

关于论文“PlanarRecon: Real-time 3D Plane Detection and Reconstruction from Posed Monocular Videos”的复现

摘要

本文复现的论文 PlanarRecon 是一种用于从已知位姿的单目视频中全局连贯检测和重建三维平面的新型框架。与以往从单幅图像中检测平面的工作不同，PlanarRecon 利用神经网络提取并融合每个视频片段（由一组关键帧组成）的场景三维体素特征表示，增量检测三维平面。基于学习的跟踪和融合模块用于合并前前后片段中的平面，以形成连贯的全局平面重建。这种设计使 PlanarRecon 能够整合每个片段中多个视图的观察结果和不同片段的时间信息，从而准确、连贯地重建具有低多边形几何结构的场景抽象。PlanarRecon 在 ScanNetv2 数据集上实现了最先进的性能，同时具有实时性。

另外，本文针对 PlanarRecon 存在无法恢复场景平面正确法向的问题进行改进，利用相机位姿修正所重建的平面的法向，从实验结果来看，算法较为合理的修正了面片方向。

关键词：三维平面重建；单目视频重建；简化多边形重建

1 引言

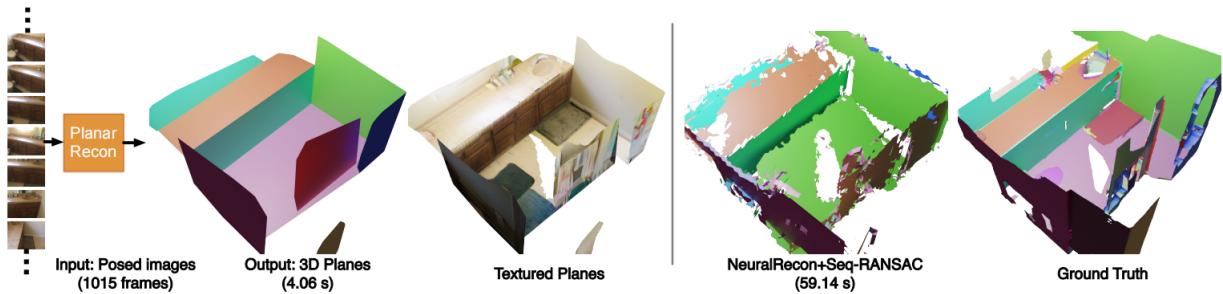


图 1. PlanarRecon 是一种采用数据驱动方式学习的方法，用于从含有位姿的单目视频中，全局性地、实时连贯地检测、跟踪和融合平面实例，最终重建场景的简化三维平面。经过 ScanNetV2 数据集 [4] 的训练，PlanarRecon 可以推广到室内场景的实时三维平面检测和重建。

随着三维成像技术和虚拟现实技术的快速发展，室内三维重建已经成为计算机视觉和图形学的研究热点。在建筑设计规划、游戏开发、无人驾驶、机器人导航等广泛的应用领域中，精确且高效的室内三维重建技术的需求日益增强。传统室内三维重建面临主要基于稠密三角面片作为基元对场景进行描述，这种表示方法存在计算存储密集、重建效率低、遮挡空洞、弱纹理区域重建易失败的问题，并且无法直接得到场景的结构信息，无法满足增强现实和虚拟现实（AR 和 VR）、室内建模和人机交互等下游关键任务的需求。

随着技术的发展和应用需求的增多，越来越多的学者开始致力于解决上述问题，研究快速、鲁棒的场景重建方法，以高度结构化特征对场景归纳和描述。平面提供了三维场景的紧凑表示和重要几何线索，被认为是最有潜力实现场景简化重建与结构理解的表示之一。以 AR 为例，要在 AR 效果和周围物理场景之间实现逼真和身临其境的交互，就需要对 AR 进行准确、一致和实时的 3D 平面检测。虽然最先进的视觉-惯性 SLAM 系统 [11, 13] 可以精确地跟踪相机姿态，并结合图像分割、点云平面拟合等方法实现三维平面重建，但由于检测质量低和计算要求高，基于图像的实时三维平面检测仍然是一个具有挑战性的问题。2022 年，Xie 等人首次提出一种基于学习的三维平面重建框架 Planrrecon [23]，如图 1 所示，该框架能够以已知姿势的单目视频作为输入，可以直接在三维空间中检测、跟踪和融合平面实例，实现了全局一致的平面检测和重建，与现有方法相比，实现了更稳健，速度更快的检测。

2 相关工作

2.1 基于多视立体几何的三维平面重建

基于已知位姿的多视图进行平面重建已有很多研究。早期的工作通常首先使用点 [2] 或线特征 [1] 执行稀疏 3D 重建，然后使用某些启发式对稀疏 3D 表示进行分组，进一步进行平面拟合。然而，这些方法的重建精度在很大程度上取决于手工制作的特征，并且对于光照变化和无纹理区域等其他因素并不鲁棒。部分工作将此问题作为图像分割任务。另一些工作基于马尔可夫随机场为每个像素分配一个平面假设 [5, 16]，采用随机扰动优化平面参数以最小化多视角光度一致性误差。其他人扩展了这个框架来处理非平面表面 [6]，并引入超像素分割来更好地处理无纹理区域 [3]。还有一些工作以单目视频作为输入，同时估计相机姿势并以 SLAM 方式重建平面。然而，这些作品通常假设世界由水平地面和一些垂直平面（例如立面或墙壁）组成 [25, 26]，和/或平面结构仅存在于低梯度区域 [3]。

