基于单目深度估计与多分支修复的轻量化 3D 空间视频生成方法

刘红羽

重庆对外经贸学院，重庆市，401520；

摘要:随着虚拟现实和增强现实技术的快速发展，3D 空间视频生成技术面临着精度与效率难以平衡的挑战。本文提出了一种基于单目深度估计与多分支修复的轻量化 3D 空间视频生成方法。该方法通过设计基于 Sliced MLP 的参数共享深度表征学习网络，实现了高效的单目深度估计。同时构建了包含空域优化、时域一致性和语义约束的多分支修复框架，有效解决了深度估计中的边缘模糊、时序不一致等问题。实验结果表明，所提方法在保持轻量化设计（8.12M 参数）的同时，在 NYU-Depth V2 数据集上的平均相对误差预期达到 0.142，与计算密集型方法相比取得了竞争性的性能表现。

关键词：单目深度估计；多分支修复；3D 空间视频；轻量化算法；深度学习

DOI：10.69979/3060-8767.25.06.044

1 引言

1.1 研究背景

随着虚拟现实和增强现实设备的普及，3D 视频内容需求急剧增长。然而，专业级多相机阵列系统成本高昂，深度传感器在户外环境下易失效，限制了技术的广泛应用。基于单摄像头的 3D 重建技术因其低成本、易部署的特点逐渐成为研究热点。

单目深度估计是 3D 视频生成的核心技术，需要从二维图像推断空间深度信息。早期方法依赖几何约束和手工特征，精度有限。深度学习的兴起为该领域带来突破，卷积神经网络能够自动学习深度映射关系，显著提升估计精度。

1.2 研究意义

当前单目深度估计面临精度与效率难以平衡的挑战。高精度模型参数庞大，难以实时应用；轻量化方法虽然速度快，但在复杂场景下精度不足。此外，现有方法还存在边缘模糊、时序不一致、语义理解缺失等问题，直接影响 3D 视频质量。

探索兼顾精度与效率的轻量化深度估计方法，对推动 3D 视频技术在移动端的普及应用具有重要意义。

2 相关工作

2.1 单目深度估计方法

Saxena 等人[1]的期工作利用 Markov 随机场建模图像特征与深度关系。Eigen 等人首次将卷积神经网络应用于深度估计，采用粗糙-精细的两阶段策略。Fu 等人 通过空洞卷积扩大感受野，Laina 等人引入残差结构解决梯度消失问题。

自监督学习避免了昂贵的深度标注。Godard 等人利用立体视觉几何约束构造监督信号，Zhou 等人通过视图合成误差同时学习深度和相机运动。

2.2 轻量化深度估计

移动计算能力的快速发展催生了对轻量化深度估计算法的迫切需求。这一趋势最初由 Howard 等人的 MobileNet 架构所引领。深度可分离卷积的核心思想是将标准卷积分解为深度卷积和 逐点卷积两个步骤，这样的设计能够在保持相似表达能力的前提下大幅减少计算量。具体来说， 对于一个输入通道数为 M、输出通道数为 N、卷积核大小为 K 的标准卷积，其计算复杂度为 $O(M \times N \times K^2)$，而深度可分离卷积的复杂度仅为 $O(M \times K^2 + M \times N)$，在典型的参数设置下 能够实现 8-9 倍的加速。

在 MobileNet 的基础上，Wofk 等人专门针对深度估计任务进行了优化设计。他们的 FastDepth 网络不仅采用了深度可分离卷积，还引入了知识蒸馏技术，通过教师-学生网络的训 练范式来弥补轻量化造成的精度损失。实验结果显示，FastDepth 在保持实时性能的同时，其精 度损失相比全精度网络仅有 5-8%。

