arXiv:2303.04508v3 [cs.CV] 7 Oct 2024

InFusionSurf：通过逐帧内参细化和 TSDF 融合先验学习优化神经 RGB-D 表面重建

Seunghwan Lee、*Gwanmo Park*、*Hyewon Son*、*Jiwon Ryu*、*Han Joo Chae*

ROKIT Healthcare, Inc.

摘要

我们提出了 InFusionSurf，这是一种创新方法，可增强利用 RGB-D 视频帧进行三维表面重建的神经辐射场（NeRF）框架。在以往通过特征编码提升优化速度的方法基础上，我们通过细化深度信息，进一步提升重建质量，同时将对优化时间的影响降至最低。InFusionSurf 采用逐帧内参优化方案，消除各深度帧中由相机运动引起的模糊。该方法还将经典的实时三维表面重建方法——截断符号距离场（TSDF）融合——用作特征网格的预训练工具，从而提升重建细节和训练速度。定量与定性对比分析表明，InFusionSurf 在保持优化效率的同时，能够高精度地重建场景。消融研究进一步验证了内参细化和基于 TSDF 融合的预训练的有效性。

索引词— RGB-D 表面重建，TSDF 融合，神经辐射场，相机运动模糊

1. 引言

将基于深度测量的隐式表面表示融入神经辐射场（NeRF）的体积渲染中，Neural RGB-D [1] by Neural RGB-D [2] 显著提升了三维表面重建中的几何估计质量。然而，与许多 NeRF 变体一样，Neural RGB-D 在新场景上的优化耗时较长，根据场景大小，需要 9 至 13 小时。尽管近期显式表示方面的进展 [3, 4, 5] 显著缩短了优化时间，但要在要求快速、准确进行三维表面重建的现实应用中实用，仍有必要进一步改进。

此外，商用图像采集设备常会在视频帧中引入运动模糊、散焦模糊和滚动快门效应等失真，使 NeRF 方法难以从这些模糊输入中生成清晰图像。尽管已有一些与 NeRF 集成的去模糊方法 [6, 7] 用于彩色帧，但它们在依赖深度的三维重建中效果较差，因为深度帧上的相机运动模糊与彩色帧上的不同，对重建结果的影响更大。

在这项工作中，我们提出了*InFusionSurf*[1]，，这是一种 NeRF 风格的 RGB-D 三维表面重建框架，通过细化深度信息来提升重建质量，同时最大限度地减少对优化时间的影响。我们的逐帧内参细化方案采用显式参数，高效优化光线投射方向，从而解决深度视频帧中的相机运动模糊问题。此外，InFusionSurf 利用截断有符号距离场（TSDF）融合 [8], 这一经典的实时三维表面重建方法作为预训练步骤，为模型提供良好的初始基础。我们证明，与已有方法 Neural RGB-D [2] 和 GO-Surf [5]相比，InFusionSurf 能够准确、高效地重建场景。我们还通过消融研究验证了 TSDF 融合先验学习和逐帧内参优化技术的有效性。

[1]http s://rokit-healthcare.github.io/InFusionS urf

2. 相关工作

2.1. 经典三维重建

TSDF 融合方法的各种变体 [8] 已广泛用于根据深度图像 [9]重建三维表面。多年来，人们在此基础上进行了大量改进，涵盖从实时应用 [10]到提升重建质量 [11]等方面。

尽管其质量并非最优，我们发现，将高效的 TSDF 融合输出用作几何先验，既能提高重建质量，又能加快收敛速度。

2.2. 神经辐射场与深度

人们已尝试将神经辐射场表示和体渲染方案 [1]应用于深度图像。一些方法引入深度先验，以改善新视角合成 [12, 13], ；另一些方法则使用隐式神经表示进行三维表面重建 [2, 5]。

鉴于 NeRF 及其变体的优化时间较长，人们提出了加快收敛的方法 [3, 4]。GO-Surf [5]将多分辨率特征网格与类似 Neural RGB-D 的混合体渲染方案相结合，从而加快优化速度。