2.2 基于学习的三维平面重建

在 Planrrecon [23] 之前，有一些研究旨在从一个或两个视图中恢复平面结构。有几项研究将从单视图重建平面视为实例分割问题 [9, 10, 19, 24, 27]，它们利用深度网络用于联合预测平面实例分割和平面参数，然后将预测的参数将分割掩模投射到三维中以进行平面重建。为了进一步提高重建精度，其他研究还检测并加强了平面实例之间的平面间关系 [14]，或利用全景输入分别对水平面和垂直面进行分割 [17]。在双视角情况下，问题变得更具挑战性，要生成统一的三维重建，模型必须正确关联各帧的平面实例。还有一些工作提出了学习每个平面实例的描述符，并通过某些优化算法进行匹配 [7, 15]，然而，目前这种方法在多视角情况下的通用性仍未得到证明。相反，Planrrecon 直接在三维空间中进行平面检测、跟踪和融合，这样可以减少匹配过程的模糊性，从而获得更高的重建精度和更一致的结果。

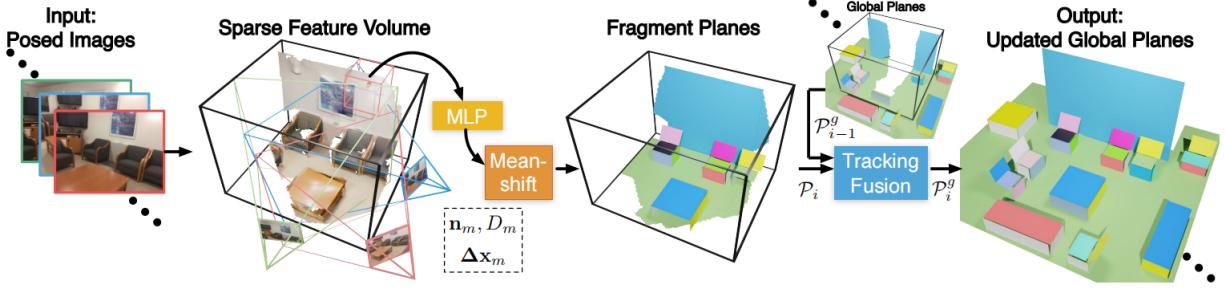


图 2. PlanarRecon 基于场景构造一个三维体素表示，首先将二维图像特征反向投影到一个片段绑定的体素 F_i 中，并以一种从粗到细的方法逐渐稀疏化该体素特征，通过一个 MLP 网络对预测每个体素的占用性、平面参数 $[n_m \ D_m]$ 及其与最近平面的距离 ΔX_m 。然后通过 Mean-shift 算法对这些混合几何原语 $[n_m \ x'm = xm + \Delta xm]$ 进行聚类形成平面实例 P_i 。跟踪和融合模块将匹配当前片段绑定的体素 F_i 的三维平面 P_i^g 和之前片段中的全局平面 P_{i-1}^g 。匹配的平面对将被融合细化，得到最终的三维平面重建

3 本文方法

3.1 本文方法概述

该论文提出了一个名为 PlanarRecon 的方法，如图1所示，该方法能够从含有位姿信息的单目视频中，全局性地、连贯地检测和重建三维平面。如图2所示，与以往从单幅图像中检测二维平面的工作不同，PlanarRecon 框架主要由两个部分组成，第一个组件是基于片段的平面检测，第二个组件是平面跟踪和融合。第一个组件利用神经网络学习场景的三维体素特征表示，为每个视频片段（由一组关键帧组成）增量检测三维平面。第二个组件融合多个片段中多个视图的观察结果和不同片段的时间信息，实现准确、连贯地重建具有低多边形几何结构的场景三维平面。实验表明，该论文所提出的方法在 ScanNetV2 [4] 数据集上实现了最先进的性能，同时具有实时性。

