Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video

Jia-Wang Bian[1,2], **Zhichao Li**[3], **Naiyan Wang**[3], **Huangying Zhan**[1,2]

Chunhua Shen[1,2], **Ming-Ming Cheng**[4], **Ian Reid**[1,2]

[1]University of Adelaide, Australia [2]Australian Centre for Robotic Vision, Australia
[3]TuSimple, China [4]Nankai University, China

Abstract

近来的工作表明，可以使用无标签的单目视频学习基于 CNN 的深度与自运动估计量。然而，由于未识别的运动物体违反了图像几何重建中的静态场景假设，导致算法性能受限。更严重的是，由于缺少适当的限制，网络在不同样本上输出尺度不一致的结果，*i.e.*，自运动网络不能提供一个长视频序列的完整摄像机轨迹由于每帧尺度的模糊性。这篇文章通过提出一种用于尺度一致性预测的几何一致性损失与一种用于处理运动物体与遮挡的引导自发现掩码来解决这个问题。不同于近期其他工作，我们并未使用多任务学习方法，故我们的框架更加简洁高效。全面的评估结果显示了我们的深度估计量在 KITTI 数据集上达到了 SOTA 的性能。另外，我们展示了我们的自运动网络能够在长视频序列上预测全局尺度一致的摄像机轨迹，并且视觉测距准确率结果与近期使用立体视频训练的模型相比具有竞争力。据我们所知，这是首个表明使用无标签单目视频训练出的深度网络能够在一段长视频序列上全局地预测尺度一致的摄像机轨迹的工作。

1 介绍

深度与自运动估计对于许多机器人和计算机视觉方面的应用是至关重要的。传统方法常为人工设计的分段系统，这种系统依赖对应的研究 [1, 2] 与用于估计的多视角几何学 [3, 4]。近来，基于深度学习的算法 [5, 6] 表明可以使用卷积神经网络（CNN）推断出单张图像的深度。特别地，无监督方法 [7–11] 显示了基于 CNN 的深度与自运动网络可以在单目视频序列上独自训练，无需使用客观事实深度或立体图像对（姿态监督）。其原理是利用预测的深度和自运动可以将一帧的图像扭曲到另一帧，之后采用图像重建损

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

失作为监督信号 [7] 以训练网络。然而，由于运动物体违反了在几何图像重建中最根本的静态场景假设，该方法的性能受限。更严重的是，由于缺少适当的约束，网络在不同样本上输出尺度不一致的结果 *i.e.*，自运动网络不能提供一个长视频序列的完整摄像机轨迹由于每帧尺度的模糊性[1]。

据我们所知，以往的工作（单目视频无监督学习）中还没有解决上述尺度不一致问题。为此，我们提出了一种用于应对这个挑战的几何一致性损失。具体来说，对于任意 2 个从一段视频中采样得到的连续帧，我们将单帧的预测深度图转换到 3 维空间，然后使用推测的自运动将其投影到另外一帧并最终最小化投影深度图与预测深度图之间的不一致性。这显式地使深度网络在连续帧上预测几何一致（同时尺度一致）的结果。通过反复采样与视频训练，每个连续图片对上的深度预测将是尺度一致的并且帧与帧之间的一致性最终能扩展到整个视频序列上。由于自运动尺度与深度的尺度紧密联系，我们提出的自运动网络能在连续的一段视频上预测尺度一致的相关摄像机姿态。我们证明了仅仅简单地积累姿态预测就能得到在一段长视频序列上全局尺度一致的摄像机轨迹 (Fig. 3)。

考虑到运动物体的挑战,近期工作通过引入额外的光流 w [9–11, 13] 或语义分割网络 [14] 以解决这个问题。虽然极大地提升了性能，这种方式也带来了训练中巨大的计算消耗。这里我们展示了我们可以自动地从提出的几何一致性项中找出一个掩码来解决这个问题且不引入新的网络。具体来说，我们可以使用提出的一致性项轻松地定位属于运动物体/遮挡区域或不同区域（*e.g.*，无明显结构的区域）的像素。通过对这些像素分配较低的权重，我们可以避免其对脆弱的图像重建损失的影响 (查看 Fig. 2 以观察掩码的可视化结果)。较于近期使用多任务学习的方法 [9–11]，我们提出的方法能简单高效。

我们进行了细致的消融实验,并清晰阐述了我们方法的效力。另外,在 KITTI [15] 数据集上全面的评估结果展示了我们的深度网络比以往在更复杂的多任务学习框架 [9–11, 16] 上训练的先进模型更加有效。同时，我们的自运动网络能够在长视频序列上推测尺度一致的摄像机轨迹，并且轨迹的准确性与使用立体视频 [17] 训练的先进的模型不相上下。

总而言之，我们的主要贡献有三:

- 我们提出了一种几何一致性约束来加强深度与自运动的尺度一致性，从而得到全局尺度一致的自运动估计量。
- 我们通过上述几何一致性约束提出了一种用于动态场景和遮挡的自探索掩码。相较于其他方法，我们提出的方法不需要二外的光流或语义分割网络，这使学习框架更简单有效。
- 我们提出的深度估计量在 KITTI 上达到 SOTA 性能，并且与使用立体视频训练的最先进模型相比，我们提出的自运动预测器显示了具有竞争力的可视化测距结果。

[1]诸如 ORB-SLAM [12] 的单目系统受到尺度模糊问题影响，但其预测是全局尺度一致的。然而，最近的使用单目视频的学习模型不仅受到尺度模糊性的影响，而且在不同的片段上预测尺度不一致的结果。

2 相关工作

传统方法依赖于同一场景多视角间的不一致性来恢复 3D 场景几何体，这类方法至少需要 2 张图片 [3]。随着深度学习的迅速发展，Eigen et al. [5] 展示了深度可以使用卷积神经网络（CNN）从单张图片中估计得到。具体来说，他们设计了一种从粗糙到精细的网络来预测单视角深度并使用从深度传感器得到的实际深度作为监督信号来训练网络。然而，纵使这些监督方法 [5, 6, 18–21] 展示出高质量流与深度估计结果。真实场景下实际值的获得却代价昂贵。

在不需要实际深度的情况下，Garg et al. [22] 展示了使用立体图像对也可以训练单视角深度网络。他们使用已建立的极线几何学 [3] 而不是深度监督。在左视图与从右视图扭曲得到的合成左视图间颜色的不一致性被用作监督信号。跟随这个想法，Godard et al. [23] 提出了约束左右视图一致性的正则化。另外，Zhan et al. [17] 将这种方法拓展到立体视频。然而，虽然基于立体配对的方法不需要实际深度，准确矫正立体摄影机在真实场景下也并不容易解决。

为了这个目的，Zhou et al. [7] 提出了一种全无监督框架，在此框架中，深度网络能够独自从单目视频中学习。其原理是他们引入一个额外的自运动网络来预测连续帧之间的相对相机姿势。在有了估计的深度和相对姿态的情况下，采用像 [22] 中的图像重建方法，以光度损失作为监督信号。然而，由于动态物体违反了图像几何重建中的静态场景假设，动态物体的出现限制了性能。更重要的是，Zhou et al. [7] 的方法存在每帧尺度的模糊性，在这一点上，一个单一的和一致的相机平移比例丢失，只有方向是已知的。结果，自运动网络不能在长视频序列上预测完全的相机轨迹。

为了应对移动物体，近来的工作 [9, 10] 引入一个额外的光流网络。更近期的工作 [11] 引入额外的动作分割网络。虽然这些方法展现了极大的性能提升，但我们在基本框架下增加了大量的额外计算消耗，然而这些方法依然存在尺度不一致问题。另外，Liu et al. [24] 采用与一致性损失类似的深度投影损失来监督密度，但他们的方法依赖于预计算的 3D 重建以进行监督。

据我们所知，这篇是第一篇展现了训练在单目视频上的自运动网络能够在长视频序列上推测出全局一致的相机轨迹的论文。这展现了在用于机器人与自动驾驶的视觉 SLAM [12] 中利用深度学习方法的强大潜力。

3 无监督下尺度一致的深度与自运动学习

3.1 方法总览

我们的目标时使用单目视频训练深度与自运动网络并限制他们来推测尺度一致的结果。给定 2 个从无标签视频中采样得到的连续帧（I_a，I_b），我们首先使用深度网络推测他们的深度图（D_a，D_b）然后使用姿态网络推测 2 帧间相对的 6D 摄像机姿态 P_{ab}。有了预测深度和相对摄像机姿态，我们可以通过插值原图像 I_b 合成参考图像 I'_a [25, 7]。然后，可以通过真实图像 I_a 与合成图像 I'_a 的光度损失监督网络。然而，由于动态场景违

反了图像重建中的几何假设，这个基本网络的性能有限。为了这个目的，我们提出了一种几何一致性损失 L_{GC} 以保证尺度一致和一种自发现掩码 M 以解决运动物体和遮挡。Fig. 1展示了我们提出的损失与掩码的说明。我们整体的目标函数如下：

$$L = \alpha L_p^M + \beta L_s + \gamma L_{GC}, \tag{1}$$

这里 L_p^M 代表通过掩码 M 加权的光度损失（L_p），L_s 代表平滑损失。我们在前向和反向共同训练网络以最大化数据使用并且为了简化我们只获取前向损失。在后续章节，我们首先在 Sec. 3.2中引入广泛使用的光度损失和平滑损失，然后在 Sec. 3.3阐述我们提出的几何一致性损失并在 Sec. 3.4说明自发现掩码。

