arXiv:2304.03560v2 [cs.CV] 5 Apr 2024

DualRefine：通过迭代极线采样与朝向平衡态的细化实现自监督深度与位姿估计

AntyantaBangunharcana[1],Ahmed Magd[2],Kyung-Soo Kim[1]1机电一体化、系统与控制实验室， [2]车辆系统设计与控制实验室 韩国科学技术院（KAIST），大韩民国{antabangun, a.magd, kyungsoo}@kaist.ac.kr

摘要

自监督多帧深度估计通过计算相邻帧之间像素对应关系的匹配代价，将几何信息注入网络，从而实现高精度。这些像素对应候选项基于帧之间的相对位姿估计计算得出。准确的位姿预测对于精确计算匹配代价至关重要，因为它们会影响极线几何。此外，改进深度估计反过来还可以用于对齐位姿估计。

受传统运动恢复结构（*SfM*）原理的启发，我们提出了 *DualRefine* 模型，通过反馈循环紧密耦合深度估计与位姿估计。我们创新的更新流程采用深度平衡模型框架，通过基于极线几何计算局部匹配代价，迭代优化深度估计和特征图的隐藏状态。重要的是，我们在每一步都利用优化后的深度估计和特征图计算姿态更新。位姿估计的这些更新会在优化过程中逐渐改变极线几何。在 *KITTI*数据集上的实验结果表明，该模型在深度预测和里程计预测方面表现出竞争力，超越了已发表的自监督基线。[1]

1. 引言

对观测到的 3D 点坐标和相机姿态进行优化，构成了运动恢复结构（SfM）的基础。对两者的估计为机器人技术奠定了基础 [37, 38, 78], 自动驾驶 [22], 或 AR/VR 应用 [63]。然而，传统的 SfM 技术容易受到无纹理区域、动态物体等场景因素的影响，等。因此，人们开始开发能够从单目图像中学习预测深度的深度学习模型 [15, 16, 19, 51, 53]。这些模型可以仅凭图像线索即可准确预测深度，无需几何信息。

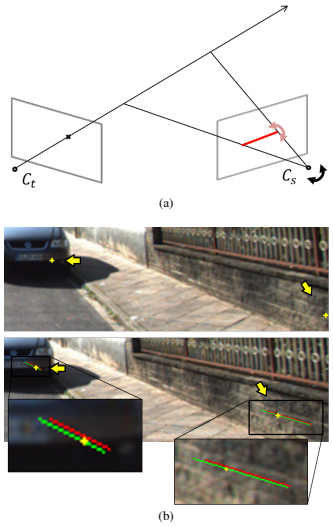


图1。 (a) 相机的估计位姿会影响极线几何。(b) 源图像中的极线，根据黄色目标图像中的点计算得出，用于基于 PoseNet的 [46]初始姿态回归（红色）和我们的优化后的位姿（绿色）。源图像中的黄色点是根据我们的最终深度和位姿估计计算得出的。

近年来，深度和位姿模型的自监督训练已成为一种颇具吸引力的方法，因为它无需真实标注，同时能够展现出与有监督模型相当的精度 [8, 21, 25, 26, 29, 31, 33, 64, 73, 77, 87, 91, 102, 110, 112]。这种方法利用深度和位姿预测在视频序列中合成相邻图像，并强制它们之间保持一致性。由于图像序列在测试时同样可用，近期的自监督方法也开始研究在推理过程中使用多帧 [95]。这些方法通常从多个视角构建代价体，以计算像素对应关系，与（多视图）立体匹配模型具有相似之处 [5, 47, 80]。通过整合多帧数据，几何信息得以融合，从而进行深度预测-

[1] *https://github.com/antabangun/DualRefine*

，从而提升性能和鲁棒性。在这种基于多帧匹配的模型中，匹配代价计算的精度至关重要。DepthFormer 的最新研究 [32]展示了其重要性，因为他们设计了一个基于 Transformer 的模块，以改进匹配代价 [88]并实现最先进的深度精度。然而，他们的方法带来了巨大的内存开销。

与立体视觉任务不同，前述自监督多帧模型不假设已知相机姿态，而是使用教师网络学习得到的估计值，通常是基于 PoseNet [46]的模型。该网络以两幅图像为输入，并回归得到六自由度姿态预测。由于估计的位姿会影响极线几何的计算（图 1(a)），因此位姿估计的准确性对于获得多帧之间准确的对应匹配至关重要。然而，正如近期研究 [75],所指出的，纯学习式姿态回归通常仍逊于传统方法，这是因为缺乏几何推理。通过优化位姿估计，我们可以提高匹配代价的准确性，并有望进一步获得更好的深度估计。如图 1(b)所示，根据回归位姿计算出的极线与优化后的估计并不对齐。反之，更好的深度预测也可能带来更好的姿态预测。因此，我们没有使用回归位姿一次性构建代价体，而是选择并行优化深度和位姿，并在每次迭代中采样更新后的局部代价体。这种方法深受传统运动恢复结构优化的启发，并与直接耦合深度和位姿预测的基于反馈的模型 [30]密切相关。

在这项工作中，我们提出了一种深度与姿态优化模型，使二者都趋向于平衡态，并在自监督框架下进行训练。我们通过以下贡献实现这一目标：**首先**，我们引入了一个基于极线几何和直接对齐的迭代更新模块。我们沿着根据当前位姿估计不断变化的极线采样候选匹配。随后，利用采样得到的匹配代价推断逐像素置信度，并据此计算深度优化结果。更新后的深度估计随后用于直接特征度量对齐，以优化姿态更新，使其趋于收敛。因此，我们的模型能够执行几何一致的深度与姿态更新。**其次**，这些更新会优化单帧模型给出的初始估计。这样一来，我们无需构建完整代价体，仅基于局部代价体进行更新，使方法更加简单、内存效率更高，也更稳健。**最后**，我们在深度平衡（DEQ）框架中设计了该方法 [4]，以隐式地使预测趋向于不动点。重要的是，DEQ 能够以较低的训练内存实现高效训练，改善了以往工作中内存消耗巨大的问题。通过这一新颖设计，我们通过实验展示了改进的深度估计，其性能可与最先进模型相媲美。此外，我们的模型展现出更强的视觉里程计结果全局一致性，优于其他基于学习的模型。

2. 相关工作

2.1. 单张图像的深度

单张图像的深度预测问题存在不适定性，因为不同的 3D 场景可能投影到同一张 2D 图像上 [35]。尽管如此，人类能够通过经验从单张图像中预测深度。受此启发，研究人员提出了大量监督式神经网络模型来解决单目深度估计任务，最早见于 Eigen 等人的论文 [16]。随后，研究人员通过修改模型架构 [15,51,67,98],、在大型数据集上训练 [14,56,99,104],、设计稳健的损失函数 [52,107], 以及将问题转化为分类任务 [55]，进一步提升了模型性能。然而，监督式深度估计需要真实深度图，而高质量、大规模地采集这类数据十分困难。这一挑战是研究人员探索半监督训练的主要原因之一。在半监督训练中，模型使用弱监督信号，例如提供相对深度 [13], 、相机姿态 [105], , 或利用合成数据进行训练 [2,50,65]。

对弱监督的需求在可泛化性和可扩展性等方面仍然存在局限。为应对这些限制，自监督训练技术的研究正逐渐受到重视。这些技术涉及在立体匹配 [21,25] 中使用几何信息，或使用一系列单目图像中的几何信息，最初由 [110]提出。Monodepth2 [26] 通过使用自动掩码和最小重投影损失来处理遮挡和自运动问题，进一步完善了利用图像序列进行训练的思想。后续工作将问题定义为分类任务 [28,45], , 修改了架构 [31], 针对低纹理区域的基于特征的损失 [77, 105], , 或减少了移动物体产生的伪影 [12,49,83]。

我们的工作基于自监督单目深度，将其用作教师和初始估计。

2.2. 多帧深度估计

在测试时依赖单帧图像，需要模型对场景的几何细节做出若干假设。相比之下，多帧方法利用可用的时序信息并结合多视图几何，从而减少了这类假设的必要性。

多帧深度预测与立体深度估计密切相关，其中神经网络将输入的立体图像转换为深度图，如 [57, 66, 86] 所示。Kendall 等。 [47] 通过构建平面扫描立体代价体取得了显著改进。一般而言，

多视图立体（MVS）研究与我们的工作更为相关，因为它利用了场景图像的非结构化集合，这意味着不同图像之间的位姿并不固定。在诸如 [39,41,61,94,101]，的研究中，通常会将前文提到的代价体与真实深度和相机姿态相结合，以提供指导。因此，这些工作在推理过程中需要相机姿态 [59,96]。尽管一些研究设法放宽了推理过程中的位姿要求，但训练仍然需要这种监督 [80,86]。

上述方法以及其他领域的进展，例如尺度域适应 [34,72,108],视图合成 [26,97], 和其他监督技术[29,33], , 显著提升了自监督方法的性能。ManyDepth [95], 与我们的工作相似，着眼于推理时使用单张图像的缺陷，并提出了一种灵活的模型，可在测试时（如果可用）利用多帧图像。ManyDepth 使用单帧模型作为教师，减少了移动物体和暂时静止的摄像机造成的伪影，从而提高了深度图的准确性。近期，DepthFormer [32] 通过使用 Transformer [88]获得改进的像素匹配代价，大幅提升了准确性。

受传统束调整的启发，我们设计了一种模型，在整合前文提到的诸多进展的同时，同时求解深度和位姿。我们的方法与DRO [30]最为相似。然而，与他们的方法相比，我们的模型将多视图几何紧密整合到迭代更新公式中，并基于局部极线几何进行细化。

2.3. 迭代细化

迭代细化已被用于提升各种学习任务的预测质量，包括目标检测 [7, 27], 光流估计 [40, 81], 语义分割 [71, 106], 等任务 [43,74]。近期一些研究尝试使用深度卷积网络迭代执行细化过程 [10,24,58,60]。另一些工作则利用前一次迭代的结果，反复训练同一个网络 [103,106]。具体而言，RAFT [81]通过其用于光流估计的迭代细化流程取得了成功。DEQ-flow [3]采用深度平衡模型（DEQ） [4] 框架，在保持准确率的同时降低了 RAFT 训练时的内存消耗。

我们模型的一个关键组件受 RAFT 和 DEQ-flow 迭代更新机制的启发。不同于光流，我们的模型并行优化深度和位姿估计。我们设计的优化模块将这两种预测紧密耦合起来，同时考虑相邻图像帧的极线几何。每次更新都会优化极线几何，从而使像素对应关系的匹配代价计算更加准确。

2.4. 位姿估计

位姿估计是自监督单目深度模型中的关键组成部分。在许多研究中，基于 PoseNet 的模型 [46] 会获取一对相邻图像帧，并输出六自由度位姿估计。这类模型较为直接，但由于缺乏几何约束，其准确性通常低于传统方法 [17,18,68] 。近期基于深度学习的定位研究通过采用直接对齐 [75, 89, 90] 或基于光流的几何对齐 [82]，在模型中引入了可微几何启发式设计。这些模型展现出更好的泛化能力和准确性。自监督单目深度也被用于改进传统里程计 [9,100]。然而，我们关注的是在自监督流程中学习同时优化深度和位姿。为提高位姿估计的准确性，我们在循环模块中集成直接对齐，确保深度与位姿之间的预测在几何上保持一致。