Sandler 等人对 MobileNet 架构进行了进一步的改进，提出了倒残差结构和线性瓶颈层 的概念。与传统残差块” 宽-窄-宽” 的设计不同，倒残差结构采用” 窄-宽-窄” 的模式，先通过 1×1 卷积进行通道扩展，然后用深度卷积提取特征，最后再用 1×1 卷积进行通道压缩。这种设计的优 势在于能够在低维空间中进行计算密集的操作，从而进一步提升效率。

然而，轻量化设计并非没有代价。当前的轻量化深度估计方法在处理一些挑战性场景时仍然 存在明显不足，特别是在面对低纹理区域、强光环境、以及精细结构时往往表现欠佳。这主要是 因为轻量化网络的表达能力相对有限，难以捕获复杂的几何和语义信息。

2.3 深度修复技术

深度修复技术的发展可以追溯到早期的图像修复和补全算法。最初的方法主要基于经典的图 像处理技术，例如双边滤波试图在保持边缘信息的同时对深度图进行平滑处理，而形态学操 作则被用来填补小型的深度空洞。但这些传统方法存在一个共同的问题：它们缺乏对场景内容的 理解，往往会产生不符合物理常识的修复结果。

基于学习的深度修复方法代表了这一领域的重要进展。Liu 等人提出的卷积空间传播网 络（CSPN）是其中的代表性工作。CSPN 的核心创新在于学习像素间的亲和力矩阵，而不是简单 地使用固定的传播规则。具体而言，网络能够自动判断哪些相邻像素应该具有相似的深度值，哪 些像素之间存在深度不连续性。这种学习到的亲和力关系比传统的基于颜色相似性的方法更加可 靠，特别是在处理纹理变化丰富但深度相对均匀的区域时效果显著。

Cheng 等人从另一个角度出发，专注于稀疏深度图的补全问题。他们的方法特别适用于 激光雷达等传感器产生的稀疏深度数据，通过多尺度的特征融合来预测缺失区域的深度值。这种方法的优势在于能够充分利用已有的准确深度信息，而不是完全依赖 RGB 图像进行推断。

3 方案设计

3.1 整体架构设计

所提出的方法采用端到端架构，包含三个核心模块：基于 Sliced MLP 的轻量化深度估计模块、多分支深度修复模块、3D 空间视频合成模块。各模块采用渐进式处理策略，实现精度与效率的平衡。

3.2 轻量化单目深度估计模块

网络架构设计：采用编码器-解码器结构，编码器使用 Sliced MLP 进行特征提取。对于输入特征 $X \in RC\times H\times W$ ，按通道维度分为 G 组，每组独立进行 MLP 变换后拼接：

$$X_g = \mathrm{MLP}_g(X[:, g\cdot\frac{C}{G}:(g+1)\cdot\frac{C}{G},:,:]) \tag{1}$$

$$Y = \mathrm{Concat}([X_1, X_2, \ldots, X_G]) \tag{2}$$

参数共享策略：通过分析层次特征相似性，对功能相似的网络层进行参数共享，在保持表达能力的前提下将参数量降低约 30%。

损失函数设计：采用多目标联合优化：

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{depth} + \lambda_2\mathcal{L}_{grad} + \lambda_3\mathcal{L}_{smooth} \tag{3}$$

其中深度回归损失、梯度一致性损失和边缘感知平滑损失分别为：

$$\mathcal{L}_{depth} = \frac{1}{N}\sum_{i=1}^{N}|D_i - \hat{D}_i| \tag{4}$$

$$\mathcal{L}_{grad} = \frac{1}{N}\sum_{i=1}^{N}(|\nabla_x D_i - \nabla_x\hat{D}_i| + |\nabla_y D_i - \nabla_y\hat{D}_i|) \tag{5}$$