3.2 光度损失与平滑损失

光度损失. 利用在经典稠密相关算法 [26] 中光照对比度和空间平滑先验，以往的工作 [7, 9–11] 使用扭曲帧与参考帧间的光照损失作为无监督损失函数来训练网络。根据预测深度图 D_a 和相对相机姿态 P_{ab}，我们通过扭曲 I_b 合成 I'_a，其中如 [7] 使用可微双线性插值 [25]。根据合成图 I'_a 与参考图 I_a，我们构造如下目标函数

$$L_p = \frac{1}{|V|} \sum_{p \in V} \|I_a(p) - I'_a(p)\|_1, \tag{2}$$

这里 V 代表成功从 I_a 投影到 I_b 上的有效点集。$|V|$ 定义为 V 中点的个数。我们选择对离群点鲁棒的 L_1 损失。然而，它仍然不是真实世界场景中照明变化的不变性。这里我们添加额外的图像相异损失 SSIM [27] 以更好的处理复杂光照改变，因为它规范化了像素照明。我们改变光度损失项 Eqn. 2为：

$$L_p = \frac{1}{|V|} \sum_{p \in V} (\lambda_i \|I_a(p) - I'_a(p)\|_1 + \lambda_s \frac{1 - \text{SSIM}_{aa'}(p)}{2}), \tag{3}$$

这里 $\text{SSIM}_{aa'}$ 代表 I_a 间 I'_a 通过 SSIM 函数 [27] 得到的逐元素相似性。根据 [23, 9, 11]，我们在框架中使用 $\lambda_i = 0.15$ 和 $\lambda_s = 0.85$。

平滑损失. 由于光度损失在场景中缺少纹理或均匀的区域不具有信息性，现有的工作在正则化估计深度图中加入了平滑度先验。我们采纳了在 [11] 中使用的边缘感知平滑损失，其形式为：

$$L_s = \sum_p (e^{-\nabla I_a(p)} \cdot \nabla D_a(p))^2, \tag{4}$$

这里 ∇ 是沿空间方向的一次导数，其保证了平滑由图像边缘引导。

3.3 几何一致性损失

如上所述，我们在预测结果上加强了几何一致性。具体来说，我们要求 D_a 与 D_b（通过 P_{ab} 关联）符合相同的 3D 场景结构并最小化差异。优化过程不仅鼓励一批中样本间的几何一致性，而且将一致性传递到整个序列。*e.g.*，I_1 的深度与 I_2 的深度在同一批次中一致；I_2 的深度与 I_3 的深度在另一个训练批次中一致。最终，一个序列中 I_i 的深度

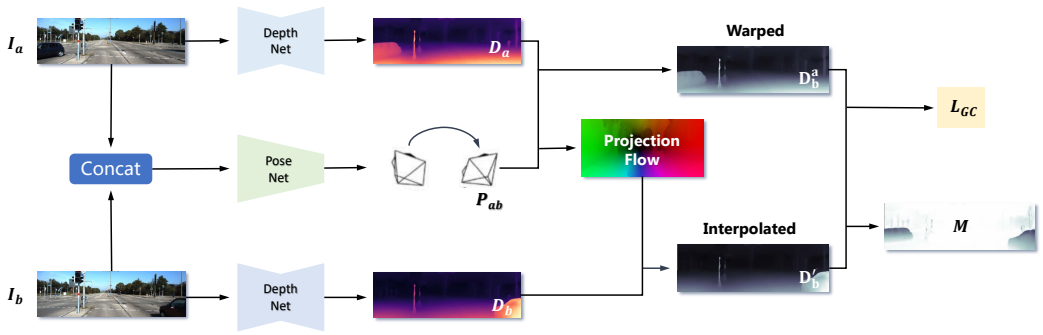


图 1: 我们提出的几何一致性损失和自发现掩码的说明。给定 2 个连续帧（I_a，I_b），我们首先使用网络预测他们的深度图（D_a，D_b）和相对姿态（P_{ab}），然后我们通过转换 D_a 到 3D 空间并使 P_{ab} 投影到 I_b 的图像平面上以获取扭曲的（D_b^a）。最终，我们如同几何一致性损失 L_{GC}（Eqn. 6）中使用 D_b^a 与从 D_b 中插值得到的 D'_b 间的不一致性监督网络训练。这里，我们将 D_b 插值是因为投影流不在 I_b 的网格上。另外，我们从不一致性图中挖掘出一个用于处理动态场景和病态估计区域（Fig. 2）的掩码 M （Eqn. 7）。为清楚起见，光度损失和平滑损失未在图中展示