3. 方法

我们的模型由两个主要子模块组成。第一个是单帧自监督深度与位姿估计器，基于先前的框架构建 [26, 109], ，我们将在第 3.1 节重新讨论。该网络既是教师，也是第二个子模块（即我们提出的多帧网络）的初始化器；后者将在第3.2 节中详细介绍。

3.1. 自监督深度与位姿

我们首先介绍经典的自监督单目深度估计流程，[26], 该流程构成了我们方法的基础。此深度训练方法假设一台内参为K 的单目相机拍摄一个场景的图像序列。在此过程中，两个网络并行训练，以估计图像D 的逐像素深度图，以及相邻图像帧之间的相对位姿。利用这两项预测将相邻图像变形至同一目标帧，并通过约束帧间的光度一致性来进行自监督训练。

给定目标图像的深度图D 及其与源图像之间的相对位姿$T_{t\to s}$，我们可以如下计算目标图像中每个像素$u = (x,\ y)$在源图像上的投影：

$$z'u' = z'\begin{bmatrix}x'\\y'\\1\end{bmatrix} = KT_{t\to s}\left(D[u]K^{-1}\begin{bmatrix}x\\y\\1\end{bmatrix}\right), \tag{1}$$

使用该像素处估计的深度$D[u]$。随后，通过在计算得到的投影处对像素值进行采样，可以将源图像变形至目标帧

$$I_{s\to t}[u] = I_s\langle u'\rangle, \tag{2}$$

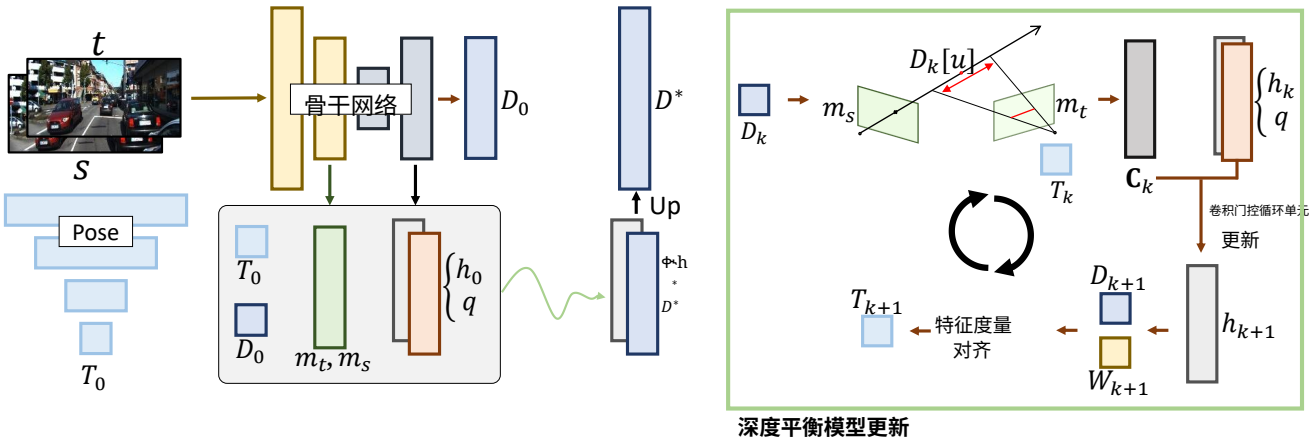


图2。（a）模型的整体流程。给定一对源图像和目标图像，教师模型会预测初始深度 D_0 和位姿 T_0,，以及将被更新的初始隐藏状态。随后执行基于深度平衡模型的对齐，以求得不动点并输出最终预测。（b）更新步骤中的每次迭代都会使用当前的深度和位姿估计。匹配代价沿着根据位姿估计不断变化的当前极线进行采样。更新由卷积门控循环单元计算。随后使用特征度量对齐，获得几何上一致的位姿更新。

其中 $\langle\cdot\rangle$表示双线性插值，该插值通过空间变换网络（STN）实现 [42]。

自监督损失由光度误差和边缘感知平滑损失组合计算:

$$\mathcal{L}_{self-sup} = \lambda_p \mathcal{L}_p + \lambda_s \mathcal{L}_s, \tag{3}$$

变形后的源图像与目标图像之间的光度误差使用结构相似性损失计算，并在多个变形后的图像之间取最小误差，以处理遮挡。感兴趣的读者可参考 [26]。自监督损失通常在多个尺度上计算，以稳定训练。

在本文中，我们使用这一流程训练单目深度估计器和位姿估计网络，目的是实现两个目标。首先，遵循 ManyDepth [95], ，我们将这些模型用作教师，在存在动态物体时约束多帧预测。此外，我们还将它们用作多帧对齐网络的初始化器。

3.1.1 单目模型

我们基于 DIFFNet [109], 这一最先进的自监督单帧深度估计器构建单目深度估计模型。我们使用 HRNet 架构 [92]从目标图像中提取多个尺度 s 的特征图。与 DIFFNet 一致，我们将多个阶段的特征图累积到$F^{(1/2^s)}$中。随后，我们使用视差解码器在尺度 $s = \{2,3\}$上生成视差预测。位姿估计网络遵循经典的 PoseNet [46] 架构，接收两幅输入图像并输出六自由度值，采用 ResNet18 [36] 骨干网络。这些网络预测的视差和位姿估计被用作教师，以训练我们的对齐子模块，这与 ManyDepth [95]中的做法一致。

3.2. 深度平衡对齐

在我们的对齐子模块中，我们假设可以获得来自源图像的额外输入，并利用这些输入细化深度和位姿估计。在本文中，我们重点使用图像序列中前一帧的图像作为源图像。

我们的对齐模块被建模为一个深度平衡模型 [4] ，它会更新隐藏状态、深度和位姿估计，使其收敛至一个不动点。具体而言，在不动点处 z^*,

$$(h^*, D^*, T^*) = z^* = \mathrm{U}(z^*, x), \tag{4}$$

其中 z 由隐藏状态 h、深度预测D和位姿预测 T组成。U 表示我们的更新函数，它交替细化深度和位姿。x 是更新模块的输入，根据每一步的极线几何获得，我们将在下一小节中讨论这一点。我们以尺度 $s = 2$执行这些迭代更新。

从特征提取输出中，我们为循环更新计算初始隐藏状态 $h^{[0]}$ =tanh(H($F^{(1/4)}$)) 和上下文特征 q = Q($F^{(1/4)}$)。H 和 Q 均由一个残差块 [36] 后接一个卷积层组成。

3.2.1 局部邻域周围的深度更新

局部极线采样。如前所述，我们的细化基于目标图像与源图像之间的像素匹配。我们使用 HRNet 特征提取器的前两个模块，从源图像 $m_s^{(1/4)}$和目标图像 $m^{(1/4)}{}_t$ 中提取单元特征，并用它们计算匹配代价。

与 RAFT [81], 类似，我们的目标是计算当前候选对应点周围的匹配值。

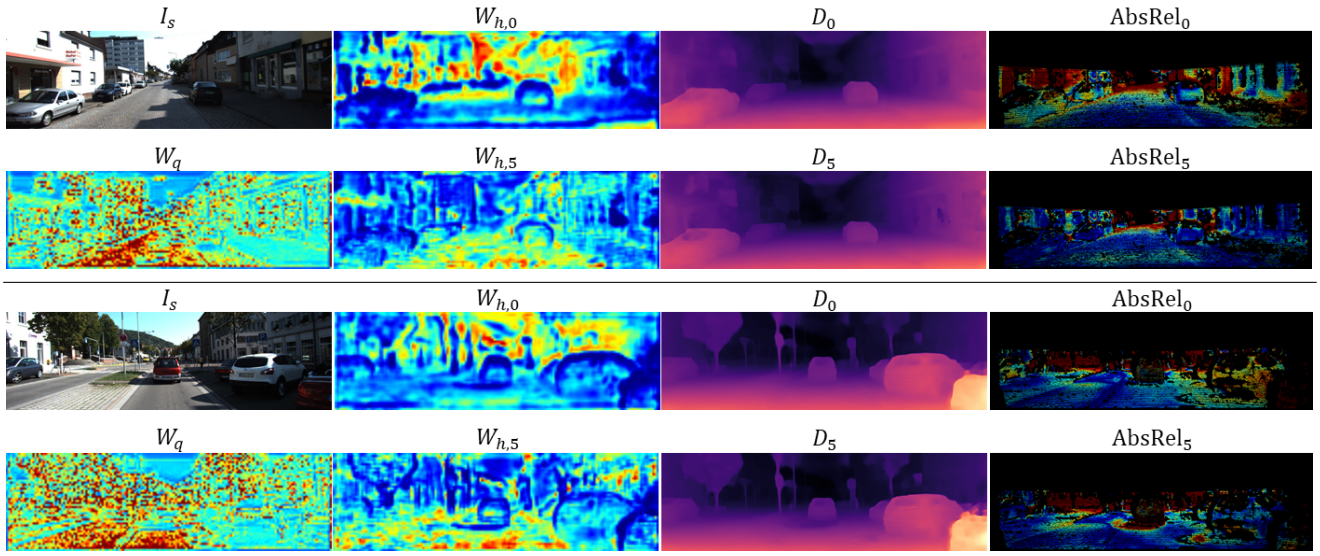


图3。KITTI 数据上的定性结果。I_S：输入图像；W_q、W_h、0 和W_h、5：置信度权重；D_0, D_5：视差估计；深度估计的绝对相对误差。

预测结果，并将它们作为更新步骤中的输入传递给我们的更新模块 k。然而，与 RAFT 不同，我们根据位姿估计 T_k沿极线进行匹配。具体做法是计算在感兴趣深度处从（1）得到的源图像中的投影坐标 u'_k 。深度候选值取为当前预测值的邻域。我们不采用 RAFT 中计算所有成对相关性的做法，而是即时计算特征匹配。

对于每个像素u，我们将局部深度候选计算为$D[u]_k \pm (i \times c \times n)$，其中 $i \in \mathbb{Z}$和$i \leq r$。在我们的实验中，我们将采样半径超参数设置为$r = 8$ 。在深度估计中，误差通常会随距离增大。为对此进行考虑，我们将$c = D[u]/C$ 定义为深度的函数，使采样范围取决于深度，其中我们将C 设为可训练参数。遵循 RAFT，为了收集更大邻域中的匹配信息，我们在多个层级$n = \{1,2,3\}$进行采样。在每个层级，我们以 $1/2^n$ 的比例对源图像$m_s^{(1/4)}$的匹配特征图进行双线性缩放。然后，在计算得到的对应坐标集合u'_k处对匹配特征进行采样，并计算其与目标特征的绝对差异

$$\mathbf{C}_k[u] = |m_t[u] - m_s\langle u'_k\rangle| \tag{5}$$

进行计算和汇总。这为我们提供了一个映射 $\mathbf{C}_k$，其中包含 $n \times (2 \times r + 1)$ 个对应深度候选值的匹配代价值。随后，我们将这些匹配代价与深度一同输入一个两层卷积神经网络（CNN）进行编码，以计算更新模块的输入

$$x_k = [\mathrm{CNN}_{\mathbf{C}}(\mathbf{C}_k), \mathrm{CNN}_{D_x}(D_k), q], \tag{6}$$