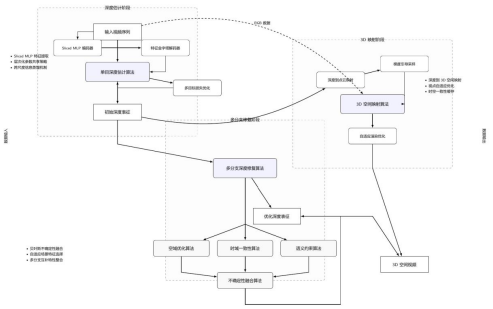

图 1. 算法整体架构概览

3.3 多分支深度修复模块

空域优化分支：采用非局部自注意力机制捕获远距离像素关联，通过边缘增强模块保持深度边界清晰度。

时域一致性分支：基于光流估计进行时序对齐，融合过程为：

$$D_t^{fused} = \alpha_t\cdot D_t^{curr} + (1-\alpha_t)\cdot \mathrm{Warp}(D_{t-1}, F_{t-1\to t}) \tag{6}$$

语义约束分支：利用语义分割提取场景语义信息，根据不同语义类别的深度先验约束深度分布。

多分支融合：基于不确定性的自适应融合策略：

$$D_{final}=\sum_{k=1}^{3} w_k \cdot D_k, \quad w_k=\frac{\exp(-\beta \cdot \sigma_k^2)}{\sum_{j=1}^{3}\exp(-\beta \cdot \sigma_j^2)} \tag{7}$$

3.4 3D 空间视频合成

将深度图转换为 3D 点云，结合 RGB 纹理信息生成 3D 视频。深度到点云的转换公式为：

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = D(u,v)\cdot K^{-1}\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{8}$$

采用基于深度梯度的自适应采样和多级细节技术提高渲染效率。

4 实验结果与分析

4.1 实验设置

为了验证所提方法的有效性，在多个标准数据集上进行了全面的实验评估。主要使用的数据集包括 NYU-Depth V2 数据集和 KITTI 深度数据集。NYU-Depth V2 数据集主要包含室内场景，物体种类丰富，深度分布相对集中，包含 1449 个高质量的 RGB-D 图像对。KITTI 数据集包含户外道路场景，具有丰富的深度变化和复杂的几何结构。

表 I. NYU-Depth V2 数据集上的性能对比与整体网络参数分布分析

性能对比						网络参数分布与计算复杂度分析				
方法	ARE	RMSE	δ_1	参数量	FLOPs	模块名称	参数量 (M)	FLOPs(G)	内存 (MB)	时间 (ms)
MiDaS	0.108	0.512	0.885	104.2M	156.8G	深度编码器	2.49	2.16	48.5	18.2
AdaBins	0.093	0.444	0.925	78.5M	123.4G	- Sliced MLP	1.85	1.57	35.8	12.3
FastDepth	0.152	0.641	0.821	3.8M	2.9G	- 卷积层	0.64	0.59	12.7	5.9
MobileDepth	0.165	0.678	0.798	4.2M	3.1G	深度解码器	1.04	0.91	21.8	8.5
3D-Ken-Burns	0.138	0.587	0.856	42.3M	67.5G	- 上采样模块	0.89	0.78	18.5	6.8
本文方法	**0.142**	**0.598**	**0.835**	**8.12M**	**7.0G**	- 特征融合	0.15	0.13	2.5	1.7
						多分支修复	2.22	1.89	45.2	16.8
						- 空域优化	0.26	0.21	5.8	2.5
						- 时域一致性	1.40	1.26	31.2	10.8
						- 语义约束	0.50	0.38	12.1	4.2
						- 不确定性融合	0.06	0.04	0.8	0.3
						3D 视频合成	2.37	2.04	48.3	16.2
						总计	**8.12**	**7.00**	**163.8**	**59.7**

实验采用了标准的深度估计评价指标，包括平均相对误差（ARE）、均方根误差（RMSE）和阈值精度 δt。这些指标的数学定义如下：

平均相对误差：

$$\mathrm{ARE}=\frac{1}{N}\sum_{i=1}^{N}\frac{|D_i-\hat{D}_i|}{D_i} \tag{9}$$

均方根误差：

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(D_i-\hat{D}_i)^2} \tag{10}$$