应全部相互一致。由于姿态网络在训练过程中与深度网络自然耦合，因此我们的方法可以在整个序列中产生尺度一致的预测。有了这个限制，我们计算深度不一致性图 D_{diff}。对于每个 $p \in V$，其定义为：

$$D_{\text{diff}}(p) = \frac{|D_b^a(p) - D'_b(p)|}{D_b^a(p) + D'_b(p)} \tag{5}$$

这里 D_b^a 是 I_b 通过使用 P_{ab} 扭曲后的 D_a 计算得到的深度图，D'_b 是从预测深度图 D_b 中插值得到的深度图（注意到我们不能直接使用 D_b，由于扭曲流不在像素网格上）。这里我们通过总和归一化他们的差异。这比绝对距离更直观，因为它在优化中平等地对待不同绝对深度的点。另外，函数是对称的，输出自然在 01 之间，其有助于训练中的数值稳定性。有了不一致性图，我们简单地定义几何一致性损失为：

$$L_{GC} = \frac{1}{|V|} \sum_{p \in V} D_{\text{diff}}(p), \tag{6}$$

其最小化每个连续对间预测深度的几何距离并强调尺度一致。通过训练，这种一致性可以传播到整个视频序列中。由于自运动和深度预测之间的紧密联系，自运动网络最终可以预测全局尺度一致的轨迹。

3.4 自发现掩码

为了处理可能损害网络训练的移动物体和遮挡，近来的工作引入了额外的光流 [9–11] 或语义分割网络 [14]。这虽然有效，但其引入了额外的计算代价和训练负担。这里，我们展示了在 Eqn. 5中的不一致性图 D_{diff} 可以有效定位这些区域。有几种场景会导致从不同视图观察到的场景结构不一致，其中包括（1）动态物体，（2）遮挡与（3）对困难区域的不准确预测。在不对这些情况显式进行分离时，我们观察到每一个都会导致 Ddiff 从