其中 $[\cdot,\cdot]$ 表示拼接。

深度更新。更新函数计算更新后的隐藏状态 h^{k+1}，使用卷积门控循环单元模块 [6,44,81]。 h^{k+1}用于计算深度更新。为稳定训练，我们使用激活函数 $\tanh(\cdot)$将深度更新值的绝对值限制在 $r \cdot c$：

$$D_{k+1} = D_k + r \cdot c \cdot \tanh(\mathrm{CNN}_{D_\mathrm{U}}(h_{k+1})). \tag{7}$$

这些更新与姿态更新交替进行，以达到不动点 D^*。使用h^*，我们计算凸上采样，以获得输入分辨率下的最终深度估计。

3.2.2 特征度量位姿对齐

在（1）中，位姿估计的准确性会影响像素坐标的计算。因此，优化后的位姿估计也会提高匹配代价的可靠性。为了在保持几何一致性的同时优化位姿，我们基于直接特征对齐执行姿态更新 [17, 18, 91]。这些更新 δ_k 可以通过求解$H_k\delta_k = b_k$，其中

$$\begin{aligned} H_k = \mathcal{J}_k^T\, diag(W_k)\mathcal{J}_k \text{ and } b_k = -\mathcal{J}_k^T\, diag(W_k) r_k \\ r_k[u] = m_s\langle u'_k\rangle - m_t[u], \end{aligned} \tag{8}$$

且 $\mathcal{J}$是关于位姿的雅可比矩阵。

为了计算这次位姿更新，可以为图像中的每个像素分配相同的置信度。然而，在公式（8）中，也可以加入每个像素的额外置信度权重W 。这样做有两个原因。首先，动态物体可能会影响位姿更新的解

，重复纹理区域中也可能出现不准确的特征对齐。为此，可以计算输入上下文特征图的置信度权重W_q。[75]具体而言，我们为目标图像和源图像计算了置信度图$W_{t,\ s}=1/(1+\mathrm{ReLU}(\mathrm{CNN}^{W_q}(F_{t,s})))$。利用公式（1）和（2），将源图像的置信度变换至目标帧。最后，置信度计算为$W_q=W_t\cdot W_{s,\ warped}$。该置信度权重只计算一次，用于为输入图像中的每个像素分配置信度。

其次，姿态更新的精度取决于深度估计的精度。为了获得更准确的姿态更新，我们希望为深度精度较高的区域赋予更高的对齐置信度。我们使用神经网络根据匹配代价推断这种置信度。由于隐藏状态h包含了这些匹配代价的历史信息，因此可以计算另一组置信度权重$W_{h,\ k}=\mathrm{CNN}^{W_h}(h_k)$。与之前的置信度图不同，这组权重会在每个更新步骤中计算并演化。这样，网络便可利用深度预测提供的信息，引导位姿估计逐步收敛。在实验中，我们研究了分别使用每种置信度权重以及将两者结合使用的效果$W_k=W_qW_{h,\ k}$。

最后，我们可以将更新后的位姿计算为$T_{k+1}=exp(\delta_k^\wedge)T_k$。这些操作均设计为可微分，以支持端到端训练。

3.2.3 深度平衡模型训练

我们采用深度平衡模型框架，重复上述步骤，直到深度值和位姿值达到不动点，此时更新值最小。在我们的实现中，深度和位姿的不动点分别选取，因此二者可能来自不同的更新步骤。最后，使用选定的深度和位姿不动点计算训练梯度。在达到不动点之前的操作无需在内存中保存梯度，从而实现高效的内存训练。

如式（1）至（3）所述，给定深度和位姿预测，通过变形输入图像即可计算自监督损失。在不动点处，我们针对每个细化后的深度D^*和位姿T^*估计额外计算两项自监督损失。与先前工作不同，先前工作将多帧深度估计D^*与教师位姿估计T_0配对，以进行变形并计算损失；而我们将D^*与T^*配对。关于这一损失配对选择的详细实验见补充材料。

在这两种损失中，我们还对D_0和D^*之间施加一致性损失，与 ManyDepth [95], 类似，以处理动态物体或遮挡区域。具体而言，我们从原始特征匹配中提取粗略的深度预测，并屏蔽差异较大的区域，同时强制其与教师深度保持一致。与 ManyDepth 不同，不过，我们的方法不会显式构建代价体。为了获得粗略深度，我们在教师深度附近的邻域内搜索匹配代价最低的位置，这与第 3.2.1 节类似，但邻域范围更大。与基于代价体的方法相比，这种方法还有一项优势：我们无需依赖估计的最小和最大深度，也无需知道估计值的尺度。此外，由于粗略深度的计算取决于匹配代价的准确性，因此可通过更准确的位姿估计加以改进（表 2）。

4. 实验

4.1. 数据集和指标

在深度估计实验中，我们使用 Eigen 训练/测试划分 [16]，数据来自 KITTI 数据集 [22]。为了评估估计的深度，我们通过一个标量对其进行缩放，使其与真实值的尺度相匹配。我们采用标准的深度评估指标 [15, 16], ，包括绝对相对误差和平方相对误差（Abs Rel、Sq Rel）、均方根误差（RMSE、RMSE log）以及阈值准确率（δ_1，δ_2和δ_3），最大深度设为 80米。前四项指标的值越低越好，其余三项则越高越好。

在视觉里程计实验中，我们使用 KITTI 里程计数据集，并沿用先前工作 [110] 中相同的训练和评估序列（序列 00–08 用于训练，序列 09–10 用于评估）。由于我们的估计依赖于单目相机，因此使用 7 DoF Umeyama 对齐 [85]将估计轨迹与真实轨迹对齐。我们采用标准的里程计评估指标，例如平移（t_{err}）和旋转（r_{err}）误差 [22],以及绝对轨迹误差（ATE） [79]。

4.2. 实现细节

我们使用 PyTorch [69] 在配备 RTX3090 GPU 的环境中进行实验，批量大小为12。按照 [26],, 我们应用颜色和翻转增强，并将输入图像调整为 640×192 的分辨率。对于高分辨率实验，我们将图像调整为 960×288。我们以 10^{-3}的学习率训练整个网络15个 epoch，之后冻结教师深度和位姿模型。随后，我们以10^{-4}的学习率继续训练网络。采用Adam 优化器 [48] ，其参数设置为 $\beta_1=0.9$ 和$\beta_2=0.999$。

如前文所述，深度骨干网络基于 HRNet 架构 [92]。对于教师位姿模型，我们遵循标准设计，使用以 ImageNet 预训练权重初始化的 ResNet18 前五层作为编码器，后接一个输出 6-DoF 位姿估计的解码器。在测试时，使用当前帧以及可用时的前一帧进行深度估计。当

	方法	测试帧	语义	宽 ×高	绝对相对误差↓	平方相对误差↓	均方根误差↓	对数均方根误差↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Low & mid res	Ranjan等. [73]	1		832 × 256	0.148	1.149	5.464	0.226	0.815	0.935	0.973
	EPC++ [62]	1		832 × 256	0.141	1.029	5.350	0.216	0.816	0.941	0.976
	Struct2depth（M） [11]	1	•	416 × 128	0.141	1.026	5.291	0.215	0.816	0.945	0.979
	真实场景视频 [29]	1	•	416 × 128	0.128	0.959	5.230	0.212	0.845	0.947	0.976
	Guizilini等。 [33]	1	•	640 × 192	0.102	0.698	4.381	0.178	0.896	0.964	**0.984**
	Johnston 等。 [45]	1		640 × 192	0.106	0.861	4.699	0.185	0.889	0.962	0.982
	Monodepth2 [26]	1		640 × 192	0.115	0.903	4.863	0.193	0.877	0.959	0.981
	Packnet-SFM [31]	1		640 × 192	0.111	0.785	4.601	0.189	0.878	0.960	0.982
	Li 等。 [54]	1		416 × 128	0.130	0.950	5.138	0.209	0.843	0.948	0.978
	DIFFNet [109]	1		640 × 192	0.102	0.764	4.483	0.180	0.896	0.965	0.983
	DualRefine-initial（D_0）	1		640 × 192	0.103	0.776	4.491	0.181	0.894	0.965	0.983
	Patil等. [70]	N†		640 × 192	0.111	0.821	4.650	0.187	0.883	0.961	0.982
	Wang等. [93]	2（-1，0）		640 × 192	0.106	0.799	4.662	0.187	0.889	0.961	0.982
	ManyDepth（MR） [95]	2（-1，0）		640 × 192	0.098	0.770	4.459	0.176	0.900	0.965	0.983
	DepthFormer [32]	2 (-1, 0)		640 × 192	0.090	**0.661**	**4.149**	0.175	0.905	**0.967**	**0.984**
	DualRefine 精炼版（D^*）	2 (-1, 0)		640 × 192	**0.087**	0.698	4.234	**0.170**	**0.914**	**0.967**	0.983
High res	DRO [30]	2 (-1, 0)		960 × 320	0.088	0.797	4.464	0.212	0.899	0.959	0.980
	Wang等. [93]	2 (-1, 0)		1024 × 320	0.106	0.773	4.491	0.185	0.890	0.962	0.982
	ManyDepth（高分辨率 ResNet50） [95]	2 (-1, 0)		1024 × 320	0.091	0.694	4.245	0.171	0.911	0.968	0.983
	DualRefine 精炼版（HR）（D^*）	2 (-1, 0)		960 × 288	**0.087**	**0.674**	**4.130**	**0.167**	**0.915**	**0.969**	**0.984**

表 1. KITTI [22] Eigen划分上的结果及与其他当前最优模型的比较。**粗体**：最佳；下划线：次佳。†：在完整序列上评估

Pose 更新	一致性 mask	绝对相对误差	平方相对误差	RMSE	δ_1	δ_2
不更新	T_0	0.097	0.713	4.462	0.898	0.964
无权重	T_0	0.091	0.694	4.271	0.909	**0.967**
不 $W_{h_,k}$	T_0	0.090	0.667	4.252	0.909	**0.967**
无 W_q	T_0	0.093	0.686	4.258	0.908	**0.967**
W_q and $W_{h,k}$	T_0	0.090	0.669	4.293	0.910	**0.967**
无权重	T^*	0.092	0.667	4.257	0.908	**0.967**
无 $W_{h_,k}$	T^*	0.091	**0.666**	4.243	0.909	**0.967**
否 W_q	T^*	0.088	0.674	4.251	0.911	0.966
W_q 和 $W_{h,_,k}$	T^*	**0.087**	0.698	**4.234**	**0.914**	**0.967**