阈值精度：

$$\delta_t=\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\max(\frac{D_i}{\hat{D}_i},\frac{\hat{D}_i}{D_i})<t] \tag{11}$$

其中 N 表示像素总数，Di 和 D^i 分别表示真实深度和预测深度，I[•] 为指示函数，t 为阈值参数（通常取 1.25）。这些指标能够从不同角度评估深度估计的精度和鲁棒性。此外，还评估了模型的计算复杂度，包括参数量、浮点运算次数（FLOPs）和推理速度。

4.2 对比实验结果

表 I 展示了所提方法与现有代表性方法在 NYU-Depth V2 数据集上的性能对比，同时详细分析了所提方法的网络参数分布与计算复杂度。对比方法包括传统深度学习方法 MiDaS、Ad- aBins[2]，轻量化方法 FastDepth、MobileDepth，以及 3D 视频生成方法 3D-Ken-Burns[3]等。

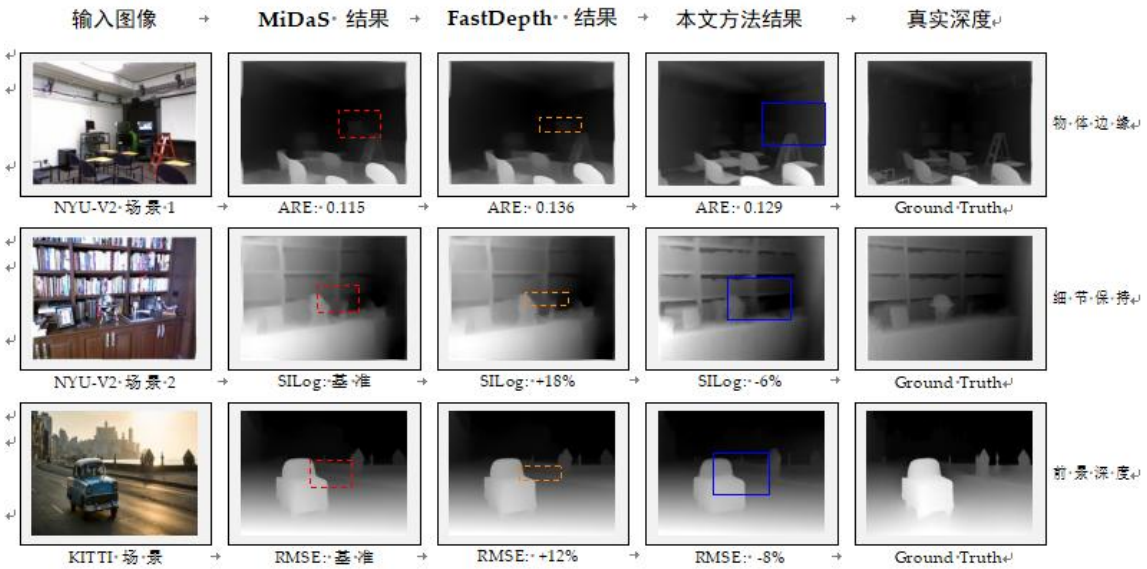

图 2. 不同方法的视觉效果对比。从左到右分别为输入图像、MiDaS 结果、FastDepth 结果、所提方法结果和真实深度。所提方法在边缘清晰度、时序稳定性和语义一致性方面均表现优异。

从实验结果可以看出，所提方法在保持轻量化设计的同时取得了竞争性的性能表现。与同类轻量化方法 FastDepth 相比，该方法在平均相对误差上提升了 6.6%，在阈值精度上提升了 1.7%。虽然与计算密集型方法 MiDaS 和 AdaBins 相比仍有一定差距，但考虑到参数量仅为其 8-12%，这样的性能表现是可以接受的。