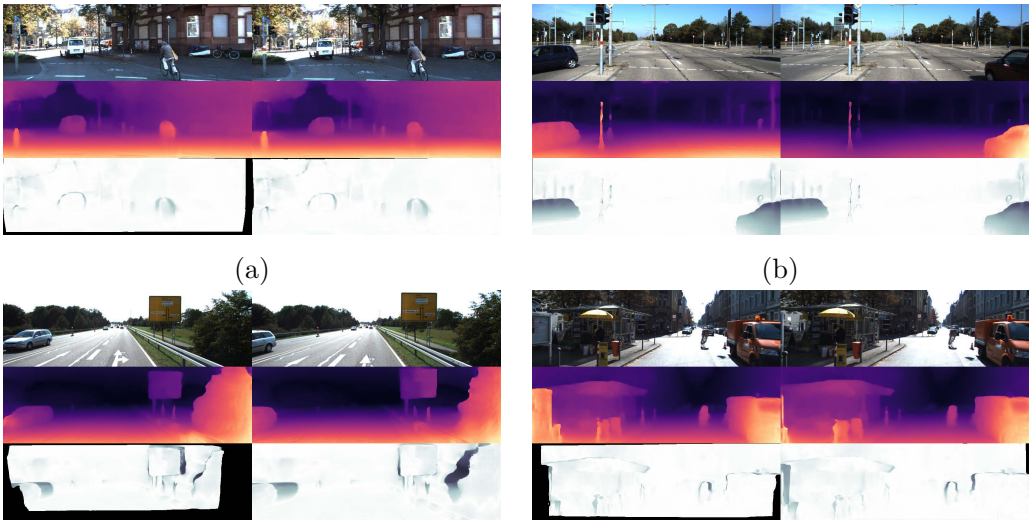


图 2: 可视化结果。从上至下：样本图，估计深度，自发现掩码。我们提出的掩码可以有效地定位遮挡和移动物体。

其理想值零开始增加。基于这些简单的观察，我们提出了一个如 D_{diff} 一样值域在 $[0, 1]$ 间权重掩码 M：

$$M = 1 - D_{\text{diff}}, \tag{7}$$

其为不一致/一致的像素分配低/高权重。它可以用于重新分配光度损失的权重。具体来说，我们改变在 Eqn. 3中的光度损失，如下

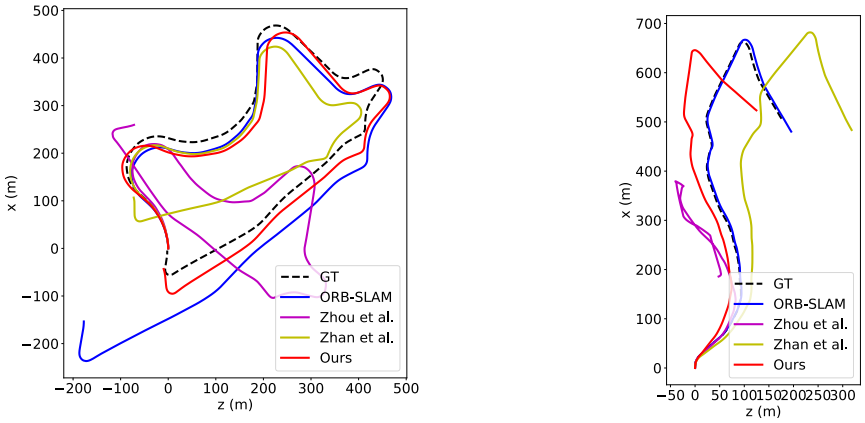
$$L_p^M = \frac{1}{|V|} \sum_{p \in V} (M(p) \cdot L_p(p)). \tag{8}$$

通过使用这个掩码，我们减轻了运动物体和遮挡物的不利影响。进一步，在误差预测区域上计算的梯度在反向传播过程中的权重较小。Fig. 2展示了所提出的掩模的视觉效果，这与我们上述的预期相吻合。

4 实验

4.1 实现细节

网络结构. 对于深度网络，我们在 DispNet [7]DispResNet [11] 与上进行实验，这二者均接受单张 RGB 图像作为输入并输出一张深度图。对于自运动网络，我们使用去除掩码分支的 PoseNet [7]。网络从拼接的 RGB 图像对中预测 6D 相对摄像机姿态。我们不在深度图的多尺度（在 [7] 中用 4 种尺度或在 [11] 中用 6 种尺度 6 ）输出上计算损失，而是通过经验发现使用单尺度监督（*i.e.*，仅在最精细的输出上计算损失）是更好的选择（Tab. 4)。我们的单尺度监督不仅提升了性能而且提供了更加精准的训练流程。我们假设这种现象的原因是在低分辨率图像中光度损失不精确，其中像素颜色被过度平滑。



(a) sequence 09 (b) sequence 10

图 3: 在 KITTI 测距数据集 [15] 测试序列上的定性结果。

单视角深度估计. 我们提出的学习框架使用 PyTorch [28] 实现。对于深度网络，我们在 KITTI raw 数据集 [15] 上使用 Eigen [5] 的分割训练与测试，这与相关工作 [10, 9, 11, 7] 的做法相同。与 [7] 相同，用我们用三个连续视频帧的片段作为训练样本，这里我们设置第二张图片为参考帧来计算与其他 3 张图片的损失然后反转他们的角色，再次计算损失以最大化数据使用。在训练中，数据同样经过随机缩放、裁减与水平反转这些数据增强，并且我们在 2 中输入分辨率（416×128 and 832×256）下进行实验。我们使用 ADAM [29] 优化器，并设置每批大小为 4，学习了为 10^{-4}。在训练中，对于 Eqn. 1我们采用 $\alpha = 1.0$，$\beta = 0.1$，与 $\gamma = 0.5$。我们共训练 200 个周期，每个周期包含 1000 张随机采用的样本，并验证一次。并且，我们在 CityScapes [30] 数据集上对网络进行预训练然后再 KITTI [15] 数据集上进行微调，每个阶段均 200 个周期。这里我们采用 Eigen et al. [5] 中的评估准则作为深度评估。

可视化测距结果. 对于姿态网络，如 Zhan et al. [17] 所做，我们在 KITTI 测距数据集 [15] 上评估可视化测距结果，这里序列 00-08/09-10 用作训练/测试。我们使用数据集的标准评估指标进行轨迹评估，而不是使用 Zhou et al. [7] 的 5 帧姿势评估，因为它们应用更广泛，更有意义。

4.2 与 SOTA 的对比

KITTI raw 数据集上的深度结果. Tab. 1 展示了 KITTI raw 数据集上的结果 [15]，当其他对比模型也在单目视频序列训练时，我们的方法达到了 SOTA 性能。注意到最近的工作 [9–11, 31] 全部是联合在多任务上训练，而我们的方法并没有。这有效地减少了训练与推理开销。而且，我们的方法与其他使用更强监督信号如校准的立体图像对（*i.e.*，姿态监督）甚至是事实深度标注的方法相比具有相当的竞争力。

KITTI 测距数据集上的可视化测距结果. 我们与 SfMLearner [7] 和在立体视频上训练的方法 [17] 进行对比。我们也报告了 ORB-SLAM [12] 系统（除去回路闭合）的结果

表 1: KITTI raw 数据集 [15] 测试部分的单视角深度估计结果。在 KITTIraw 数据集上 [15] 训练的方法标记为 K. Models, 在 CityScapes [30] 上预训练的标注为 CS+K。(D) 表示深度监督，(B) 表示双目. 立体输入对，(M) 表示单目视频片段。(J) 表示多任务联合训练。表格每块中最好的结果进行了加粗。