3.2 三维稀疏特征体素的构建

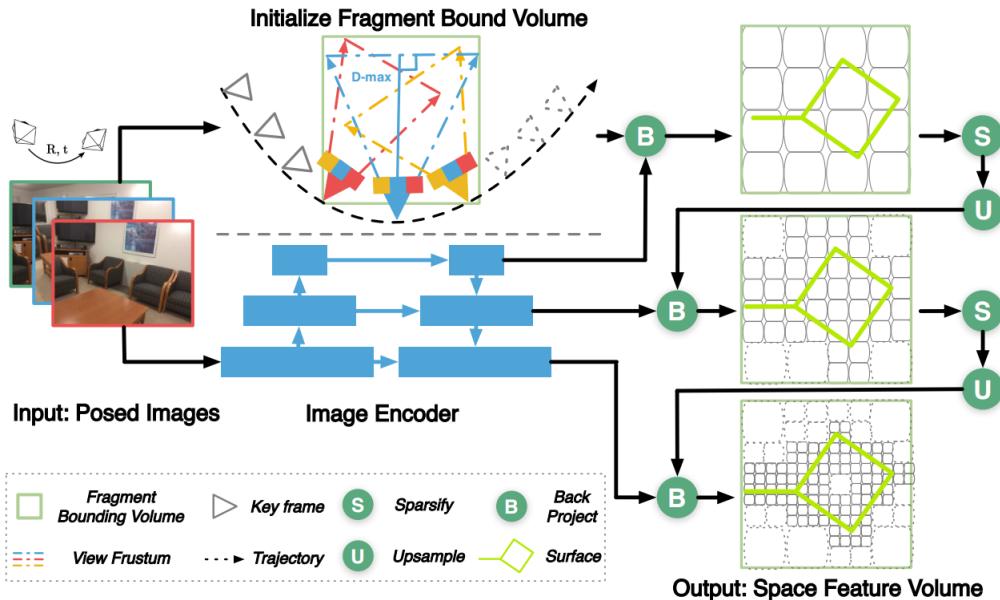


图 3. 三维稀疏体素的构建过程示意图。

给定一个视频输入，PlanarRecon 首先依次将其分割成多个不重叠的视频片段，对每一

个视频片段进行关键帧检测得到视频片段的关键帧集。如图??所示，PlanarRecon 首先利用 MnasNet 网络 [20] 对每一个关键帧进行特征提取，进一步将所有的特征通过相机位姿反投影到三维空间。三维空间被划分为一组离散体素，每个体素的特征表示是其在不同关键帧中观察到的像素表示的平均值。PlanarRecon 按照计算图像表征时的网络层次结构，以从粗到细的金字塔方式构建三维稀疏体素特征，其中三维空间被逐渐划分为更细的体素。在每个金字塔中，PlanarRecon 对每个体素进行占位分类，只有被占据的体素才会在下一级金字塔中得到进一步处理，最终会在最细的层次上得到一组被占据的体素，用于下一步的平面检测。

3.3 基于三维稀疏特征体素的平面检测

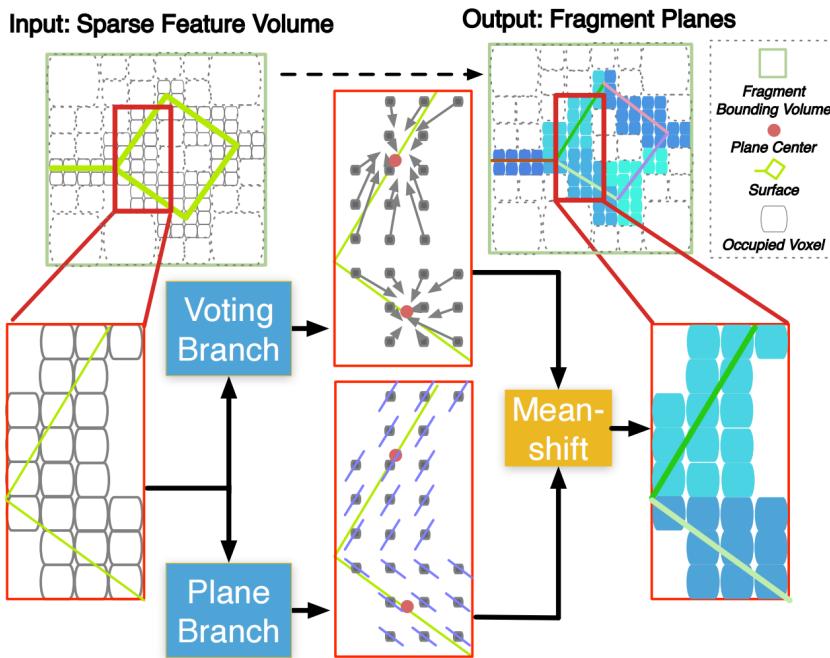


图 4. 基于片段的平面检测过程示例图。彩色体素网格表示它们所属不同的平面实例。

PlanarRecon 进一步通过一个多层次感知器 (MLP, Multilayer Perceptron) [21] 对每个体素的占用性进行预测。如图4所示，对于每个被占据的体素，PlanarRecon 利用两个并行分支进行处理，其中平面分支用于预测平面参数，一个投票分支用于预测体素到所属平面中心点的位移。PlanarRecon 基于这些参数对所占体素进行 Mean-Shit 聚类，将具有相似平面参数和位移的体素组合在一起，以获得三维平面检测结果。