表 2。姿态更新对 KITTI [22]Eigen划分影响的消融实验。**粗体**：最佳。

前一帧不可用时，我们会跳过细化模块，直接使用初始估计。

4.3. 消融实验

姿态更新。我们分析了针对深度的姿态优化效果，并在表2中展示了研究结果。不执行姿态更新的模型准确率最差。

不断演化的置信度加权。 我们还展示了置信度加权的影响。有趣的是，对于所有采用位姿更新的模型，即使不使用任何加权来引导位姿计算，也可以观察到相似的性能。

一致性掩码。 这里我们使用优化后的位姿提取一致性掩码，以用于训练。整体结果略有提升，这表明改进后的位姿估计有助于计算更准确的匹配代价。

深度平衡模型迭代。 我们研究了深度平衡模型迭代更新的相关方面，并在表 3 中展示了训练和测试时改变迭代次数所得的结果。结果表明，6 次迭代足以找到不动点。我们推测，初始估计提供了可靠的起点，因此能够快速收敛。尽管如此，仍需进一步研究模型训练的稳定性

DEQ # 次迭代	绝对相对误差	平方相对误差	RMSE	δ_1	Time (ms)
3→3	0.094	0.725	4.355	0.906	53
6→3	0.097	0.711	4.312	0.908	53
6→6	0.087	0.698	4.234	0.914	68
12→3	0.098	0.73	4.370	0.900	53
12→6	0.093	0.695	4.310	0.906	68
12→12	0.089	0.692	4.242	0.910	99

表 3. KITTI [22] Eigen划分上的深度平衡模型迭代消融实验。 $a \rightarrow b$ 表示 a：训练时的迭代次数，以及 b：测试时的迭代次数。

使用 DEQ 技术进行更多次迭代 [3]是合理的。在我们的机器上，6 次更新迭代使基线模型的推理时间从 37.90 毫秒（其中也包括初始位姿回归）增加了 ∼31 毫秒，总计达到 68.49 毫秒，运行速度接近 15 fps。不过，我们的实现仅使用 PyTorch 基础函数，因此还可以进一步优化代码。

4.4. 深度结果

表1 展示了我们的方法与最先进的自监督模型在深度估计方面的比较。我们将模型与在单目视频上训练的模型进行比较。我们的方法优于大多数以往模型，并且与基于 Transformer 的 [88]DepthFormer [32] 模型相比也具有竞争力。具体而言，我们的方法在 δ_1方面取得了显著提升，表明其内点精度很高。此外，与每批训练需要 16 GB 训练内存的 DepthFormer 相比，我们的方法处理 12 批数据仅需 24 GB 内存，约为1/8×。这主要是因为我们的方法基于局部采样来优化二维隐藏状态，而 [32]则使用自注意力和交叉注意力沿每个深度区间优化二维特征图和三维特征体。

图3 展示了我们模型的视差和误差图定性结果。我们可以观察到改进

方法	序列 9			序列 10		
	t_{err}(%) ↓	r_{err}(°/100m) ↓	ATE (m) ↓	t_{err}(%) ↓	r_{err}(°/100m) ↓	ATE（米） ↓
ORB-SLAM2 [68](无回环)	9.67	0.3	44.10	4.04	0.3	6.43
ORB-SLAM2 [68]	3.22	0.4	8.84	4.25	0.3	8.51
SfMLearner [110]	19.15	6.82	77.79	40.40	17.69	67.34
GeoNet [102]	28.72	9.8	158.45	23.90	9.0	43.04
DeepMatchVO [76]	9.91	3.8	27.08	12.18	5.9	24.44
单目深度2 [26]	17.17	3.85	76.22	11.68	5.31	20.35
DW [29]-学习得到	-	-	20.91	-	-	17.88
DW [29]-校正后	-	-	19.01	-	-	14.85
SC-Depth [8]	7.31	3.05	23.56	7.79	4.90	12.00
Zou等。 [111]	3.49	**1.00**	11.30	**5.81**	1.8	11.80
P-RGBD SLAM [9]	5.08	1.05	13.40	4.32	2.34	**7.99**
DualRefine-initial (T_0)	9.06	2.59	39.31	9.45	4.05	15.13
DualRefine 精炼版（T^*）	**3.43**	1.04	**5.18**	6.80	**1.13**	10.85

表4. 测量09和10序列。ORB-SLAM2 已包含 KITTI里程计数据中的数据片段 . 我们将其与其他状态进行比较 -of-the- 艺术自身 -监督深度与里程计作为一种具有代表性的非学习方法。**粗体**：最佳，下划线：一次佳。

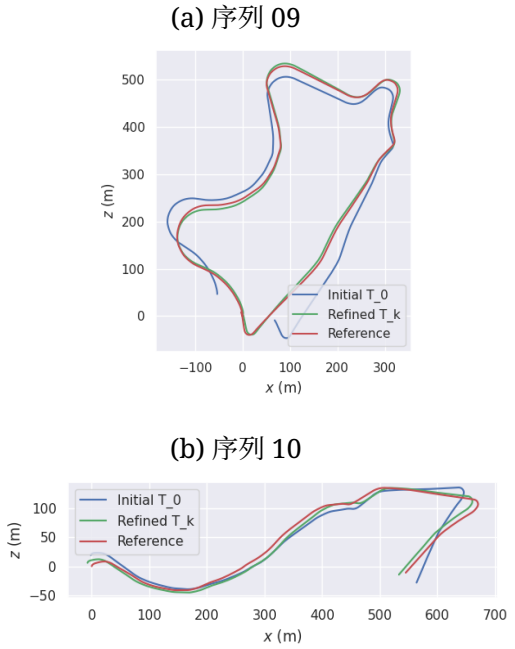


图4.在KITTI里程计数据的（a）序列09和（b）序列10上，由初始位姿估计器估计的轨迹，以及使用我们的位姿优化模块得到的优化后的轨迹。即使没有针对全局一致性进行显式训练，优化后的位姿估计也能改善全局轨迹。

优化后深度的误差图。我们还额外展示了模型计算得到的置信度权重输出。我们观察到，只计算一次的置信度w_q会稀疏地赋予高置信度。有趣的是，随着每次迭代而变化的置信度权重最初会对远处的点赋予高置信度，并随着迭代次数增加而逐渐转向较近的点。

局限性： 我们注意到，在较低的图像集中的移动车辆上，误差有所增加。与DepthFormer 相比，我们模型较高的均方根误差也表明其离群预测更多。这可能是由于反复执行我们的迭代更新可能会进一步加剧离群- 值问题。即使采用一致性掩码，我们提出的模型在处理动态物体方面仍存在局限性。我们将这一问题的进一步讨论留待未来研究。

4.5. 里程计结果

我们在表 4 中展示了教师模型和细化模型的视觉里程计结果。我们也展示了此前在单目视频上训练的模型的结果。如图 4 所示，我们的细化模块大幅改善了初始里程计结果。尽管本研究旨在提升局部位姿的估计精度，以实现准确匹配，但在所有指标上，我们的表现都优于大多数其他模型。即使没有像 [8, 9]，那样进行明确训练以确保尺度一致性，我们的细化输出仍能展现出全局一致的里程计预测。此外，不同于 Zou等人 [111] 通过长期几何信息推断位姿，我们仅使用两个输入帧来推断位姿估计。我们的模型还取得了与传统 ORB-SLAM2 相当的 ATE，尽管传统 ORB-SLAM2 会执行全局几何优化，而我们的结果在r_{err} 中仍稍逊一筹。

5. 结论

在本文中，我们提出了一种用于多帧深度和位姿估计与细化的自监督流程。通过结合神经网络表示和几何约束的能力来细化深度和位姿，我们的方法在这两项任务中都取得了最先进的性能。与竞争方法相比，我们的方法还展现出更高的效率，并具有进一步改进的潜力。然而，我们仍然观察到，在动态场景中深度精度较差。

