Unsupervised Learning of Depth and Ego-Motion from Video

Tinghui Zhou UC Berkeley
Matthew Brown Google
Noah Snavely Google
David G. Lowe Google

摘要

我们提出了一种无监督学习框架，用于从非结构化视频序列中进行单目深度和相机运动估计。与最近的工作[1][2][3]一样，我们使用端到端学习方法，将视图合成作为监控信号。与之前的工作相比，我们的方法是完全无监督的，只需要单目视频序列进行训练。我们的方法使用单视图深度和多视图姿态网络，损失基于使用计算的深度和位姿将附近的视图扭曲到目标。因此，网络在训练期间通过损耗耦合，但可以在测试时独立应用。对 KITTI 数据集的经验评估证明了我们方法的有效性：1）单目深度与使用地面真实位姿或深度进行训练的监督方法相比表现良好，2）位姿估计与在可比输入设置下建立的 SLAM 系统相比表现良好。

关键词：无监督学习; 单目深度估计和相机位姿估计; SfM ; SLAM

1 引言

Struct from motion, 从运动中恢复三维结构，这是这篇文章完成的主要任务，具体来说就是完成了无监督的单目深度估计和相机位姿估计任务，而无监督的单目深度估计是这篇文章的亮点，这在 2017 年是最早一批将有监督的单目深度估计向无监督学习推进的工作，并且在当时与其他有监督方法相比也颇具竞争力, 所完成的单目深度估计也是我的主要关注点。深度估计是计算机视觉中一个很重要的基础性问题，如果我们能够获取三维物体的精确的深度信息，那么我们就能更高效更准确地完成大量计算机视觉中的下游任务，比如即时定位和建图、AR、VR、三维重建、自动驾驶等等。获取深度信息的途径有很多，比如激光雷达、双目相机，这两种方案能够直接获取深度信息，但是缺点明显，比如激光雷达，体积大而且成本高；而双目相机通过视差几何计算出深度，计算量大、难以应对复杂场景并且相机基线限制了测量范围。最理想的途径是采用单目相机采集图像，然后通过这些图像通过某种方法能直接获取到每个像素点的深度信息，而这种方法就是单目深度估计。一个理想的单目深度估计模型可以做到以任意图像作为输入，由模型给出它的深度图，这样的端到端模型结合模型压缩等方法更是有望能够集成到移动端硬件，相比激光雷达等传统方法，具有成本低、精度高、运算快等优势。最早的基于深度学习的单目估计模型是有监督的，监督学习可能会比无监督学习更容易实现，效果也可能更好，但是监督学习的弊端在单目深度估计任务中会被放得更大：带有深度信息的 ground truth 难以获取，因而无监督的单目深度估计更有发展潜力，也更符合理想模型所需满足的条件。《Unsupervised Learning of Depth and Ego-Motion from Video》正是最早一批提出无监督的单目深度估计和相机位姿估计的经典论文，该论文以 view synthesis(视图合成) 作为监督信号，通过视图合成构建了一个损失函数，迫使模型输出对深度和相机位姿的良好估计[4]。通过对这篇文章进行复现，我认为可以让我更好地理解无监督单目深度估计和位姿估计，并且学习以视图合成作为监督信号构建损失函数这一重要思想。

2 相关工作

2.1 Struct from motion

Struct from motion 可以被译为从运动中恢复三维结构，它的意思是通过二维的图像对或者视频序列中恢复出相应的三维信息，其中包括成像摄像机的运动参数以及场景的结构信息。Sfm 已经是一个经过充分研究的问题，有很长的研究历史，并且已经有一系列已建立的技术工具链[5][6][7]。尽管这些传统方法在很多情况下是高效的，但是它对准确的图像的对应关系的依赖可能会在以下情况出现问题：弱纹理、复杂的几何/光度测量、薄结构和遮挡。而深度学习具有强大拟合能力，并且能够在训练期间利用外部监督，在测试数据时可能克服上述的问题。因而，也出现了很多将深度学习应用在 Sfm 相关问题的工作，比如特征匹配、位姿估计、深度估计等等。