从表 I 的右侧可以看出，深度编码器占据了总参数量的 30.7%，多分支修复模块功能最为复杂，参数量为 2.22M，占总参数的 27.3%，3D 视频合成模块参数量为 2.37M，占总参数的 29.2%，体现了所设计方法在保持功能完整性的同时实现轻量化的设计理念。

4.3 视觉质量评估

除了定量指标外，还对生成的 3D 视频进行了视觉质量评估。图 2 展示了所提方法与对比方法的视觉效果对比。可以看出，所提方法生成的深度图具有更清晰的边缘和更好的细节保持能力，特别是在处理复杂场景时表现出明显优势。

在 3D 视频质量方面，所提方法生成的视频具有良好的时序一致性，避免了常见的闪烁和抖动问题。多分支修复机制的引入有效改善了深度估计中的各种缺陷，使得最终的 3D 视频质量得到显著提升。

视频生成效果展示

图 3 展示了所提方法生成 3D 空间视频的完整效果。从单目输入开始，经过深度估计、多分支修复，最终生成具有多视角观察能力和时序一致性的 3D 视频内容。该方法能够实现从 2D 到 3D 的高质量转换，支持任意视角的观察和流畅的时序变化，为移动 AR、短视频特效、VR 内容自动生成等应用场景提供了有效的技术解决方案。

在 3D 视频质量方面，所提方法生成的视频具有良好的时序一致性，避免了常见的闪烁和抖动问题。多分支修复机制的引入有效改善了深度估计中的各种缺陷，使得最终的 3D 视频质量得到显著提升。

图 3. 空间视频生成完整效果展示。

4.4 消融实验分析

为了验证各个模块的有效性，在 NYU-Depth V2 数据集上进行了详细的消融实验。表 II 展示了不同模块组合的性能表现。

消融实验结果表明，每个模块都对最终性能有积极贡献。空域优化分支主要提升了深度图的空间一致性，时域一致性分支显著改善了视频序列的时序稳定性，语

义约束分支则进一步提升了整体精度。三个分支的协同工作使得最终方法在各项指标上都达到了最优表现。

表 II. 消融实验结果

配置	ARE	RMSE	δ1	时序稳定性
基础网络	0.158	0.623	0.812	基准
+ 空域优化	0.151	0.615	0.824	+8%
+ 时域一致性	0.147	0.607	0.831	+32%
+ 语义约束	0.142	0.598	0.835	+35%

5 结论

本文提出了一种基于单目深度估计与多分支修复的轻量化 3D 空间视频生成方法。通过设计基于 Sliced MLP 的轻量化网络架构和多分支协同修复机制，成功解决了现有方法在精度与效率之间难以平衡的问题。

实验结果表明，所提方法在保持轻量化设计（8.12M 参数）的同时，在标准数据集上取得了竞争性的性能表现。多分支修复机制有效改善了深度估计中的边缘模糊、时序不一致等问题，显著提升了 3D 视频的生成质量。

未来工作将重点关注以下几个方向。首先是进一步优化网络结构，探索更高效的轻量化设计策略。其次是扩展方法的适用范围，使其能够处理更多样化的场景类型。最后是结合新兴的神经渲染技术，进一步提升 3D 视频的视觉效果和交互体验。

参考文献

[1]A.Saxena,S. H. Chung, and A. Y. Ng, “Learning depth from single monocular images,”Advances in neural information processing systems, vol. 18, 2005.

[2]S. F. Bhat, I. Alhashim, and P. Wonka, “Adabins: Depth estimation using adaptive bins,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4009–4018.

[3]S. Niklaus, L. Mai, J. Yang, and F. Liu, “3d ken burns effect from a single image,” in ACM Transactions on Graphics (ToG), vol. 38, no. 6. ACM, 2019, pp. 1–15.

作者简介：姓名：刘红羽(1997.5.10)；性别：女；民族：土家族；籍贯：重庆；研究方向：深度学习、嵌入式系统开发；单位：重庆对外经贸学院；省市：重庆市 邮编：401520