3.3.1 平面分支

特别地，PlanarRecon 并不直接预测平面参数，而是通过给定一组法向锚点 (6 个，其中两个与地面平行，其余 4 个与地面垂直)，在训练过程中，网络只需要预测哪一个法向锚点最接近真实值，以该法向锚点作为输出，并与真实值计算损失回归残差。进一步，PlanarRecon 预测体素中心 \mathbf{x}_m 到它所属平面的距离 D_m ，那么我们将该体素移动到它预测所属平面上，如式1所示，利用点法式方程我们可以计算得到平面偏置 d_m 。

$$d_m = -\langle \tilde{\mathbf{x}}_m, \mathbf{n}_m \rangle, \quad \tilde{\mathbf{x}}_m = \mathbf{x}_m + D_m \mathbf{n}_m. \quad (1)$$

其中 $\langle \cdot, \cdot \rangle$ 表示两个向量的内积

3.3.2 投票分支

经过平面分支的参数预测，我们已经可以根据平面参数对体素进行聚类，但是，这种方法可能无法分离具有相似平面参数的两个平面，即使它们在空间中的位置不同。因此，文中加入了投票分支，用于预测体素中心 x_m 到它所属平面的质心距离 ΔX_m ，则每个体素多了一个聚类可用的变量，即所属平面质心的位置 $\mathbf{x}'_m = \mathbf{x}_m + \Delta \mathbf{x}_m$ ，以此可以对具有相似平面参数的多个平面进行区分。

3.3.3 基于 MeanShift 聚类得到平面实例

经过平面分支和投票分支后，每一个三维体素被赋予了一些所属平面特征的变量，包括所属平面的参数、所属平面质心的位置，因此最终的平面实例是根据这些平面参数对体素进行 MeanShift 聚类得到的，最终每个平面实例的参数由 MeanShift 聚类中心给出。

3.4 平面的匹配、跟踪与融合

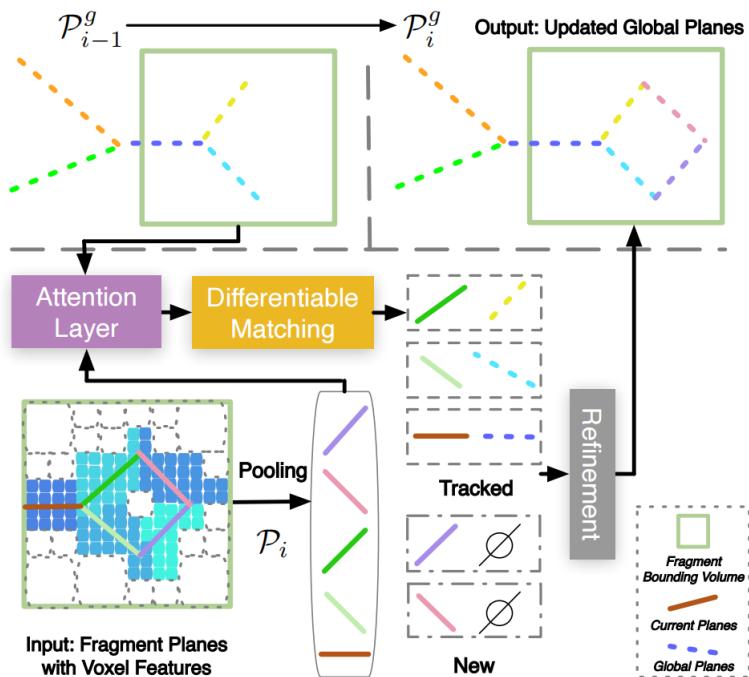


图 5. 三维平面的跟踪与匹配过程示意图。彩色线段表示不同的平面实例及其描述符。

经过基于片段的平面检测模块后，算法可以得到每一个视频片段的平面实例。为了实现实时的平面重建，PlanrRecon 融合对以前所有片段与当前片段的平面检测结果来维护平面的全局重建。当一个新的片段被计算得到平面实例参数时，PlanrRecon 首先利用式2计算当前检测平面与全局重建平面之间的相似性，

$$\mathbf{S}(m, n) = \langle \mathbf{h}_{i-1,m}^g, \mathbf{h}_{i,n} \rangle, \quad (2)$$

其中 m, n 表示两个平面实例， $\mathbf{h}_{i-1,m}^g$ 和 $\mathbf{h}_{i,n}$ 相应的经过网络提取的增强特征向量

在得到两两平面的相似性度量值后，PlanarRecon 利用可微匈牙利匹配算法 [8] 对平面的对应关系进行优化，目标是最大化当前平面集与全局平面集的相似性度量值之和 \mathbf{M}^* 。

$$\mathbf{M}^* = \arg \min_{\mathbf{M}} \sum_{m,n} \mathbf{S}(m, n) \mathbf{M}(m, n). \quad (3)$$

得到匹配关系后，PlanarRecon 利用 Attention-GRU(注意力机制融合门控循环单元) [22] 来自动学习匹配平面的融合，根据式 4 得到融合细化的平面，最终对全局重建进行更新。

$$P_{i,m}^g = \frac{\gamma P_{i-1,m}^g + P_{i,n}}{\gamma + 1}, \quad (4)$$

其中， γ 是控制 GRU 更新速度的参数。GRU 将匹配平面的特征进行融合，并通过 MLP 层来预测 γ 。 $P_{i,m}^g$ 可以是表面法线 $\mathbf{n}_{i,m}^g$ ，也可以是平面偏移 $d_{i,m}^g$ ，或者平面的特征向量 $\mathbf{f}_{i,m}^g$ 。