2.2 基于扭曲的视图合成

几何场景理解的一个重要应用是新视图合成，也就是从相机的新视角合成场景的外观。经典的视图合成首先要么是显式估计三维场景的底层几何结构要么建立输入视图的像素级别的对应关系，然后通过从输入视图合成图像块来合成新视图（比如[8][9]）。而一些方法则是采用了端到端的学习方法来进行新视图合成，基于深度或者光流对输入进行转换，比如 DeepStereo[10]，Deep3D[11]和 Appearance Flows[10.1007/978-3-319-46493-0_18] 。对于这些方法，场景的底层几何由量化的深度平面（DeepStereo）、概率视差图（Deep3D）和视图依赖流场（Appearance Flows）表示。这些方法不同于直接从输入视图映射到目标视图的方法，基于扭曲变形的方法被迫学习几何信息或者是对应关系的中间预测。而在复现论文最终，目标就是从 CNN 中提炼出这种几何推理能力，并训练为可以实现这种基于扭曲的视图合成的模型。

2.3 通过二维观测学习单视图三维信息

此外，复现论文与一系列的从二维观测学习单视图三维信息密切相关。这些相关工作有：Garg 等人[13]提出使用投影误差来学习单视图深度估计 CNN，以对校准的立体孪生进行监督。同时，Deep3D[11]预测了第二个立体视点从使用立体电影片段作为训练的输入图像数据 Godard 等人[3]采用了类似的方法添加了左右一致性约束，以及更好的架构设计，从而带来了令人印象深刻的性能。与复现论文类似，这些技术仅会从图像观察中学习而与需要真实深度进行训练的方法不同，通过调整相机参数和场景深度来最小化基于像素的误差来完成结构和运动的估计。然而，与其直接将损失最小化来获得估计，基于 CNN 的方法只会对每批输入实例采取梯度下降，这将允许网络从大量相关图像中学习隐式先验。而要构建这样的学习框架，前提就是这个前向传播过程必须是可微的，而有部分作者已经探索了在以这种方式训练的模型中构建可微渲染操作，比如[14][15][16]。复现论文中也采取了可微渲染操作，并从二维的视频序列中学习到了场景的深度信息。

2.4 基于视频的无监督/自监督学习

另一项相关工作是基于视频的无监督/自监督学习，目标通常是从视频数据中学习通用的视觉特征然后用于其他计算机视觉任务，比如目标检测和语义分割。当我们专注于推断出现实的场景几何和自我运动时，直觉上，从深度网络中学习而来的内部推断（特别是在单视图深度 CNN）应该能够捕获到某些层次的语义，而这些语义可以推广到其他任务。和复现论文同一时期的另一项工作，Vijayanarasimhan 等人[17]也提出了一个基于视频的深度、相机运动和场景运动联合训练框架。该工作与复现论文在概念

上相似，但是复现论文的重点是无监督，而提到的这个框架则是采用了监督。并且在训练过程中场景动力学建模方式也是有显著差异的，他们明确地解决了物体运动，而复现论文的可解释性掩模不考虑运动、遮挡和其他因素。

3 本文方法

3.1 本文方法概述

本文采取了一个双网络联合训练的结构如图 1，由 Depth CNN 负责估计深度图，Pose CNN 负责估计送入的连续帧之间的相对位姿，利用估计的深度以及相对位姿完成视图合成任务，构建损失函数，通过优化损失函数实现更加正确的视图合成，从而迫使模型作出更准确的深度估计和位姿估计。其中，Depth CNN 和 Pose CNN 会进行联合训练，使用它们的输出构建损失函数实现共同优化，在模型训练完成后，Depth CNN 和 Pose CNN 则可以单独使用，完成各自的任务。

