在训练之前对训练数据的数据增强，使得模型在相似环境中仍然能够拥有强大的性能。

Table 3. Experimental results on the DDAD dataset. “OM” denotes whether moving objects are considered, and “-” indicates that the data is not provided in the paper
表 3. DDAD 数据集上的实验结果。“OM”表示是否考虑运动物体，“-”表示文章中没有提供该数据

方法	OM	Abs Rel	Sq Rel	RMSE	RMSElog	$\xi < 1.25\uparrow$	$\xi < 1.25^2\uparrow$	$\xi < 1.25^3\uparrow$
Monodepth2 [18]		0.239	12.547	18.392	0.316	0.752	0.899	0.949
Packnet [19]		0.182	7.945	15.021	0.259	0.828	0.925	0.961
zeroDepth [11]		0.156	-	10.678	-	0.814	-	-
Lite-Mono [6]		0.199	10.851	15.151	0.261	0.802	0.919	0.959
SurroundDepth [7]		0.200	3.392	12.270	-	0.740	0.894	0.947
ZoeDepth [22]		0.647	-	16.320	-	0.265	-	-
Ours	√	0.162	3.592	14.083	0.252	0.798	0.918	0.965

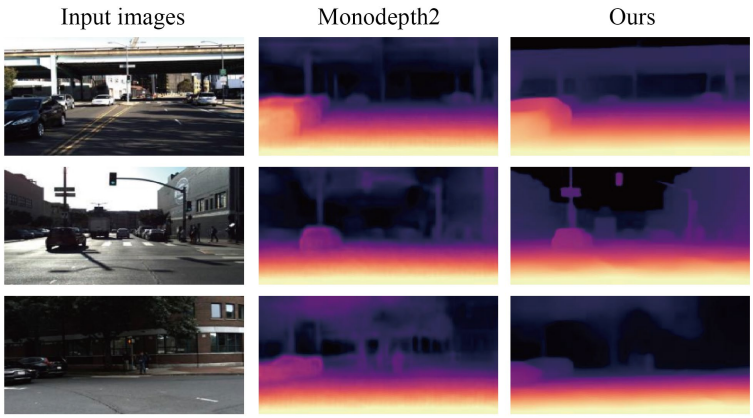

Figure 5. Comparison of visualization results on the DDAD dataset
图 5. DDAD 数据集上可视化比较结果

3.4. 消融实验

为了全面评估本文提出的模块的有效性，本文在具有挑战性的 nuScenes 数据集上对动态物体解耦(Dynamic Object Disentanglement, DOD)、动态物体重构损失(Dynamic Object Reconstruction Loss, DRL)以及深度一致损失(Depth Consistency Loss, DCL)三个关键部分进行了消融实验。实验的基线模型采用的是 Lite-Mono 模型，具体结果如表 4 所示。实验表明，当引入动态物体解耦模块时，深度估计性能已显著提升。进一步地，加入动态区域一致性损失和深度一致损失后，性能得到了进一步的改善。因此，可以得出结论，这三个模块对模型的性能提升均有显著贡献。

Table 4. Results of ablation experiment
表 4. 消融实验结果

方法	Abs Rel	Sq Rel	RMSE	RMSElog	$\xi < 1.25\uparrow$	$\xi < 1.25^2\uparrow$	$\xi < 1.25^3\uparrow$
Baseline	0.425	16.267	10.392	0.416	0.722	0.839	0.889
Baseline + DOD	0.202	2.944	7.549	0.274	0.769	0.911	0.957
Baseline + DOD + DRL	0.167	1.944	7.349	0.264	0.774	0.912	0.961
Baseline + DOD + DRL + DCL	0.152	1.627	7.083	0.252	0.792	0.916	0.965