3.5 损失函数定义

损失函数由三大部分组成，第一部分是占用性预测损失，第二部分是基于片段的平面检测损失，第三部分是平面跟踪匹配损失。

占用性预测损失 L_o 被定义为每个体素的预测占用值 (0/1) 与真实占用值的交叉熵损失。

基于片段的平面检测损失 L_d 包括 4 个部分，监督锚点法线选择的学习的交叉熵损失 L_{ac} ，监督锚点法向残差向量的 smooth-L1 损失 L_{ar} 、平面偏移距离回归的 smooth-L1 损失 L_{or} 以及体素到所属平面质心的距离的 L1 损失 L_{dc} 。

$$L_d = \alpha_1 L_{ac} + \alpha_2 L_{ar} + \alpha_3 L_{or} + \alpha_4 L_{dc}. \quad (5)$$

平面跟踪匹配损失 L_m 被定义为匹配矩阵计算函数 \mathbf{M}^* 的负最大对数似然，并对 \mathcal{P}_{i-1}^g 和 \mathcal{P}_i 中未匹配的平面进行惩罚。

$$L_m = - \sum_{(i,j) \in \mathcal{M}} \log \mathbf{M}_{i,j}^* - \sum_{i \in \mathcal{I}} \log \mathbf{M}_{i,N+1}^* - \sum_{j \in \mathcal{J}} \log \mathbf{M}_{M+1,j}^*, \quad (6)$$

其中 M 和 N 是 \mathcal{P}_{i-1}^g 和 \mathcal{P}_i 的数量， \mathcal{I} 和 \mathcal{J} 是 \mathcal{P}_{i-1}^g 和 \mathcal{P}_i 中未被匹配的平面集。

最终的损失函数被定义为三个损失的联合损失

$$L = \beta o L_o + \beta d L_d + \beta m L_m. \quad (7)$$

4 复现细节

4.1 与已有开源代码对比

本文复现了 PlanarRecon 中基于多视图的三维稀疏体素特征的构建和基于三维稀疏体素特征的平面检测部分，平面的匹配跟踪与融合部分、Baseline 使用了官方所发布的源代码 [23]。另外，PlanarRecon 只能恢复平面的参数，但重建面片的法向存在不合理的情况（例如地面的法向应该朝上而不是朝下），本文通过引入相机位姿对恢复出的面片的法向进行了检查修复，使其具有合理的法向。

4.2 实验设置

实验所使用的数据集是 ScanNetv2 [4]，ScanNetv2 数据集包含由移动设备从 1613 个室内场景中拍摄的 RGB-D 视频，训练集、验证集和测试集的划分遵循 ScanNetv2 官方的划分。由于 ScanNetv2 没有提供平面标签，根据原文的方法，遵循 PlaneRCNN [9]，使用在网格面片上采样点并拟合平面的方法生成 3D 平面标签。由于训练数据较多，为了节约计算资源，我们基于官方提供的预训练参数开始训练，继续迭代 10 个 epoch。

重建平面与真实平面标签的量化对比分析是通过计算重建平面采样点云与真实平面标签的采样点云的 Recall, Prec, F-score 指标 [12] 完成的。

实验环境及参数如表1所示。

表 1. 实验环境信息

项目	详细信息
操作系统	Ubuntu 20.04 LTS
CPU	Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20GHz
GPU	NVIDIA GeForce GTX 3090 × 8
内存	252GB
依赖项	Python 3.7, Pytorch 1.7, Numpy 1.16.4, etc.

4.3 创新点

本文针对 PlanarRecon 只能恢复平面的参数，但重建面片的法向存在不合理的情况（例如地面的法向应该朝上而不是朝下）进行改进，通过引入相机位姿对恢复出的面片的法向进行了检查修复，使其具有合理的法向。具体来说，基于几何和光学的基本原理，光线从物体表面反射到相机的角度应该等于相机视线与物体表面法线的角度，在相机视角小于 180° 的情况下，这个角度应该小于 90° ，也就是说，当我们从相机的视点观察一个物体表面，并反射到相机的光线 r 应该大致沿着与表面法线 n 相反（夹角 θ 小于 90° ）的方向传播，由于我们有位姿，那么就相当于知道了 r ，因此我们可以根据这一原理对平面的法向 n 进行检查和纠正。