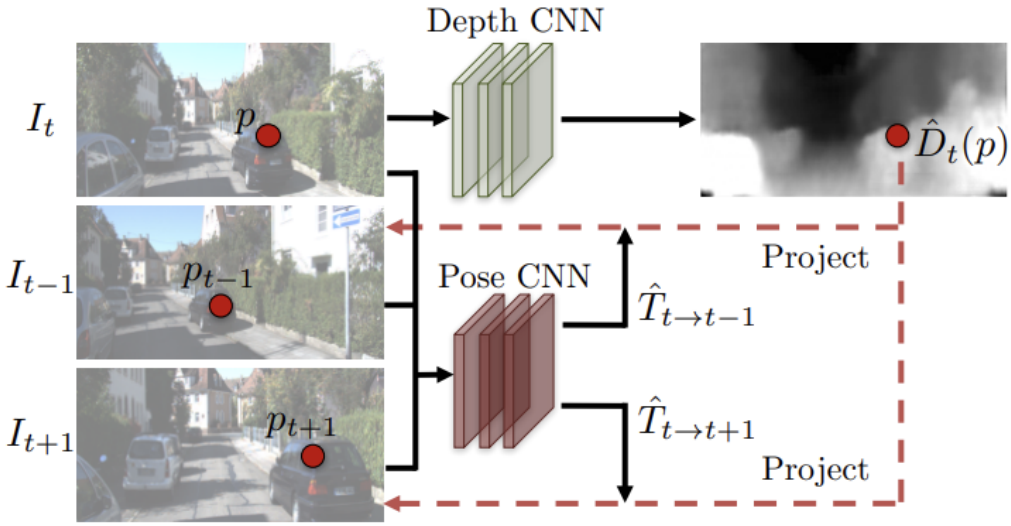

图 1: 整体结构

3.2 Depth CNN

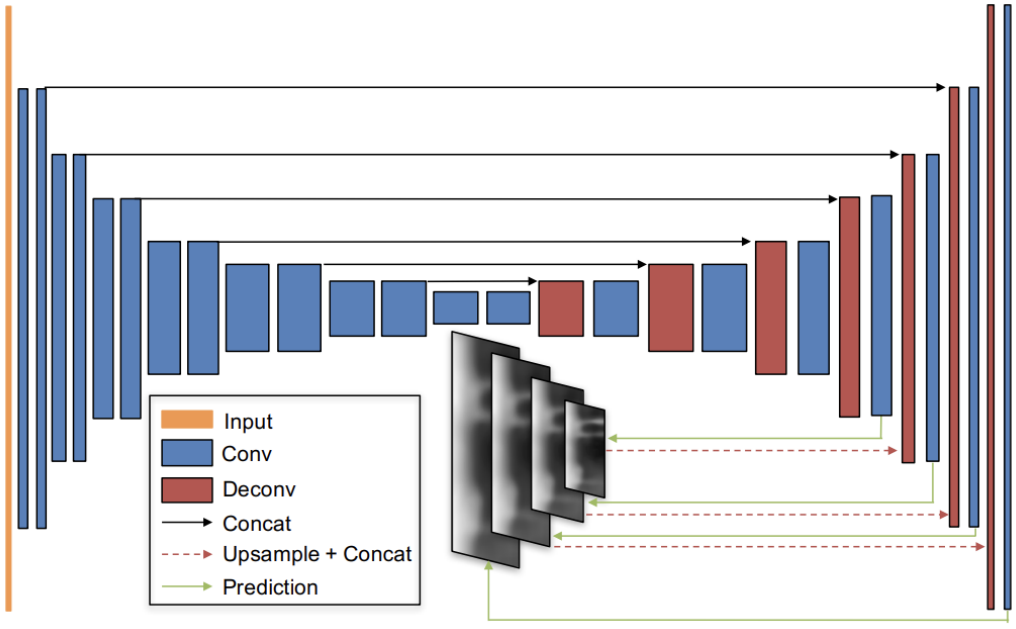

图 2: Depth CNN

如上图 2所示为 Depth CNN 的结构，它采用了类似与 UNet 的编解码模型，以单幅图片作为输入，最终以原始尺寸输出输入图片的深度图（视差图）。值得注意的是，Depth CNN 在训练过程中会输出四张尺寸不一样的视差图，这四张尺寸不同的视差图取自解码的不同阶段，在计算损失函数的时候会充分利用到不同尺寸视差图和可解释性掩模，目的是应对弱纹理、错误估计距离实际情况很远、梯度局部性的问题，利用不同尺寸的数据计算损失函数最后在累加起来得到最终的损失值。

3.3 Pose CNN

如下图 3所示为 Pose CNN 的结构，也是类似于 UNet 的编解码模型，不过其以三张图片（一般情况下，其中一张为目标视图，其余两张为源视图）作为输入，在编码阶段就会输出两个 6 维的位姿向量（表示目标视图到两个源视图之间的变换关系）。而在解码阶段，类似于 Depth CNN，在训练过程会输出不同尺度的可解释掩模，可解释掩模的作用就是对图像的异常区域进行过滤，在计算损失的时候忽视异常区域。