参考文献

[1] Donald G Anderson。非线性积分方程的迭代过程。ACM期刊（JACM），12(4):547–560，1965。14[2] Amir Atapour-Abarghouei 和 Toby P Breckon。使用合成数据并通过图像风格迁移进行域适应的实时单目深度估计。载于 IEEE计算机视觉与模式识别会议论文集，第2800–2810页，2018年。2[3] Shaojie Bai、Zhengyang Geng、Yash Savani 和 J Zico Kolter。深度平衡光流估计。载于IEEE/CVF计算机视觉与模式识别会议论文集，第620–630页，2022年。3, 7, 14[4] Shaojie Bai、J Zico Kolter 和 Vladlen Koltun。深度平衡态模型。神经信息处理系统进展，32，2019年。2, 3, 4, 14[5] Antyanta Bangunharcana、Jae Won Cho、Seokju Lee、In So Kweon、Kyung-Soo Kim 和 Soohyun Kim。相关与激励：通过引导式代价体激励实现实时立体匹配。载于2021 IEEE/RSJ智能机器人与系统国际会议（IROS），第3542–3548页。IEEE，2021年。1[6] Antyanta Bangunharcana、Soohyun Kim 和 Kyung-Soo Kim。重新审视DROID-SLAM中卷积门控循环单元的感受野。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第1906–1916页，2022年。5[7] Sayanti Bardhan。通过上下文细化进行显著目标检测。载于 IEEE/CVF计算机视觉与模式识别会议研讨会论文集，第356–357页，2020年。3[8] Jiawang Bian、Zhichao Li、Naiyan Wang、Huangying Zhan、Chunhua Shen、Ming-Ming Cheng 和 Ian Reid。从单目视频中进行无监督尺度一致深度与自运动学习。神经信息处理系统进展，32:35–45，2019年。1, 8[9] Jia-Wang Bian、Huangying Zhan、Naiyan Wang、Zhichao Li、Le Zhang、Chunhua Shen、Ming-Ming Cheng 和 Ian Reid。从视频中进行无监督尺度一致深度学习。《国际计算机视觉期刊》，129 2564，2021 9 2548 3, 8 ( ): – 。[10] Arantxa Casanova、Guillem Cucurull、Michal Drozdzal、Adriana Romero 和 Yoshua Bengio。关于用于语义分割的密集连接表示层级的迭代细化。载于 IEEE计算机视觉与模式识别会议研讨会论文集，第978–987页，2018年。3[11] Vincent Casser、Soeren Pirk、Reza Mahjourian 和 Anelia Angelova。无需传感器的深度预测：利用结构从单目视频中进行无监督学习。载于 AAAI人工智能会议论文集，第33卷，第8001–8008页，2019年。7[12] Vincent Casser、Soeren Pirk、Reza Mahjourian 和 Anelia Angelova。利用结构与语义进行无监督单目深度和自运动学习。载于 IEEE/CVF计算机视觉与模式识别会议研讨会论文集，第0–0页，2019年。2[13] Weifeng Chen、Zhao Fu、Dawei Yang 和 Jia Deng。野外单图像深度感知。《神经信息处理系统进展》，第 29 卷，2016 年。2[14] Weifeng Chen、Shengyi Qian、David Fan、Noriyuki Kojima、Max Hamilton 和 Jia Deng。Oasis：用于野外单图像 3D 的大型数据集。发表于《IEEE/CVF计算机视觉与模式识别会议》（CVPR）论文集，2020 年 6 月。2[15] David Eigen 和 Rob Fergus。利用通用多尺度卷积架构预测深度、表面法线和语义标签。发表于《IEEE国际计算机视觉会议》论文集，第 2650–2658 页，2015 年。1、2、6 – 。[16] David Eigen、Christian Puhrsch 和 Rob Fergus。利用多尺度深度网络从单幅图像预测深度图。《神经信息处理系统进展》，第 27 卷，2014 年。1、2、6、14、15[17] Jakob Engel、Vladlen Koltun 和 Daniel Cremers。直接稀疏里程计。《IEEE模式分析与机器智能汇刊》，40(3):611–625，2017 年。3、5[18] Jakob Engel、Thomas Sch¨ops 和 Daniel Cremers。LSD-SLAM：大规模直接单目 SLAM。发表于《欧洲计算机视觉会议》论文集，第 834–849 页。Springer，2014 年。3、5[19] Huan Fu、Mingming Gong、Chaohui Wang、Kayhan Batmanghelich 和 Dacheng Tao。用于单目深度估计的深度序数回归网络。发表于《IEEE计算机视觉与模式识别会议》论文集，第 2002–2011 页，2018 年。1[20] Samy Wu Fung、Howard Heaton、Qiuwei Li、Daniel McKenzie、Stanley J Osher 和 Wotao Yin。不动点网络：具有无雅可比矩阵反向传播的隐式深度模型。2021 年。14[21] Ravi Garg、Vijay Kumar Bg、Gustavo Carneiro 和 Ian Reid。用于单视图深度估计的无监督 CNN：几何来助力。发表于《欧洲计算机视觉会议》论文集，第 740–756 页。Springer，2016 年。1、2、14[22] Andreas Geiger、Philip Lenz 和 Raquel Urtasun。我们准备好自动驾驶了吗？KITTI 视觉基准套件。发表于 2012 年 IEEE计算机视觉与模式识别会议论文集，第 3354–3361 页。IEEE，2012 年。1、6、7、14、15、17[23] Zhengyang Geng、Meng-Hao Guo、Hongxu Chen、Xia Li、Ke Wei 和 Zhouchen Lin。注意力机制是否优于矩阵分解？arXiv 预印本 arXiv:2109.04553，2021 年。14[24] Golnaz Ghiasi 和 Charless C Fowlkes。用于语义分割的拉普拉斯金字塔重建与细化。发表于《欧洲计算机视觉会议》论文集，第 519–534 页。Springer，2016 年。3[25] Cl´ement Godard、Oisin Mac Aodha 和 Gabriel J Brostow。利用左右一致性进行无监督单目深度估计。发表于《IEEE计算机视觉与模式识别会议》论文集，第 270–279 页，2017 年。1、2 – 。[26] Cl´ement Godard、Oisin Mac Aodha、Michael Firman 和 Gabriel J Brostow。深入探究自监督单目深度估计。发表于《IEEE/CVF国

际计算机视觉会议》论文集，第 3828–3838 页，2019 年。1、2、3、4、6、7、8、15 – 。[27] Jicheng Gong、Zhao Zhao 和 Nic Li。通过迭代提案细化改进多阶段目标检测。发表于 BMVC，第 223 页，2019 年。3[28] Juan Luis GonzalezBello 和 Munchurl Kim。别再惦记激光雷达：基于中概率体的自监督深度估计器。《神经信息处理系统进展》，第 33 卷：12626–12637，2020 年。2[29] Ariel Gordon、Hanhan Li、Rico Jonschkowski 和 Anelia Angelova。野外视频深度：从未知摄像机进行无监督单目深度学习。发表于《IEEE/CVF国际计算机视觉会议》论文集，第 8977–8986 页，2019 年。1、3、7、8[30] Xiaodong Gu、Weihao Yuan、Zuozhuo Dai、Chengzhou Tang、Siyu Zhu 和 Ping Tan。DRO：用于运动恢复结构的深度循环优化器。arXiv 预印本 arXiv:2103 13201，2021 年。2、3、7、15 。[31] Vitor Guizilini、Rares Ambrus、Sudeep Pillai、Allan Raventos 和 Adrien Gaidon。用于自监督单目深度估计的 3D Packing。发表于《IEEE/CVF计算机视觉与模式识别会议》论文集，第 2485–2494 页，2020 年。1、2、7、15[32] Vitor Guizilini、Rareș Ambruș, Dian Chen、Sergey Zakharov 和 Adrien Gaidon。利用 Transformer 进行多帧自监督深度估计。发表于《IEEE/CVF计算机视觉与模式识别会议》论文集，第 160–170 页，2022 年。2、3、7、14、15[33] Vitor Guizilini、Rui Hou、Jie Li、Rares Ambrus 和 Adrien Gaidon。用于自监督单目深度的语义引导表征学习。arXiv 预印本 arXiv:2002.12319，2020 年。1、3、7[34] Vitor Guizilini、Jie Li、Rareș Ambruș, 和 Adrien Gaidon。用于语义分割的几何无监督域适应。发表于《IEEE/CVF国际计算机视觉会议》论文集，第 8537–8547 页，2021 年。3[35] Richard Hartley 和 Andrew Zisserman。计算机视觉中的多视图几何。剑桥大学出版社，2003 年。2[36] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。发表于《IEEE计算机视觉与模式识别会议》论文集，第 770–778 页，2016 年。4[37] Sungchul Hong、Antyanta Bangunharcana、Jae-Min Park、Minseong Choi 和 Hyu-Soung Shin。基于视觉 SLAM 的行星建设机器人测绘方法。《传感器》，21(22):7715，2021 年。1[38] Sungchul Hong、Pranjay Shyam、Antyanta Bangunharcana 和 Hyuseoung Shin。行星建设场地光照变化环境下的机器人测绘方法。《遥感》，14(4):1027，2022 年。1[39] Po-Han Huang、Kevin Matzen、Johannes Kopf、Narendra Ahuja 和 Jia-Bin Huang。DeepMVS：学习多视图立体视觉。发表于《IEEE计算机视觉与模式识别会议》论文集，第 2821–2830 页，2018 年。3 – 。[40] Junhwa Hur 和 Stefan Roth。用于联合光流与遮挡估计的迭代残差细化。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 5754–5763 页，2019 年。3[41] Sunghoon Im、Hae-Gon Jeon、Stephen Lin 和 In So Kweon。Dpsnet：端到端深度平面扫描立体视觉。arXiv 预印本 arXiv:1905.00538，2019 年。3[42] Max Jaderberg、Karen Simonyan、Andrew Zisserman 等。空间变换网络。神经信息处理系统进展，28：2017–2025，2015 年。4[43] Qi Jia、Shuilian Yao、Yu Liu、Xin Fan、Risheng Liu 和 Zhongxuan Luo。分割、放大并反复处理：以困难方式检测伪装物体。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 4713–4722 页，2022 年。3[44] Shihao Jiang、Dylan Campbell、Yao Lu、Hongdong Li 和 Richard Hartley。通过全局运动聚合学习估计隐藏运动。载于 IEEE/CVF国际计算机视觉会议论文集，第 9772–9781 页，2021 年。5[45] Adrian Johnston 和 Gustavo Carneiro。使用自注意力机制和离散视差体进行自监督单目训练深度估计。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 4756–4765 页，2020 年。2，7，15[46]Alex Kendall、Matthew Grimes 和 Roberto Cipolla。PoseNet：用于实时 6 自由度相机重新定位的卷积网络。载于 IEEE 国际计算机视觉会议论文集，第 2938–2946 页，2015 年。1，2，3，4，18[47] Alex Kendall、Hayk Martirosyan、Saumitro Dasgupta、Peter Henry、Ryan Kennedy、Abraham Bachrach 和 Adam Bry。端到端学习深度立体回归的几何与上下文信息。载于 IEEE 国际计算机视觉会议论文集，第 66–75 页，2017 年。1，2[48] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。arXiv 预印本 arXiv:1412 6980，2014 6 。 。[49] Marvin Klingner、Jan-Aike Term¨ohlen、Jonas Mikolajczyk 和 Tim Fingscheidt。自监督单目深度估计：通过语义引导解决动态物体问题。载于欧洲计算机视觉会议，第 582–600 页。Springer，2020 年。2[50] Jogendra Nath Kundu、Phani Krishna Uppala、Anuj Pahuja 和 R Venkatesh Babu。Adadepth：用于深度估计的无监督内容一致适配。载于 IEEE计算机视觉与模式识别会议论文集，第 2656–2665 页，2018 年。2[51] Iro Laina、Christian Rupprecht、Vasileios Belagiannis、Federico Tombari 和 Nassir Navab。通过全卷积残差网络实现更深层次的深度预测。载于 2016 年第四届国际三维视觉会议（3DV）论文集，第 239–248 页。IEEE，2016 年。1，2[52] Jin Han Lee、Myung-Kyu Han、Dong Wook Ko 和 Il Hong Suh。从大到小：用于单目深度估计的多尺度局部平面引导。arXiv 预印本 arXiv:1907.10326，2019 年。2

[53] Bo Li、Chunhua Shen、Yuchao Dai、Anton Van Den Hengel 和 Mingyi He。利用深度特征回归和分层条件随机场，从单目图像估计深度与表面法线。载于 IEEE计算机视觉与模式识别会议论文集，第 1119–1127 页，2015 年。1。[54] Hanhan Li、Ariel Gordon、Hang Zhao、Vincent Casser 和 Anelia Angelova。动态场景中的无监督单目深度学习。arXiv 预印本 arXiv:2010.16404，2020 年。7[55] Ruibo Li、Ke Xian、Chunhua Shen、Zhiguo Cao、Hao Lu 和 Lingxiao Hang。用于鲁棒深度预测的深度注意力分类网络。载于亚洲计算机视觉会议，第