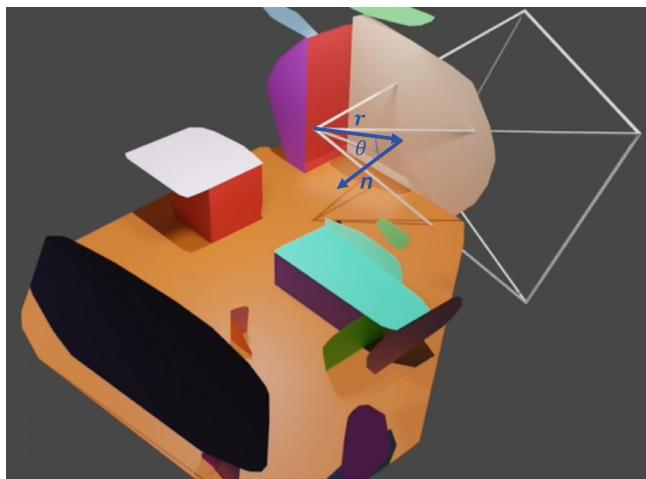


图 6. 示意图：从相机的视点观察一个物体表面，并反射到相机的光线 r 应该大致沿着与表面法线 n 相反（夹角 θ 小于 90° ）的方向传播

5 实验结果分析

5.1 测试结果分析

实验测试主要基于 ScanNet 数据集展开。图7展示了利用 PlanarRecon 对 ScanNetV2 部分测试场景进行重建的结果以及误差可视分析，可以看到 PlanarRecon 较好地完成了场景的抽象表示。表2展示了复现的代码的测试结果与 PlanarRecon 原论文中的测试结果进行的对比，可以看到重建场景评估结果基本一致，但是 PlanarRecon 在本文测试环境中的时间消耗较长，这可能是本文的测试环境与论文中的测试环境不同所致。

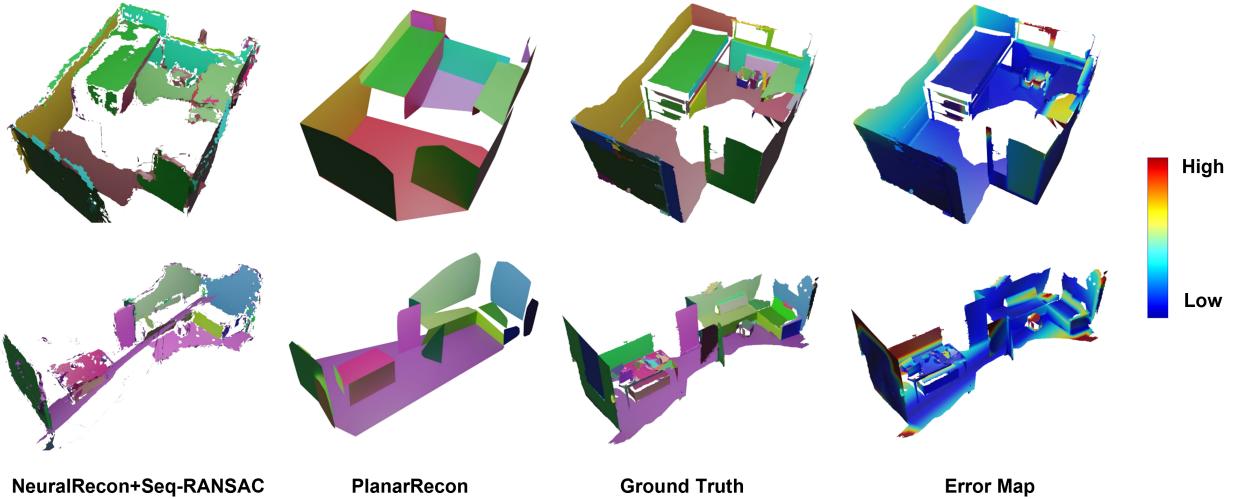


图 7. 部分 ScanNet 场景测试结果

表 2. 在 ScanNetv2 测试集的评估结果

	Method	Recall↑	Prec↑	F-score↑	Time(ms/frame)↓
Reproduced	NeuralRecon [18]+Seq-RANSAC	0.292	0.303	0.297	723
	PlanarRecon [23]	0.376	0.407	0.391	81
Original	NeuralRecon [18]+Seq-RANSAC	0.296	0.306	0.301	586
	PlanarRecon [23]	0.372	0.412	0.389	40