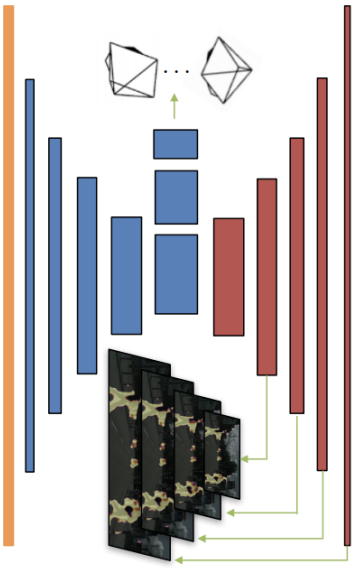

图 3: Pose CNN

3.4 损失函数定义

当通过 Depth CNN 得到目标视图的深度图，通过 Pose CNN 得到目标视图和源视图之间的位姿变换后，我们就可以构建损失函数了。该方法的损失函数主要由三个部分组成：光度一致性损失、平滑损失、可解释性掩模正则项，其中最关键的则是光度一致性损失。

$$\mathcal{L}_{final} = \sum_{l} \mathcal{L}_{vs}^{l} + \lambda_s \mathcal{L}_{smooth}^{l} + \lambda_e \sum_{s} \mathcal{L}_{reg}(\hat{E}_s^l) \tag{1}$$

1式是最终的损失函数，可以见到其由三个部分组成，其中第一部分就是光度一致性损失，是最重要的部分，光度一致性损失如式 2所示。

$$\mathcal{L}_{vs} = \sum_{<I_1,\ldots,I_N>\in S} \sum_{p} \hat{E}_S(p) |I_t(p) - \hat{I}_s(p)| \tag{2}$$

要计算光度一致性误差，首先我们得通过生成的深度图和位姿变换得到 $\hat{I}_s(p)$，主要是通过一个可微深度图像渲染过程获得的，如图 4所示。

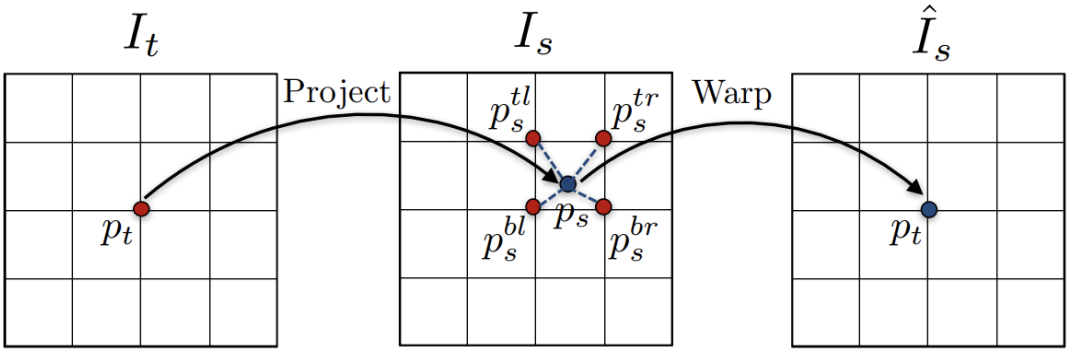

图 4: 损失函数定义

$$p_s \sim K\hat{T}_{t\to s}\hat{D}_t(p_t)K^{-1}p_t \tag{3}$$

通过上式，利用可微双边采样，线性插值 p_s 四个相邻像素的值来近似 $I_s(p_s)$，通过投影几何学得到的像素 warping 的坐标能够充分利用估计的深度和估计的相机位姿。得到 $\hat{I}_s(p)$ 后，与 $I_t(p)$ 相减就是光度的相差，我们期待这两个值是一样的，那么就代表估计出来的深度和相机位姿正确，因此最小化这个损失来达到我们的目的。此外，$\hat{E}S(p)$ 则是可解释性掩模，只有在合法区域该项才为 1，从而达到减少异常值对训练的影响。

第二项 $\lambda_s\mathcal{L}^l_{smooth}$ 是平滑损失，主要是为了克服梯度的局部性，允许梯度能直接从更大的空间区域得到。通过该思路，由于它对网络结构框架不敏感，预测深度图时，我们通过最小化二阶梯度的 L_1 范数来得到。

第三项是可解释性掩模的正则项，由于在最小化光度一致性损失的过程中，合法区域的可解释性掩模值为 1，我们想要最小化损失函数的过程中，就会不可避免地使得可解释性掩模 $\hat{E}_S$ 趋向于 0，为了解决这个问题，通过增加正则项来解决，使其成为 $\mathcal{L}_{reg}(\hat{E}_{sl})$，这个正则项通过最小化交叉熵损失来鼓励非零的预测。