663 678 2018 2 – 页。

[56] Springer，。Zhengqi Li 和 Noah Snavely。Megadepth：利用互联网照片学习单视图深度预测。载于 IEEE计算机视觉与模式识别会议论文集，第 2041–2050 页，2018 年。2[57] Zhengfa Liang、Yiliu Feng、Yulan Guo、Hengzhu Liu、Wei Chen、Linbo Qiao、Li Zhou 和 Jianfeng Zhang。通过特征恒常性学习视差估计。载于 IEEE计算机视觉与模式识别会议论文集，第 2811–2820 页，2018 年。2[58] Guosheng Lin、Anton Milan、Chunhua Shen 和 Ian Reid。Refinenet：用于高分辨率语义分割的多路径细化网络。载于 IEEE计算机视觉与模式识别会议论文集，第 1925–1934 页，2017 年。3[59] Chao Liu、Jinwei Gu、Kihwan Kim、Srinivasa G Narasimhan 和 Jan Kautz。神经 RGB（R）D 感知：从摄像机视频中获取深度与不确定性。载于 IEEE/CVF 计算机视觉与模式识别会议论文集，第 10986–10995 页，2019 年。3[60] Jiawei Liu、Jing Zhang 和 Nick Barnes。用于伪装物体检测的置信度感知学习。arXiv 预印本 arXiv:2106.11641，2021 年。3[61] Xiaoxiao Long、Lingjie Liu、Christian Theobalt 和 Wenping Wang。具有自适应法线约束的遮挡感知深度估计。载于欧洲计算机视觉会议，第 640–657 页。Springer，2020 年。3[62] Chenxu Luo、Zhenheng Yang、Peng Wang、Yang Wang、Wei Xu、Ram Nevatia 和 Alan Yuille。每个像素都至关重要++：通过三维整体理解联合学习几何与运动。《IEEE模式分析与机器智能汇刊》，42(10)：2624–2641，2019 年。7，15[63] Xuan Luo、Jia-Bin Huang、Richard Szeliski、Kevin Matzen 和 Johannes Kopf。一致的视频深度估计。《ACM图形学汇刊》（ToG），39(4)：71–1，2020 年。1[64] Reza Mahjourian、Martin Wicke 和 Anelia Angelova。利用三维几何约束，从单目视频无监督学习深度与自运动。载于 IEEE计算机视觉与模式识别会议论文集，第 5667–5675 页，2018 年。1[65] Nikolaus Mayer、Eddy Ilg、Philipp Fischer、Caner Hazirbas、Daniel Cremers、Alexey Dosovitskiy 和 Thomas Brox。什么样的合成训练数据有助于学习视差与光流估计？《国际计算机视觉期刊》，126(9)：942–960，2018 年。2

[66] Nikolaus Mayer、Eddy Ilg、Philip Hausser、Philipp Fischer、Daniel Cremers、AlexeyDosovitskiy 和 ThomasBrox。用于训练视差、光流和场景流估计卷积网络的大型数据集。载于《IEEE计算机视觉与模式识别会议论文集》，第 4040–4048页，2016 年。2[67] Arsalan Mousavian、Hamed Pirsiavas 和JanaKoˇseck´a。使用深度卷积网络进行联合语义分割和深度估计。载于 2016 年第四届国际三维视觉会议（3DV），第 611–619 页。IEEE，2016 年。2[68] Raul Mur-Artal 和Juan D Tard´os。Orb-slam2：用于单目、双目和 RGB-D相机的开源同步定位与建图系统。IEEE机器人学汇刊，33(5)：1255–1262，2017 年。3，8[69] Adam Paszke、Sam Gross、FranciscoMassa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga 等。PyTorch：一种命令式风格的高性能深度学习库。《神经信息处理系统进展》，32，2019 年。6[70] Vaishakh Patil、Wouter Van Gansbeke、Dengxin Dai和Luc Van Gool。不要忘记过去：从单目视频进行循环深度估计。IEEE机器人与自动化快报，5(4)：6813–6820，2020 年。7，15[71] Pedro O Pinheiro、Tsung-Yi Lin、RonanCollobert和 Piotr Doll´ar。学习细化目标分割区域。载于《欧洲计算机视觉会议》，第 75–91页。Springer，2016 年。3[72] Koutilya PNVR、Hao Zhou和 David Jacobs。Sharingan：结合合成数据和真实数据进行无监督几何估计。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 13974–13983 页，2020 年。3[73] Anurag Ranjan、VarunJampani、Lukas Balles、Kihwan Kim、Deqing Sun、JonasWulff和Michael J Black。竞争性协作：深度、相机运动、光流和运动分割的联合无监督学习。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第12240–12249 页，2019 年。1，7，15[74] Chitwan Saharia、Jonathan Ho、William Chan、Tim Sali- mans、DavidJ Fleet和Mohammad Norouzi。通过迭代细化实现图像超分辨率。arXiv:2104.07636，2021 年。3[75] Paul-Edouard Sarlin、Ajaykumar Unagar、MansLarsson、Hugo Germain、Carl Toft、Viktor Larsson、Marc Pollefeys、Vincent Lepetit、LarsHammarstrand、Fredrik Kahl 等。回到特征：从像素到位姿学习鲁棒的相机定位。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 3247–3257页，2021 年。2，3，6[76] Tianwei Shen、Zixin Luo、Lei Zhou、Hanyu Deng、RunzeZhang、TianFang 和 Long Quan。超越光度损失：自监督自运动估计。载于 2019 年国际机器人与自动化会议（ICRA），第 6359–6365 页。IEEE，2019 年。8[77] Chang Shu、KunYu、ZhixiangDuan 和 Kuiyuan Yang。用于深度自监督学习的特征度量损失以及

自运动。载于《欧洲计算机视觉会议》，第 572–588 页。Springer，2020。1, 2[78] Pranjay Shyam、Antyanta Bangunharcana 和 Kyung-Soo Kim。在低光照条件下保持图像特征匹配性能。载于《2020年第20届国际控制、自动化与系统会议（ICCAS）》，第 1079–1085 页。IEEE，2020。1[79] J¨urgen Sturm、Nikolas Engelhard、Felix Endres、Wolfram Burgard 和 Daniel Cremers。RGB-D SLAM系统评测基准。载于《2012 IEEE/RSJ智能机器人与系统国际会议》，第 573–580 页。IEEE，2012。6[80] Zachary Teed 和 Jia Deng。DeepV2D：通过可微分运动结构实现从视频到深度。arXiv 预印本 arXiv:1812.04605，2018。1, 3[81] Zachary Teed 和 Jia Deng。RAFT：用于光流的循环全对场变换。载于《欧洲计算机视觉会议》，第

402 419 2020 3, 4, 5, 16 –[82]

页。Springer，。Zachary Teed 和 Jia Deng。DROID-SLAM：适用于单目、立体和 RGB-D相机的深度视觉 SLAM。神经信息处理系统进展，34 16569，2021 16558 3 : –。[83] Fabio Tosi、Filippo Aleotti、Pierluigi Zama Ramirez、Matteo Poggi、Samuele Salti、Luigi Di Stefano 和 Stefano Mattoccia。通过视频蒸馏语义，实现全面的场景理解。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第 4654–4665 页，2020。2[84] Jonas Uhrig、Nick Schneider、Lukas Schneider、Uwe Franke、Thomas Brox 和 Andreas Geiger。稀疏不变卷积神经网络。载于《2017年国际三维视觉会议（3DV）》，第 11–20 页。IEEE，2017。14, 15[85] Shinji Umeyama。两个点模式之间变换参数的最小二乘估计。《IEEE模式分析与机器智能汇刊》，13(04):376–380，1991。6[86] Benjamin Ummenhofer、Huizhong Zhou、Jonas Uhrig、Nikolas Mayer、Eddy Ilg、Alexey Dosovitskiy 和 Thomas Brox。DEMON：用于学习单目立体视觉的深度与运动网络。载于《IEEE计算机视觉与模式识别会议论文集》，第 5038–5047 页，2017。2, 3[87] Igor Vasiljevic、Vitor Guizilini、Rares Ambrus、Sudeep Pillai、Wolfram Burgard、Greg Shakhnarovich 和 Adrien Gaidon。用于深度和自运动自监督学习的神经射线表面。载于《2020年国际三维视觉会议（3DV）》，第 1–11 页。IEEE，2020。1[88] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力机制就是你所需要的一切。神经信息处理系统进展，30，2017。2, 3, 7, 14[89] Lukas Von Stumberg、Patrick Wenzel、Qadeer Khan 和 Daniel Cremers。GN-Net：用于多天气重定位的高斯–牛顿损失。《IEEE机器人与自动化快报》，5(2):890–897，2020。3[90] Lukas Von Stumberg、Patrick Wenzel、Nan Yang 和 Daniel Cremers。LM-Reloc：基于 Levenberg–Marquardt的直接视觉重定位。载于《2020年国际三维视觉会议（3DV）》，第 968–977 页。IEEE，2020。3

[91] Chaoyang Wang、Jos´e Miguel Buenaposada、Rui Zhu 和 Simon Lucey。使用直接方法从单目视频学习深度。载于 IEEE计算机视觉与模式识别会议论文集，第 2022–2030 页，2018 年 1、5 –。[92] Jingdong Wang、Ke Sun、Tianheng Cheng、Borui Jiang、Chaorui Deng、Yang Zhao、Dong Liu、Yadong Mu、Mingkui Tan、Xinggang Wang 等。用于视觉识别的深度高分辨率表征学习。《IEEE 模式分析与机器智能汇刊》，43(10)：3349–3364，2020 年。4、6[93] Jianrong Wang、Ge Zhang、Zhenyu Wu、XueWei Li 和 Li Liu。通过隐式线索进行深度估计的自监督联合学习框架。arXiv 预印本 arXiv:2006 09876，2020 年 7、15。。[94] Kaixuan Wang 和 Shaojie Shen。Mvdepthnet：实时多视图深度估计神经网络。载于 2018 年国际三维视觉会议（3DV），第 248–257 页。IEEE，2018 年。3[95] Jamie Watson、Oisin Mac Aodha、Victor Prisacariu、Gabriel Brostow 和 Michael Fir man。时间机会主义者：自监督多帧单目深度。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 1164–1174 页，2021 年。1、3、4、6、7、15[96] Xingkui Wei、Yinda Zhang、Zhuwen Li、Yanwei Fu 和 Xiangyang Xue。Deepsfm：通过深度捆绑调整实现运动恢复结构。载于欧洲计算机视觉会议论文集，第 230–247 页。Springer，2020 年。3[97] Yi Wei、Shaohui Liu、Yongming Rao、Wang Zhao、Jiwen Lu 和 Jie Zhou。Nerfingmvs：面向室内多视图立体的神经辐射场引导优化。载于 IEEE/CVF国际计算机视觉会议论文集，第 5610–5619 页，2021 年。3[98] Diana Wofk、Fangchang Ma、Tien-Ju Yang、Sertac Karaman 和 Vivienne Sze。Fastdepth：嵌入式系统上的快速单目深度估计。载于 2019 年国际机器人与自动化会议（ICRA），第 6101–6108 页。IEEE，2019 年。2[99] Ke Xian、Chunhua Shen、Zhiguo Cao、Hao Lu、Yang Xiao、Ruibo Li 和 Zhenbo Luo。使用网络立体数据监督的单目相对深度感知。载于 IEEE计算机视觉与模式识别会议论文集，第 311–320 页，2018 年。2[100] Nan Yang、Lukas von Stumberg、Rui Wang 和 Daniel Cremers。D3vo：用于单目视觉里程计的深度、位姿和不确定性估计。载于 IEEE/CVF计算机视觉与模式识别会议论文集，第 1281–1292 页，2020 年。3[101] Yao Yao、Zixin Luo、Shiwei Li、Tian Fang 和 Long Quan。Mvsnet：面向非结构化多视图立体的深度推断。载于欧洲计算机视觉会议论文集（ECCV），第 767–783 页，2018 年。3[102] Zhichao Yin 和 Jianping Shi。Geonet：密集深度、光流和相机位姿的无监督学习。载于 IEEE计算机视觉与模式识别会议论文集，第 1983–1992 页，2018 年。1、8