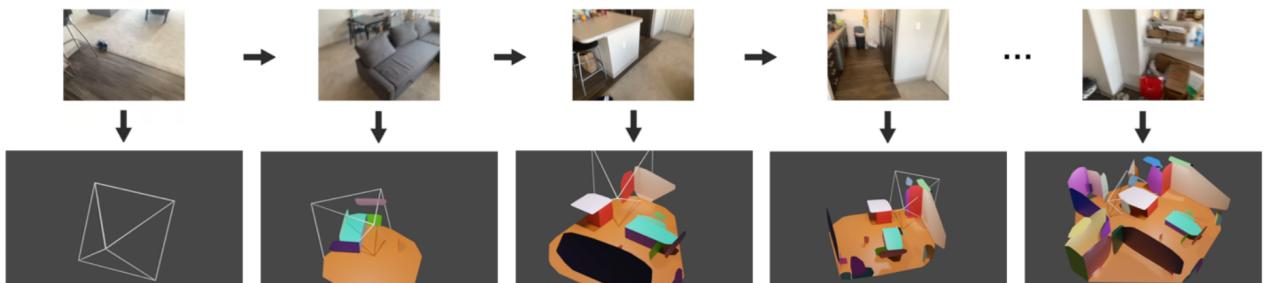


图 8. 从含有位姿信息的单目视频中，全局性地、实时连贯地检测、跟踪和融合平面实例，重建场景的简化三维平面描述

图8可视化了使用 PlanarRecon 从含有位姿信息的单目视频中，全局性地、实时连贯地检测、跟踪和融合平面实例，重建场景的简化三维平面描述的过程。

5.2 面片法向修正对比分析

针对 PlanarRecon 只能恢复平面的参数，但重建面片的法向存在不合理的情况（例如地面的法向应该朝上而不是朝下）进行改进，本文通过引入相机位姿对恢复出的面片的法向进行了检查修复，使其具有合理的法向。图9展示了，修正前后的对比，可以观察到修正后的法向更加合理。

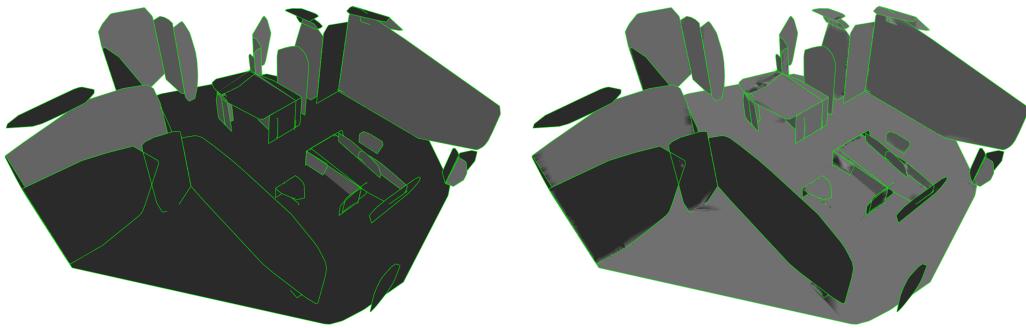


图 9. 法向修复前后对比结果（左为修复前，右为修复后）：亮色表示法向朝向观测点，暗色表示法向背向观测点

6 延伸工作

PlanarRecon 能够重建室内三维场景的抽象（低多边形）表示，但是它无法恢复平面与平面之间的连接关系，基于这个缺陷出发，我们拟定了一个新的，更加具有挑战性，且创新性也更高的课题，为了验证这个任务的可行性，我们首先在 CAD 模型上进行重建测试，未来计划推广到更泛化的重建任务上。如图10所示，我们希望以给定位姿的图片作为输入，重建出对象的低多边形平面及其连接关系。图11展示了目前的部分结果，我们的算法可以以带位姿的多视角图像作为输入，重建出多个局部平面模型，进一步我们组合这些局部平面重建并优化得到全局平面重建结果。

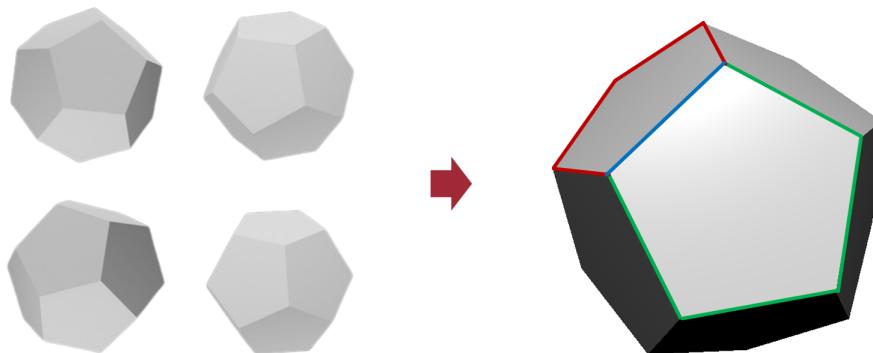


图 10. 给定位姿的图片作为输入，重建出对象的低多边形平面及其连接关系

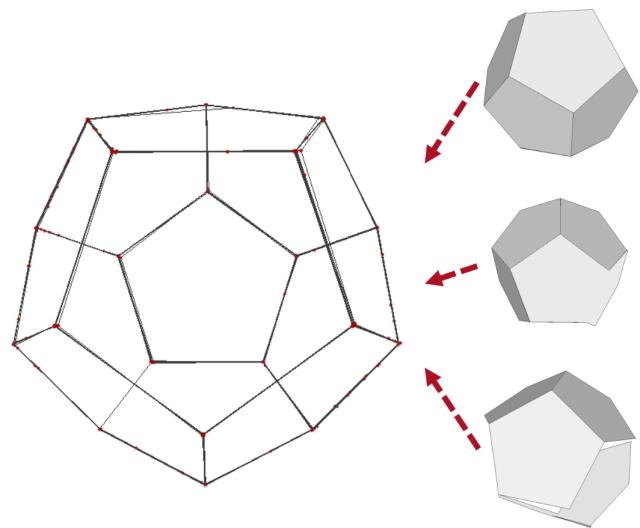


图 11. 算法部分结果展示

7 总结与展望

PlanarRecon 是一种用于从已知位姿的单目视频中全局连贯检测和重建三维平面的新型框架，它首次将这个任务实现为一个完整网络并进行端到端学习，同时为场景的抽象结构表示重建提供了一种新的思路。本文完成了该方法的复现与基本的测试实验，并针对它无法恢复平面的法向的问题进行改进，加入了相机位姿辅助修正法向的模块。