4 复现细节

4.1 与已有开源代码的不同点（创新点）

整个复现工作主要参考了本篇论文 sfmlearner 由 Clement Pinard 实现的非官方的 pytorch 实现。在完成主体部分的基础上，对损失函数进行了修改，以期达到更好的效果。对现有的光度一致性损失部分，增添了一个 SSIM[18]损失项，新的光度一致性损失如下式 4。

$$\mathcal{L}_{vs} = \sum_{<I_1,\dots,I_N>\in S}\sum_p \hat{E}_S(p)(\lambda|I_t(p)-\hat{I}_s(p)| + (1-\lambda)\frac{1-SSIM_{ts'}(p)}{2}) \tag{4}$$

SSIM 是一种测量图片相似性的方法，称为结构相似性，其目的是通过光度、对比度、结构三个方面评判两幅图片的相似性，其公式如下：

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c1)(\sigma_{xy}+c2)}{(\mu_x^2+\mu_y^2+c1)(\sigma_x^2+\sigma_y^2+c2)} \tag{5}$$

原作中的光度一致性损失通过两幅图像的光度（也就是像素值）直接判断两幅图像的相似性，而

目的就是优化模型使得通过估计的深度和估计的位姿生成的图像与目标图像一致，但仅通过光度进行判断不能处理特殊情况，比如光照的变化，因此可以考虑对图像中的其他隐含信息也进行对比，使得两幅图像的相似更加合理，从而使得模型向更准确地估计深度和位姿的目标优化。这一思路参考了sc-depth[19]。

4.2 实验过程

本实验主要使用 KITTI 数据集[20]进行基准测试，以评估系统的性能，详细参数均在代码中展示。实验的平台是 ubuntu20.04，cpu 为 intel512400，gpu 为 3070laptop。

4.2.1 模型训练

主要训练两个模型，第一个是根据原作的 pytorch 实现提供的参数进行复现的模型，第二个是改进了损失函数之后的模型。对于复现的模型，采用 batchsize=4，可解释性掩模权重为 0.2，平滑损失权重为 0.1，epochsize=3000，sequence-length=3 的参数进行模型的训练，选取的数据集为经过处理和划分的 kitti 数据集，将这个复现的模型称为复现模型。第二个是在改进了损失函数之后的模型，基本参数不变，SSIM 权重为 0.15，选取的数据集也不变，将这个改进后的模型称为改进模型。

4.2.2 模型评估

训练后的模型将会得到两个子模型（depth 和 pose），使用对应的评估代码对这两个子模型进行评估，对 depth 模型进行评估，采用 kitti 数据集中划分为测试集的部分，而对 pose 模型进行评估，采用 kitti_odometry 数据集的 sequence09 和 sequence10。

5 实验结果分析

对复现和改进模型的 depth 子模型进行评估以及比较，得到如下数据：

	Abs Rel	Sq Rel	RMSE	RMSE(log)	Acc.1	Acc.2	Acc.3
作者	0.181	1.341	6.236	0.262	0.733	0.901	0.964
复现模型	0.1986	1.4834	6.2817	0.2701	0.6959	0.8972	0.9645
改进模型	0.3458	6.8177	8.7599	0.4116	0.6178	0.8305	0.9188

表 1: depth 模型评估与原作对比

如表 1 所示，复现模型和原作的 pytorch 实现的评估数据非常接近，可见复现工作基本成功，而改进后的模型并没有得到效果上的提升，这次改进只能算是一次尝试，由于损失函数的改变，相关的最优超参数应该也会发生变化，但是完整地进行一次训练需要花费十小时以上，时间成本巨大，目前还没探索出最优的超参数。

对复现和改进模型的 pose 子模型进行评估以及比较，得到如下数据：

	ATE mean(std)	RE mean(std)
作者	0.0179(0.0110)	0.0018(0.0009)
复现模型	0.0190(0.0137)	0.0036(0.0026)
改进模型	0.0354(0.0162)	0.0036(0.0023)

表 2: pose 模型评估与原作对比 Seq.09

	ATE mean(std)	RE mean(std)
作者	0.0141(0.0115)	0.0018(0.0011)
复现模型	0.0156(0.0143)	0.0037(0.0029)
改进模型	0.0254(0.0170)	0.0036(0.0026)

表 3: pose 模型评估与原作对比 Seq.10

如表 2 和 3 所示，分别是模型 Seq.09 和 Seq.10 数据集上的表现，可见复现模型与原作的模型也是十分接近，但是改进模型的问题依然一样，效果不太好。