[103] Qihang Yu、Lingxi Xie、Yan Wang、Yuyin Zhou、ElliotK Fishman 和AlanLYuille。循环显著性变换网络：结合多阶段视觉线索进行小器官语义分割。在IEEE计算机视觉与模式识别会议论文集，第8280–8289页，2018年。3[104] Amir R Zamir、Alexander Sax、William Shen、Leonidas J Guibas、Jitendra Malik 和Silvio Savarese。Taskonomy：解耦任务迁移学习。在IEEE计算机视觉与模式识别会议论文集，第3712–3722页，2018年。2[105] Huangying Zhan、RaviGarg、Chamara Saroj Weerasekera、KejieLi、HarshAgarwal和 IanReid。使用深度特征重建的单目深度估计和视觉里程计无监督学-习。在IEEE计算机视觉与模式识别会议论文集，第340–349页，2018年。2[106] Chi Zhang、Guosheng Lin、Fayao Liu、RuiYao 和Chun-huaShen。Canet：具有迭代细化和注意力少样本学习的类无关语义分割网络。在IEEE/CVF计算机视觉与模式识别会议论文集，第5217–5226页，2019年。3[107] Zhenyu Zhang、Zhen Cui、Chunyan Xu、ZequnJie、XiangLi 和 JianYang。用于语义分割和深度估计的联合任务递归学习。在欧洲计算机视觉会议（ECCV）论文集，第235–251页，2018年。2[108] Shanshan Zhao、HuanFu、MingmingGong 和 Dacheng Tao。面向几何的单目深度估计对称域适应。在IEEE/CVF计算机视觉与模式识别会议论文集，第9788–9798页，2019年。3[109] Hang Zhou、David Greenwood和SarahTaylor。具有内部特征融合的自监督单目深度估计。arXiv预印本arXiv:2110.09482，2021年。3, 4,7[110] Tinghui Zhou、Matthew Brown、NoahSnavely 和 David GLowe。通过视频进行深度和自运动的无监督学习。在IEEE计算机视觉与模式识别会议论文集，第1851–1858页，2017年。1, 2, 6, 8[111] Yuliang Zou、Pan Ji、Quoc-Huy Tran、Jia-Bin Huang 和Manmohan Chandraker。通过自监督长期建模学习单目视觉里程计。在欧洲计算机视觉会议，第710–727页。Springer，2020年。8, 16[112] Yuliang Zou、Zelun Luo 和Jia-Bin Huang。Df-net：使用跨任务一致性的深度和光流无监督联合学习。在欧洲计算机视觉会议（ECCV）论文集，第36–53页，2018年。1

补充材料

6. DEQ框架

我们遵循深度平衡模型（DEQ） [3,4] 的一般框架，并采用拟牛顿求解器来加速收敛。在实验中，我们使用安德森求解器 [1]。深度平衡模型在不动点 z^* 处计算 $A = I - \partial \mathrm{U} / \partial z^*$ 以获得梯度。通常通过执行另一次不动点迭代来实现这一点。然而，按照 [3,20,23]，我们近似计算$A = I$，并使用非精确梯度进行训练。

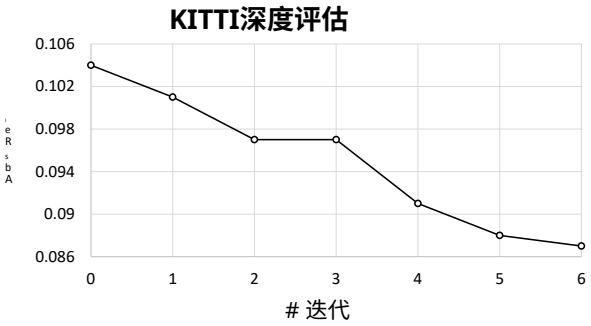


图5. KITTI深度中每次DualRefine迭代的绝对相对误差变化。

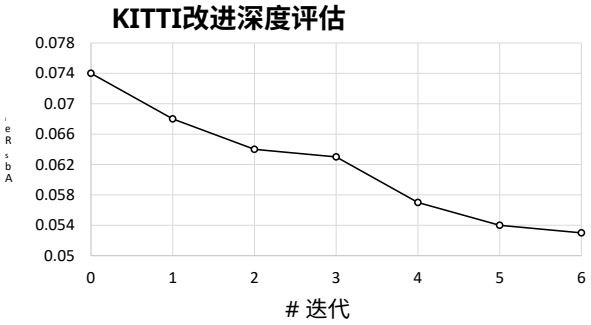


图6. KITTI改进深度中每次DualRefine迭代的绝对相对误差变化。

	损失对 D^*	损失对 T^*	绝对相对误差 ↓	平方相对误差 ↓	均方根误差 ↓	δ_1 ↑
1	T_0	D_0	0.99	0.765	4.449	0.898
2	T^*	D_0	0.093	0.698	4.342	0.907
3	T_0	D^*	0.092	0.657	4.34	0.908
4	T^*	D^*	0.089	0.632	4.305	0.907

表 5.KITTI [22]Eigen划分上姿态更新和自监督配对影响的消融实验。

7. 训练损失组合

确定在精炼后的不动点处计算自监督损失的最佳配对并不简单。对于每个精炼后的估计值（D^* 和 T^*），我们可以使用分离的初始估计值（$[D^* \leftrightarrow$ 与分离的$T_0]$ 配对，以及 $[T^* \leftrightarrow$ 与分离的$D_0]$ 配对）或对应的精炼估计值（$[D^* \leftrightarrow T^*]$配对和 $[D^* \leftrightarrow T^*]$）来计算自监督损失。

比较结果见表格。5。我们发现，当最终估计值均与对应的初始估计值配对时，准确度较差。我们推断，将最终估计值与初始估计值配对，会对模型施加较强的约束，从而限制输出范围。当至少一个最终估计值与对应的初始估计值配对时，结果最佳。例如，深度损失使用$[D^* \leftrightarrow$ 分离的 $T_0]$配对计算，而位姿损失使用 $[T^* \leftrightarrow D^*]$ 配对计算。从这项实验来看，将精炼后的估计值彼此配对似乎能取得最佳准确度。

8. KITTI深度上的更多结果

8.1. KITTI改进深度

在表格6中，我们展示了在稠密真值改进版 [84] KITTI [22] Eigen划分 [16]上的评估结果。我们采用Garg裁剪 [21]并报告距离不超过80米时的误差。我们的细化模块改进了初始估 [88]计，并优于大多数先前模型，同时仍可与基于Transformer的DepthFormer [32] 模型相媲美。

8.2. 深度平衡模型结果

在表 7中，我们展示了模型在每次深度平衡模型迭代后的输出误差。第 0 次迭代对应初始深度估计器生成的深度估计。可以看出，我们的模型在第 6^{th} 次迭代左右收敛。我们还在图 5中绘制了绝对相对误差。

8.3. KITTI改进深度的深度平衡模型结果

我们还在表 8 中展示了详细的深度平衡模型误差，并在图 6 中绘制了 KITTI改进深度真值对应的各次迭代绝对相对误差。类似地，我们的模型在第 6^{th} 次迭代左右收敛。

8.4. 其他定性结果

我们通过图 7 和 8 展示了 KITTI数据集上的其他结果。一个有趣的现象是，模型学会了对车辆和缺乏纹理的图像区域赋予较低的置信度。我们还在图 8 中展示了初始估计与精化后估计之间极线几何的差异，这可能导致光度损失和匹配代价不准确。

	方法	测试帧	宽 ×高	绝对相对误差↓	平方相对误差↓	均方根误差↓	对数均方根误差↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Low & mid res	兰詹 [73]	1	832 × 256	0.123	0.881	4.834	0.181	0.860	0.959	0.985
	EPC++ [62]	1	832 × 256	0.120	0.789	4.755	0.177	0.856	0.961	0.987
	Johnston [45] 等.	1	640 × 192	0.081	0.484	3.716	0.126	0.927	0.985	0.996
	单目深度2 [26]	1	640 × 192	0.090	0.545	3.942	0.137	0.914	0.983	0.995
	Packnet-SFM [31]	1	640 × 192	0.078	0.420	3.485	0.121	0.931	0.986	0.996
	DualRefine-initial (D_0)	1	640 × 192	0.075	0.379	3.490	0.117	0.936	0.989	0.997
	Patil等. [70]	N†	640 × 192	0.087	0.495	3.775	0.133	0.917	0.983	0.995
	王等. [93]	2 (-1, 0)	640 × 192	0.082	0.462	3.739	0.127	0.923	0.984	0.996
	ManyDepth [95]	2 (-1, 0)	640 × 192	0.064	0.320	3.187	0.104	0.946	0.990	0.995
	DepthFormer [32]	2 (-1, 0)	640 × 192	0.055	**0.271**	**2.917**	0.095	0.955	0.991	**0.998**
	DualRefine精炼版（D^*）	2 (-1, 0)	640 × 192	**0.053**	0.290	2.974	**0.092**	**0.962**	**0.992**	**0.998**
High res	DRO [30]	2 (-1, 0)	960 × 320	0.057	0.342	3.201	0.123	0.952	0.989	0.996
	ManyDepth（高分辨率 ResNet50） [95]	2 (-1, 0)	1024 × 320	0.062	0.343	3.139	0.102	0.953	0.991	0.997
	DualRefine 精炼版（D^*）	2 (-1, 0)	960 × 288	**0.050**	**0.250**	**2.763**	**0.087**	**0.965**	**0.993**	**0.998**

表6. 在 KITTI [22] Eigen划分 [16] 上使用改进深度图与其他最先进模型的结果对比 [84]. **粗体**：最佳，下划线：次佳。†：在完整序列上评估。

# Iters	绝对相对误差↓	平方相对误差↓	均方根误差↓	对数均方根误差↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
0	0.104	0.778	4.495	0.181	0.894	0.965	0.983
1	0.101	0.743	4.405	0.179	0.902	0.966	0.983
2	0.097	0.708	4.302	0.176	0.909	0.967	0.983
3	0.097	0.711	4.312	0.176	0.908	0.967	0.983
4	0.091	0.700	4.259	0.172	0.913	0.967	0.983
5	0.088	0.697	4.239	0.170	0.914	0.967	0.983
6	0.087	0.698	4.234	0.170	0.914	0.967	0.983
7	0.088	0.696	4.230	0.171	0.913	0.967	0.983
8	0.088	0.695	4.229	0.172	0.912	0.966	0.983
9	0.089	0.693	4.234	0.173	0.911	0.966	0.983