		Error ↓				Accuracy ↑		
Methods	Dataset	AbsRel	SqRel	RMS	RMSlog	< 1.25	$< 1.25^2$	$< 1.25^3$
Eigen et al. [5]	K (D)	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Liu et al. [6]	K (D)	0.202	1.614	6.523	0.275	0.678	0.895	0.965
Garg et al. [22]	K (B)	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Kuznietsov et al. [18]	K (B+D)	**0.113**	**0.741**	**4.621**	**0.189**	**0.862**	**0.960**	**0.986**
Godard et al. [23]	K (B)	0.148	1.344	5.927	0.247	0.803	0.922	0.964
Godard et al. [23]	CS+K (B)	0.124	1.076	5.311	0.219	0.847	0.942	0.973
Zhan et al. [17]	K (B)	0.144	1.391	5.869	0.241	0.803	0.928	0.969
Zhou et al. [7]	K (M)	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Yang et al. [31] (J)	K (M)	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian et al. [8]	K (M)	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Wang et al. [16]	K (M)	0.151	1.257	5.583	0.228	0.810	0.936	0.974
Geonet-VGG [9] (J)	K (M)	0.164	1.303	6.090	0.247	0.765	0.919	0.968
Geonet-Resnet [9] (J)	K (M)	0.155	1.296	5.857	0.233	0.793	0.931	0.973
DF-Net [10] (J)	K (M)	0.150	1.124	5.507	0.223	0.806	0.933	0.973
CC [11] (J)	K (M)	0.140	**1.070**	**5.326**	**0.217**	0.826	0.941	**0.975**
Ours	K (M)	**0.137**	1.089	5.439	**0.217**	**0.830**	**0.942**	**0.975**
Zhou et al. [7]	CS+K (M)	0.198	1.836	6.565	0.275	0.718	0.901	0.960
Yang et al. [31] (J)	CS+K (M)	0.165	1.360	6.641	0.248	0.750	0.914	0.969
Mahjourian et al. [8]	CS+K (M)	0.159	1.231	5.912	0.243	0.784	0.923	0.970
Wang et al. [16]	CS+K (M)	0.148	1.187	5.496	0.226	0.812	0.938	0.975
Geonet-Resnet [9] (J)	CS+K (M)	0.153	1.328	5.737	0.232	0.802	0.934	0.972
DF-Net [10] (J)	CS+K (M)	0.146	1.182	5.215	0.213	0.818	0.943	**0.978**
CC [11] (J)	CS+K (M)	0.139	**1.032**	**5.199**	0.213	0.827	0.943	0.977
Ours	CS+K (M)	**0.128**	1.047	5.234	**0.208**	**0.846**	**0.947**	0.976

表 2: KITTI 数据集 [15] 上的可视化测距结果。我们将 ORB-SLAM [12] 的性能作为参考并与最近的深度方法进行对比。K 表示模型在 KITTI 上训练，CS+K 表示模型在 Cityscapes [30] 进行预训练。

Methods	Seq. 09		Seq. 10	
	t_{err} (%)	r_{err} (°/100m)	t_{err} (%)	r_{err} (°/100m)
ORB-SLAM [12]	15.30	0.26	3.68	0.48
Zhou et al. [7]	17.84	6.78	37.91	17.78
Zhan et al. [17]	11.93	3.91	12.45	**3.46**
Ours (K)	11.2	3.35	**10.1**	4.96
Ours (CS+K)	**8.24**	**2.19**	10.7	4.58

作为参考。虽然强调这导致了一个简单的帧到帧姿态估计框架与视觉 SLAM 系统的比较注记，后者有一个强大的后端优化系统 (*i.e.*, bundle adjustment [32]) 来提高性能。这里，我们忽略序列 (09 和 10) 中的帧 (分别为前 9 和 30)，对于这些帧，ORB-SLAM [12]

由于初始化失败而无法输出相机姿势。Tab. 2显示测试序列 09 和 10 的平均平移和旋转误差，Fig. **??**显示定性结果。注意到这种比较对所提出的方法非常不利：*i)* 由于 [7] 的尺度不一致，我们将每帧尺度与客观事实尺度对齐，而我们的方法只对齐一个全局尺度；*ii)* [17] 需要立体视频训练，而我们只使用单目视频。即使在对我们方法不公平的情况下，我们的结果相比 [17] 仍有竞争力。甚至当与 ORB-SLAM [12] 系统对比时，在序列 09 上我们的方法显示了更低的平移误差和更好的视觉效果。在未标记的单目视频上训练的深度模型可以预测全局尺度一致的视觉测距，这是一个显著的进步。

4.3 消融实验

在这节中，我们首先验证了我们提出的几何一致性损失 L_{GC} 和自发现掩码 M 的有效性。然后我们在不同尺度数量、网络结构与图像分辨率下进行实验。

验证 L_{GC} 与 M. 我们使用 DispNet [7] 进行消融实验，图像分辨率为 416×128。Tab. 3显示了单尺度与多尺度监督下的深度结果。结果清晰地阐释了我们提出的项对整体性能的贡献。并且，Fig. 4展示了训练中的验证误差，这表明 L_{GC} 能够有效地避免模型过拟合。

表 3: L_{GC} 与 M 的消融实验. 括号中显示多尺度（4）监督的结果.