利用复现得到的 depth 模型，对部分测试集进行视差图的预测，效果如下：

图 5: 左原图，右视差图

6 总结与展望

单目深度估计是具有重要意义的计算机视觉经典问题，如果能仅采用单目相机进行精确的深度估计，那么大量下游应用的性能和效率也会得到提升，同时，成本的大量下降也会催生更多的实际应用。本文给出了一种基于单目相机的深度估计和相机姿态估计的深度学习方法，并且与前人的工作不同，该方法是无监督的，所以具有开创性。尽管该方法能达到不错的效果，但仍有一些可以进一步改进的地方，该方法要求相机的内参已知，如果能将相机内参作为优化变量一同学习的话，就能够省去相机标定的工作，同时，网上的大量的不知道相机参数的数据也能够使用。此外，该方法对移动物体等异常区域提出的可解释性掩模效果一般，所以对于动态场景等难以处理的数据仍然有很大的进步空间。针对复现工作来说，复现基本达到要求，但是改进的损失函数并没有使得模型有很好的提升，具体原因还需要进一步探索，但通过这些尝试能使我更加熟悉深度学习的基本流程以及多视图几何、损失函数等立体视觉以及深度学习的相关知识。

参考文献

[1] FLYNN J, NEULANDER I, PHILBIN J, et al. Deep Stereo: Learning to Predict New Views from the World's Imagery[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

2016.

[2] GARG R, BG V K, CARNEIRO G, et al. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue[J]. Springer, Cham, 2016.

[3] GODARD C, AODHA O M, BROSTOW G J. Unsupervised Monocular Depth Estimation with Left-Right Consistency[C]//Computer Vision & Pattern Recognition. 2017.

[4] ZHOU T, BROWN M, SNAVELY N, et al. Unsupervised Learning of Depth and Ego-Motion from Video[J]., 2017.

[5] FURUKAWA Y, CURLESS B, SEITZ S M, et al. Towards Internet-scale multi-view stereo[J]., 2010.

[6] WU. C. VisualSFM: A visual structure from motion system[J]., 2011.

[7] NEWCOMBE R A, LOVEGROVE S J, DAVISON A J. DTAM: Dense tracking and mapping in real-time[C]//IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011. 2011.

[8] CHEN S E, WILLIAMS L. View Interpolation Image Synthesis[J]. Proc Siggraph, 1993: 279-288.

[9] C., Lawrence, Zitnick, et al. High-quality video view interpolation using a layered representation[J]. ACM Transactions on Graphics, 2004, 23(3): 600-608.

[10] FLYNN J, NEULANDER I, PHILBIN J, et al. Deep Stereo: Learning to Predict New Views from the World's Imagery[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[11] XIE J, GIRSHICK R, FARHADI A. Deep3D: Fully Automatic 2D-to-3D Video Conversion with Deep Convolutional Neural Networks[J]. Springer International Publishing, 2016.

[12] ZHOU T, TULSIANI S, SUN W, et al. View Synthesis by Appearance Flow[C]//LEIBE B, MATAS J, SEBE N, et al. Computer Vision – ECCV 2016. Cham: Springer International Publishing, 2016: 286-301.

[13] GARG R, BG V K, CARNEIRO G, et al. Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue[J]. Springer, Cham, 2016.

[14] HANDA A, BLOESCH M, PATRAUCEAN V, et al. gvnn: Neural Network Library for Geometric Computer Vision[J]. Springer, Cham, 2016.

[15] KULKARNI T D, WHITNEY W, KOHLI P, et al. Deep Convolutional Inverse Graphics Network[J]. MIT Press, 2015.

[16] LOPER M M, BLACK M J. OpenDR: An Approximate Differentiable Renderer[C]//European Conference on Computer Vision. 2014.

[17] VIJAYANARASIMHAN S, RICCO S, SCHMID C, et al. SfM-Net: Learning of Structure and Motion from Video[J]., 2017.

[18] WANG Z. Image Quality Assessment : From Error Visibility to Structural Similarity[J]. IEEE Transactions on Image Processing, 2004.

[19] BIAN J W, ZHAN H, WANG N, et al. Unsupervised Scale-Consistent Depth Learning from Video[J]., 2021.

[20] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//IEEE Conference on Computer Vision & Pattern Recognition. 2012.