4. 总结

为了解决传统自监督单目深度估计在动态场景中精度不足的问题，本文提出了一种新型的自监督单目深度估计模型。该模型通过引入动态物体掩码将动态物体从目标图像中分离出来，并引入光流网络估计目标图像到源图像的光流。利用动态物体、源图像及其光流信息重构新的图像，以提高动态物体深度估计的准确性。为了确保光流网络的估计精度，本文提出了动态物体重构损失。最后，通过共享权重的深度估计网络对重构源图像和原始目标图像进行深度估计，并利用这两个深度图进行相互监督，以保证深度信息的一致性，为确保深度能够有效地监督，提出了深度一致损失。本文在 nuScenes，KITTI，DDAD 数据集上的实验结果表明，模型能够有效解决动态场景深度估计不准确的问题，并且在 KITTI 这种静态场景较多的数据集也有一定的增益。尽管实验验证了所提方法的有效性和适用性，但仍存在一定的局限性。由于使用光流重构图像会导致重构源图像出现缺失，我们为解决这一问题在本文中采用的是使用目标图像前后帧共同监督的方法。我们认为这一方法并不是解决此问题的最优解，因此这也是我们未来研究的重点方向之一。

参考文献

[1] Borghi, G., Venturelli, M., Vezzani, R. and Cucchiara, R. (2017) POSEidon: Face-From-Depth for Driver Pose Estimation. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 5494-5503. https://doi.org/10.1109/cvpr.2017.583

[2] Biswas, J. and Veloso, M. (2012) Depth Camera Based Indoor Mobile Robot Localization and Navigation. 2012 *IEEE International Conference on Robotics and Automation*, Saint Paul, 14-18 May 2012, 1697-1702. https://doi.org/10.1109/icra.2012.6224766

[3] Swan, J.E., Jones, A., Kolstad, E., Livingston, M.A. and Smallman, H.S. (2007) Egocentric Depth Judgments in Optical, See-Through Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, **13**, 429-442. https://doi.org/10.1109/tvcg.2007.1035

[4] Godard, C., Aodha, O.M. and Brostow, G.J. (2017) Unsupervised Monocular Depth Estimation with Left-Right Consistency. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6602-6611. https://doi.org/10.1109/cvpr.2017.699

[5] Song, S. and Chandraker, M. (2014) Robust Scale Estimation in Real-Time Monocular SFM for Autonomous Driving. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 1566-1573. https://doi.org/10.1109/cvpr.2014.203

[6] Zhang, N., Nex, F., Vosselman, G. and Kerle, N. (2023) Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 18537-18546. https://doi.org/10.1109/cvpr52729.2023.01778

[7] Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Rao, Y., Huang, G., Lu, J. and Zhou, J. (2023) SurroundDepth: Entangling Surrounding Views for Self-Supervised Multi-Camera Depth Estimation. *Proceedings of the PMLR Conference on Robot Learning*, Atlanta, 6-9 November 2023, 539-549.

[8] Sun, Y. and Hariharan, B. (2024) Dynamo-Depth: Fixing Unsupervised Depth Estimation for Dynamical Scenes. *Advances in Neural Information Processing Systems*, **36**, 54987-55005.

[9] Geiger, A., Lenz, P., Stiller, C. and Urtasun, R. (2013) Vision Meets Robotics: The KITTI Dataset. *The International Journal of Robotics Research*, **32**, 1231-1237. https://doi.org/10.1177/0278364913491297

[10] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., et al. (2020) NuScenes: A Multimodal Dataset for Autonomous Driving. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 11618-11628. https://doi.org/10.1109/cvpr42600.2020.01164

[11] Guizilini, V., Vasiljevic, I., Chen, D., Ambruș, R. and Gaidon, A. (2023) Towards Zero-Shot Scale-Aware Monocular Depth Estimation. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 9199-9209. https://doi.org/10.1109/iccv51070.2023.00847

[12] Kumar, A.C.S., Bhandarkar, S.M. and Prasad, M. (2018) DepthNet: A Recurrent Neural Network Architecture for Monocular Depth Prediction. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Salt Lake City, 18-22 June 2018, 3960-3968. https://doi.org/10.1109/cvprw.2018.00066