	Error ↓				Accuracy ↑		
Methods	AbsRel	SqRel	RMS	RMSlog	< 1.25	$< 1.25^2$	$< 1.25^3$
Basic	0.161 (0.185)	1.225	5.765	0.237	0.780	0.927	0.972
Basic+SSIM	0.160 (0.163)	1.230	5.950	0.243	0.775	0.923	0.969
Basic+SSIM+GC	0.158 (0.161)	1.247	5.827	0.235	0.786	0.927	0.971
Basic+SSIM+GC+M	**0.151** (**0.158**)	**1.154**	**5.716**	**0.232**	**0.798**	**0.930**	**0.972**

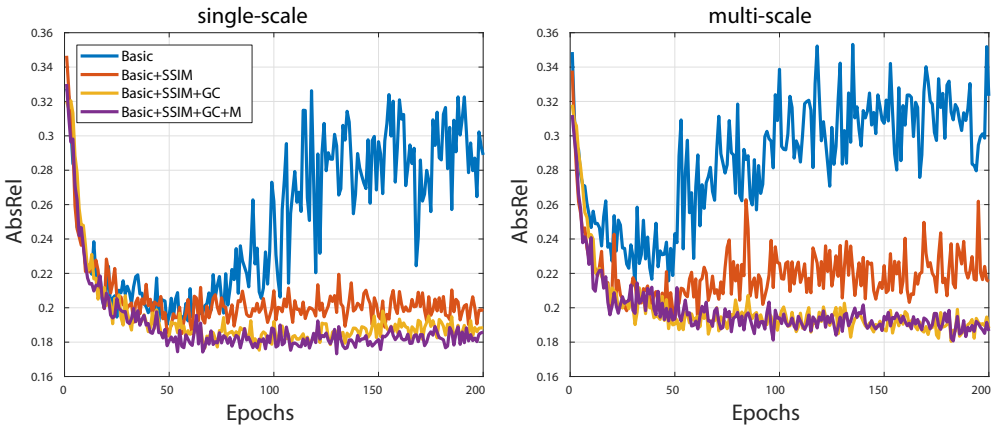


图 4: 验证误差. *Basic* 与 *Basic+SSIM* 均在越 50 周期后过拟合，而其他的模型由于包含 L_{GC} 不会过拟合。另外，单尺度监督下的模型在训练中超越了那些多尺度（4）监督的模型。

单尺度 vs 多尺度监督. 如 Sec. 4.1中所述，我们实证发现，使用单尺度监督比使用广泛使用的多尺度解决方案具有更好的性能。Tab. 4展示了结果。我们假设的原因是光度损失是不准确的低分辨率图像，其中像素的颜色过于平滑。此外，由于两个连续视图之间的位移很小，因此不需要进行多尺度求解。

表 4: 不同尺度监督的消融实验.

	Error ↓				Accuracy ↑		
#Scales	AbsRel	SqRel	RMS	RMSlog	< 1.25	$< 1.25^2$	$< 1.25^3$
1	**0.151**	**1.154**	**5.716**	**0.232**	**0.798**	**0.930**	**0.972**
2	0.152	1.192	5.900	0.235	0.795	0.927	0.971
3	0.159	1.226	5.987	0.240	0.780	0.921	0.969
4	0.158	1.214	5.898	0.239	0.782	0.925	0.971

网络结构与图像分辨率. Tab. 5显示了不同网络结构不同尺度图像下的结果,这里 DispNet 与 DispResNet 都从 CC [11] 借鉴，DispNet 也在 SfMLearner [7] 用到。结果表明，高分辨率的图像和更深的网络可以产生更好的性能。

表 5: 不同网络结构与图像分辨率上的消融实验.