虽然 PlanarRecon 在 ScanNetv2 上取得相对不错的性能，但是对计算资源要求仍然相对较高，后续可以进一步改进提高性能。此外，目前三维平面的真实标签仍然是通过拟合点云得到的，室外场景的真实标签获取将会更具有挑战性，未来可以探索此方法的泛化性，在室外场景上测试和改进，还可以针对基于多视角图片的无监督学习的三维平面重建进行探索和研究。

从 PlanarRecon 中延伸出的第六小节所介绍的任务，以给定位姿的图片作为输入，重建出对象的低多边形平面及其连接关系，也是一个非常有前景的课题，目前也正在研究中。

参考文献

- [1] Caroline Baillard and Andrew Zisserman. Automatic reconstruction of piecewise planar models from multiple views. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 2, pages 559–565. IEEE, 1999.
- [2] Adrien Bartoli. A random sampling strategy for piecewise planar scene segmentation. *Computer Vision and Image Understanding*, 105(1):42–59, 2007.
- [3] Alejo Concha and Javier Civera. Dpptam: Dense piecewise planar tracking and mapping from a monocular sequence. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5686–5693. IEEE, 2015.
- [4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.
- [5] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Manhattan-world stereo. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1422–1429. IEEE, 2009.
- [6] David Gallup, Jan-Michael Frahm, and Marc Pollefeys. Piecewise planar and non-planar stereo for urban scene reconstruction. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1418–1425. IEEE, 2010.
- [7] Linyi Jin, Shengyi Qian, Andrew Owens, and David F Fouhey. Planar surface reconstruction from sparse views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12991–13000, 2021.
- [8] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.
- [9] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4450–4459, 2019.
- [10] Chen Liu, Jimei Yang, Duygu Ceylan, Ersin Yumer, and Yasutaka Furukawa. Planenet: Piece-wise planar reconstruction from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2579–2588, 2018.
- [11] Raúl Mur-Artal and Juan D. Tardós. Visual-inertial monocular slam with map reuse. *IEEE Robotics and Automation Letters*, 2(2):796–803, 2017.

- [12] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 414–431. Springer, 2020.
- [13] Zainab Oufqir, Abdellatif El Abderrahmani, and Khalid Satori. Arkit and arcore in serve to augmented reality. In *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–7, 2020.
- [14] Yiming Qian and Yasutaka Furukawa. Learning pairwise inter-plane relations for piecewise planar reconstruction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 330–345. Springer, 2020.
- [15] Yifei Shi, Kai Xu, Matthias Nießner, Szymon Rusinkiewicz, and Thomas Funkhouser. Planematch: Patch coplanarity prediction for robust rgb-d reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–766, 2018.
- [16] Sudipta Sinha, Drew Steedly, and Rick Szeliski. Piecewise planar stereo for image-based rendering. In *2009 International Conference on Computer Vision*, pages 1881–1888, 2009.
- [17] Cheng Sun, Chi-Wei Hsiao, Ning-Hsu Wang, Min Sun, and Hwann-Tzong Chen. Indoor panorama planar 3d reconstruction via divide and conquer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11338–11347, 2021.
- [18] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021.
- [19] Bin Tan, Nan Xue, Song Bai, Tianfu Wu, and Gui-Song Xia. Planetr: Structure-guided transformers for 3d plane recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4186–4195, 2021.
- [20] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2820–2828, 2019.
- [21] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

- [22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [23] Yiming Xie, Matheus Gadelha, Fengting Yang, Xiaowei Zhou, and Huaizu Jiang. Planar-recon: Real-time 3d plane detection and reconstruction from posed monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6219–6228, 2022.
- [24] Fengting Yang and Zihan Zhou. Recovering 3d planes from a single image via convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.
- [25] Shichao Yang and Sebastian Scherer. Monocular object and plane slam in structured environments. *IEEE Robotics and Automation Letters*, 4(4):3145–3152, 2019.
- [26] Shichao Yang, Yu Song, Michael Kaess, and Sebastian Scherer. Pop-up slam: Semantic monocular plane slam for low-texture environments. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1222–1229. IEEE, 2016.
- [27] Zehao Yu, Jia Zheng, Dongze Lian, Zihan Zhou, and Shenghua Gao. Single-image piecewise planar 3d reconstruction via associative embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1029–1037, 2019.