[13] Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G. and Firman, M. (2021) The Temporal Opportunist: Self-

Supervised Multi-Frame Monocular Depth. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 1164-1174. https://doi.org/10.1109/cvpr46437.2021.00122

[14] Zhou, T., Brown, M., Snavely, N. and Lowe, D.G. (2017) Unsupervised Learning of Depth and Ego-Motion from Video. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6612-6619. https://doi.org/10.1109/cvpr.2017.700

[15] Godard, C., Aodha, O.M. and Brostow, G.J. (2017) Unsupervised Monocular Depth Estimation with Left-Right Consistency. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6602-6611. https://doi.org/10.1109/cvpr.2017.699

[16] Loshchilov, I. and Hutter, F. (2017) Decoupled Weight Decay Regularization. arXiv: 1711.05101.

[17] Teed, Z. and Deng, J. (2020) RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Computer Vision—ECCV* 2020, Springer, 402-419. https://doi.org/10.1007/978-3-030-58536-5_24

[18] Godard, C., Aodha, O.M., Firman, M. and Brostow, G. (2019) Digging into Self-Supervised Monocular Depth Estimation. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 3827-3837. https://doi.org/10.1109/iccv.2019.00393

[19] Guizilini, V., Ambrus, R., Pillai, S., Raventos, A. and Gaidon, A. (2020) 3D Packing for Self-Supervised Monocular Depth Estimation. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 2482-2491. https://doi.org/10.1109/cvpr42600.2020.00256

[20] Guizilini, V., Vasiljevic, I., Ambrus, R., Shakhnarovich, G. and Gaidon, A. (2022) Full Surround Monodepth from Multiple Cameras. *IEEE Robotics and Automation Letters*, **7**, 5397-5404. https://doi.org/10.1109/lra.2022.3150884

[21] Kim, J.H., Hur, J., Nguyen, T.P. and Jeong, S.G. (2022) Self-Supervised Surround-View Depth Estimation with Volumetric Feature Fusion. *Advances in Neural Information Processing Systems*, **35**, 4032-4045.

[22] Bhat, S.F., Birkl, R., Wofk, D., Wonka, P. and Müller, M. (2023) ZoeDepth: Zero-Shot Transfer by Combining Relative and Metric Depth. arXiv: 2302.12288.

[23] Vankadari, M., Golodetz, S., Garg, S., Shin, S., Markham, A. and Trigoni, N. (2023) When the Sun Goes Down: Repairing Photometric Losses for All-Day Depth Estimation. *Conference on Robot Learning*, Atlanta, 6 November 2023, 1992-2003.

[24] Marsal, R., Chabot, F., Loesch, A., Grolleau, W. and Sahbi, H. (2024) MonoProb: Self-Supervised Monocular Depth Estimation with Interpretable Uncertainty. 2024 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 3-9 January 2024, 3625-3634. https://doi.org/10.1109/wacv57701.2024.00360

[25] Yin, Z. and Shi, J. (2018) GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 1983-1992. https://doi.org/10.1109/cvpr.2018.00212

[26] Zhao, W., Liu, S., Shu, Y. and Liu, Y. (2020) Towards Better Generalization: Joint Depth-Pose Learning without PoseNet. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 9148-9158. https://doi.org/10.1109/cvpr42600.2020.00917

[27] Li, H., Gordon, A., Zhao, H., Casser, V. and Angelova, A. (2021) Unsupervised Monocular Depth Learning in Dynamic Scenes. *Proceedings of the PMLR Conference on Robot Learning*, London, 8-11 November 2021, 1908-1917.

[28] Saunders, K., Vogiatzis, G. and Manso, L.J. (2023) Dyna-DM: Dynamic Object-Aware Self-Supervised Monocular Depth Maps. 2023 *IEEE International Conference on Autonomous Robot Systems and Competitions* (*ICARSC*), Tomar, 26-27 April 2023, 10-16. https://doi.org/10.1109/icarsc58346.2023.10129564