		Error ↓				Accuracy ↑		
Methods	Resolutions	AbsRel	SqRel	RMS	RMSlog	< 1.25	$< 1.25^2$	$< 1.25^3$
DispNet	416×128	0.151	1.154	5.716	0.232	0.798	0.930	0.972
DispResNet	416×128	0.149	1.137	5.771	0.230	0.799	0.932	0.973
DispNet	832×256	0.146	1.197	5.578	0.223	0.814	0.940	**0.975**
DispResNet	832×256	**0.137**	**1.089**	**5.439**	**0.217**	**0.830**	**0.942**	**0.975**

4.4 Timing and memory analysis

Training time and parameter numbers. 我们与 CC [11] 进行对比，这两种方法都是在一个 16GB 的 Tesla V100 上训练的。我们使用批大小为 4 来测量每个训练迭代（包括向前和向后传递）所花费的时间。图像分辨率为 832×256。CC [11] 需要 3 个部分，包括（深度，姿态）、流与掩码。相反，我们的方法只训练（深度，姿态）。总体上，如其作者报道，CC 消耗大概 *7 days* 进行训练，而我们的方法消耗大概 *32 hours*。Tab. 6展示了每次迭代的时间与每个网络的参数。

表 6: 每次迭代的训练时间与每个网络的模型参数.

	CC [11]			Ours
Network	(Depth, Pose)	Flow	Mask	(Depth, Pose)
Time	0.96s	1.32s	0.48s	0.55s
Parameter Numbers	(80.88M, 2.18M)	39.28M	5.22M	(80.88M, 1.59M)

推理时间. 我们在 RTX 2080 测试模型。批大小为 1，时间为 100 次迭代时间取平均。DispNet 与 DispResNet 的结构与 SfMLearner [7] 和 CC [11] 中各自相同。所以他们的速度理论上是一样的。Tab. 7展示了结果。

表 7: 每个图像/图像对上的推理时间.

	DispNet	DispResNet	PoseNet
128×416	4.9 ms	9.6 ms	0.6 ms
256×832	9.2 ms	15.5 ms	1.0 ms

5 Conclusion

本文提出了一种尺度一致深度和自运动估计的无监督学习框架。该方法的核心是尺度一致性的几何一致性损失和处理动态场景的自发现掩模。在所提出的学习框架下，我们的深度模型在 KITTI [15] 数据集上达到了最先进的性能，并且我们的自运动网络可以得出与训练在立体视频上的模型具有竞争力的可视化测距结果。据我们所知，这是首个表明在未标记的单目视频上进行深度模型训练可以预测长序列上全局尺度一致的相机轨迹的工作。在未来的工作中，我们将致力于改善视觉测距的准确性，将漂移校正解决方案纳入目前的框架。

Acknowledgments

The work was supported by the Australian Centre for Robotic Vision, the Major Project for New Generation of AI (No. 2018AAA0100400), and NSFC (NO. 61922046). Jiawang would also like to thank TuSimple, where he started research in this field.

References

[1] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal on Computer Vision (IJCV)*, 60(2), 2004.

[2] Jia-Wang Bian, Wen-Yan Lin, Yun Liu, Le Zhang, Sai-Kit Yeung, Ming-Ming Cheng, and Ian Reid. GMS: Grid-based motion statistics for fast, ultra-robust feature correspondence. *International Journal on Computer Vision (IJCV)*, 2019.

[3] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[4] Jia-Wang Bian, Yu-Huan Wu, Ji Zhao, Yun Liu, Le Zhang, Ming-Ming Cheng, and Ian Reid. An evaluation of feature matchers for fundamental matrix estimation. In *British Machine Vision Conference (BMVC)*, 2019.

[5] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Neural Information Processing Systems (NIPS)*, 2014.

[6] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Recognition and Machine Intelligence (PAMI)*, 38(10), 2016.

[7] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[8] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] Zhichao Yin and Jianping Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In *European Conference on Computer Vision (ECCV)*, 2018.

[11] Anurag Ranjan, Varun Jampani, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive Collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular slam system. *IEEE Transactions on Robotics (TRO)*, 31(5), 2015.

[13] Yang Wang, Zhenheng Yang, Peng Wang, Yi Yang, Chenxu Luo, and Wei Xu. Joint unsupervised learning of optical flow and depth by watching stereo videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] Jianbo Jiao, Ying Cao, Yibing Song, and Rynson Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *European Conference on Computer Vision (ECCV)*, 2018.

[15] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets Robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013.

[16] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[17] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[19] Chengzhou Tang and Ping Tan. BA-Net: Dense bundle adjustment network. In *International Conference on Learning Representations (ICLR)*, 2019.

[20] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[21] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[22] Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision (ECCV)*. Springer, 2016.

[23] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] Xingtong Liu, Ayushi Sinha, Mathias Unberath, Masaru Ishii, Gregory D Hager, Russell H Taylor, and Austin Reiter. Self-supervised learning for dense depth estimation in monocular endoscopy. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 128–138. Springer, 2018.

[25] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Neural Information Processing Systems (NIPS)*, 2015.

[26] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal on Computer Vision (IJCV)*, 56(3), 2004.

[27] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image Quality Assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*, 13(4), 2004.

[28] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[29] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[31] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry with edge-aware depth-normal consistency. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[32] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*. Springer, 1999.