表7. KITTI [22] Eigen划分上每次DualRefine迭代的误差变化。

# 迭代次数	绝对相对误差↓	平方相对误差↓	均方根误差↓	对数均方根误差↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
0	0.074	0.389	3.390	0.115	0.940	0.990	0.997
1	0.068	0.344	3.201	0.106	0.950	0.991	0.997
2	0.064	0.311	3.100	0.101	0.956	0.992	0.998
3	0.063	0.314	3.105	0.101	0.956	0.992	0.998
4	0.057	0.299	3.029	0.096	0.960	0.992	0.998
5	0.054	0.293	2.995	0.093	0.961	0.992	0.998
6	0.053	0.290	2.974	0.092	0.962	0.992	0.998
7	0.052	0.287	2.962	0.092	0.962	0.992	0.998
8	0.053	0.285	2.963	0.093	0.961	0.992	0.998
9	0.054	0.286	2.979	0.094	0.960	0.992	0.998

表8。KITTI [22] Eigen划分中，每次DualRefine迭代 [16] 使用改进深度图 [84] 后的误差变化。

方法	t_{err}(%) ↓	r_{err}(°/100m) ↓	绝对轨迹误差 (m) ↓
ORB-SLAM2 [61]	12.96	**0.7**	44.09
单目深度2 [22]	12.28	3.1	99.36
Zou等人. [102]	7.28	1.4	71.63
DualRefine-initial（T_0）	12.50	4.04	118.29
DualRefine 精炼版（T^*）	**5.82**	1.51	**17.27**

表 9. 使用 ORB-SLAM2 双目结果作为伪真值，在 KITTI 里程计测试集（序列 11 ∼ 21）上的更多结果。我们与具有代表性的最先进自监督深度和里程计方法进行了比较。ORB-SLAM2 作为一种有代表性的非学习型方法纳入比较。

9. KITTI 里程计上的更多结果

我们使用 ORB-SLAM2 双目版本作为伪真值，按照 Zou 等人的方法，对 KITTI 里程计数据集的序列 11–21 进行了额外评估。 [111] 我们在表 9 中给出了平均结果。精炼大幅改善了初始预测，即使与启用了闭环检测的 ORB-SLAM2 相比，ATE 也表现更优。

10. 卷积门控循环单元更新实现

在我们的方法中，我们使用标准的Conv-GRU模块 [81] 来计算更新，具体如下：

$$\begin{aligned} z_{k+1} &= \sigma(\mathrm{CNN}_z([h_k, x_k])) \\ r_{k+1} &= \sigma(\mathrm{CNN}_r([h_k, x_k])) \\ \tilde{h}_{k+1} &= \tanh(\mathrm{CNN}_{\tilde{h}}([r_{k+1} \odot h_k, x_k])) \\ h_{k+1} &= (1 - z_{k+1}) \odot h_k + z_{k+1} \odot \tilde{h}_{k+1} \end{aligned} \tag{9}$$

其中， σ 表示sigmoid激活函数。未来将考虑探索Conv-GRU模块的其他变体。

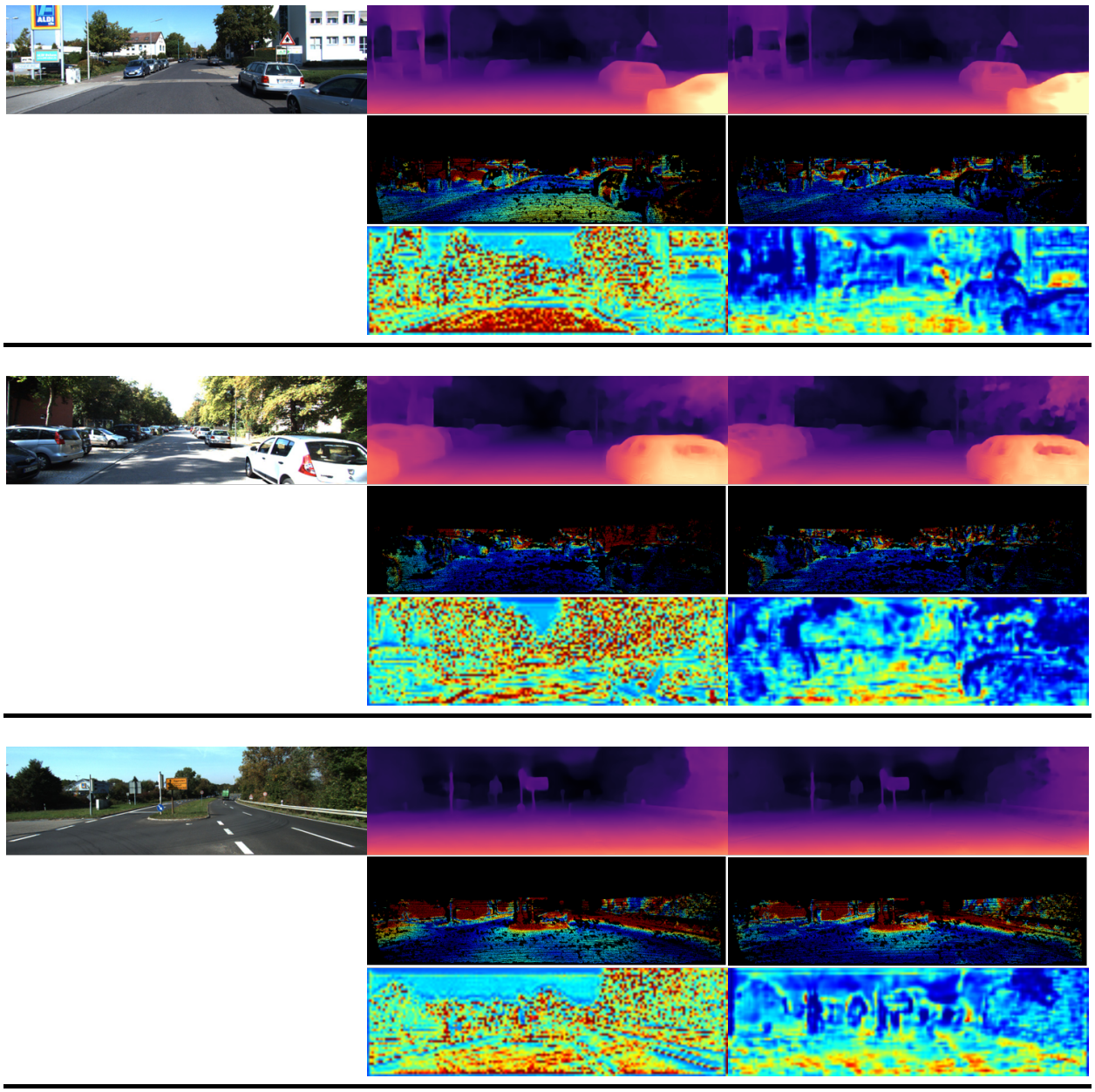


图 7. KITTI 数据集上的定性结果 [22] 。左上：输入图像，中上：初始视差，右上：细化后的视差，中间：初始误差图，右中：细化后的误差图，下中：固定置信度权重，右下： 6^{th} 迭代置信度权重。

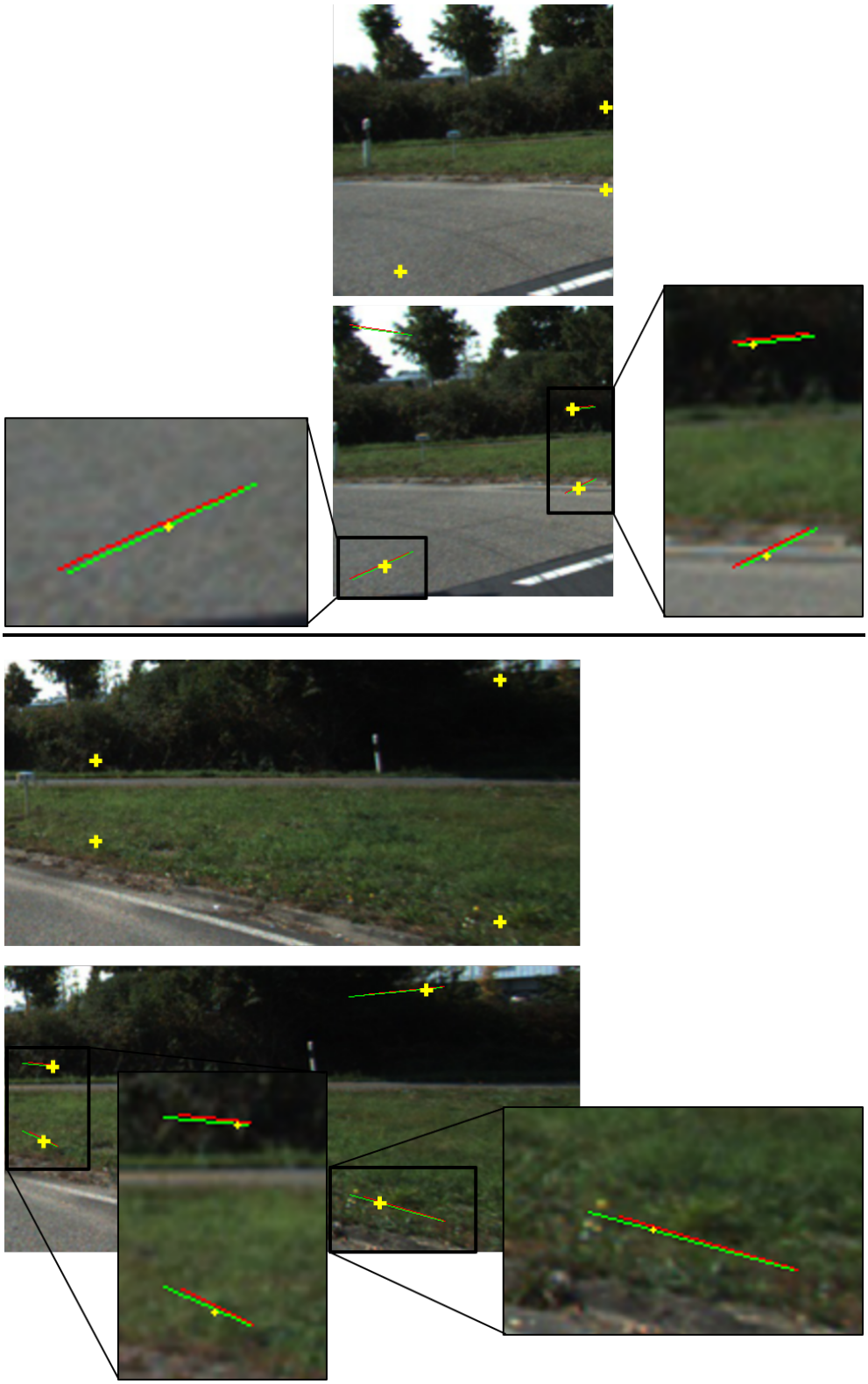


图8。源中的极线（绿色）。源图像中的黄色点图像，计算自 黄色 目标图像中的点，用于 PoseNet [46] 初始位姿（红色）和我们优化后的位姿图像是根据我们的最终深度和位姿估计计算得出的。