我们采用 Neural RGB-D [2] 和稠密特征网格表示，以实现精准且高效的三维场景重建。我们进一步引入逐帧内参优化方案和 TSDF 融合引导的预训练阶段，以提升重建质量；在对优化时间影响极小的情况下，重建质量优于 GO-Surf 和 Neural RGB-D。

2.3. 相机运动模糊处理

已有许多彩色图像去模糊方法被提出，包括基于卷积神经网络 [14] 或生成模型 [15] 的方法。多项研究将去模糊过程整合到 NeRF 中，采用可变形核 [6]，或在可学习的曝光时间 [7]内对虚拟清晰彩色图像进行平均，以合成模糊图像。

我们的方法与其他方法不同，重点在于校正深度帧中的失真；这种失真与彩色帧中的失真有显著差异 [16]。为解决深度特有的运动模糊，我们采用逐帧内参优化技术，通过受约束的可变形核优化渲染射线方向，

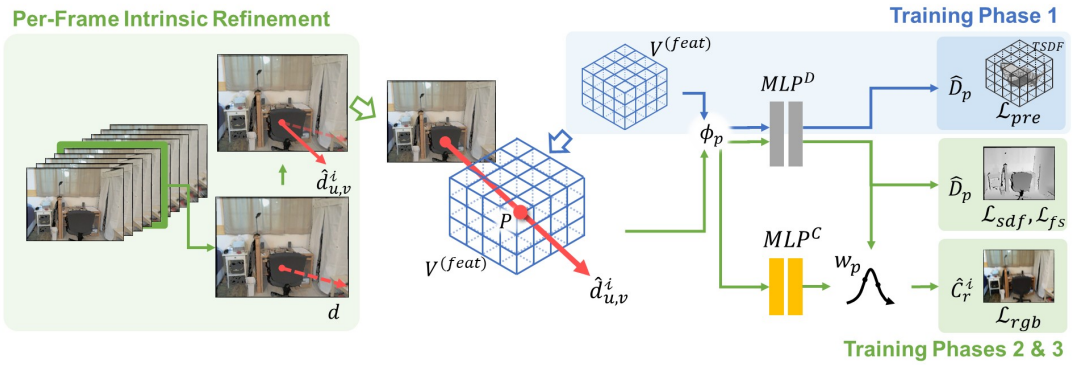


图 1：我们提出了逐帧内参细化方案和经典 TSDF 融合先验学习方案，以在尽量减少对优化时间影响的同时实现高质量的三维表面重建。我们采用 Neural RGB-D 方法，并对其进行改进，使用稠密特征网格和浅层多层感知机。我们的逐帧内参优化方案可补偿相机运动造成的特定帧畸变效应。训练的第一阶段使用 TSDF 融合算法学习几何先验，后续阶段则采用渐进式学习技术。

使其在整个图像平面上应用平移和缩放变换。

3. 方法

我们的方法基于 [2] 提出的神经 RGB-D 表面重建方案，并采用类似 DVGO [3] 的稠密特征网格，以加快优化速度。我们使用逐帧内参细化来校正相机运动误差，并采用包含 TSDF 融合的三阶段训练方案，以在降低计算负担的同时取得更好的结果。

3.1. 混合几何表示

我们使用稠密特征网格显式学习局部特征，与使用多层感知机相比，可显著降低计算复杂度和训练时间。我们根据场景大小动态调整每个场景的特征网格 $V^{(feat)}$ ——L_x、L_y、L_z，其大小根据深度帧和估计的相机位姿计算得出：

$$V^{(feat)} : (N_x \times N_y \times N_z) \mapsto \mathbb{R}^F, \tag{1}$$

其中，F 是特征向量长度的固定超参数，网格尺寸 N_x、N_y、N_z 根据单元格大小 gs 设定，计算方式为 $\lceil L_x/gs \rceil$，$\lceil L_y/gs \rceil$，$\lceil L_z/gs \rceil$。

对于一个三维点 $\mathbf{p}$，其特征向量 $\phi_\mathbf{p}$ 由最近的 8 个网格顶点 $\mathbf{P}_{near}$ 进行三线性插值得到。

$$\phi_\mathbf{p} = interp[V^{(feat)}(\mathbf{P}_{near})] \in \mathbb{R}^F \tag{2}$$

我们使用浅层多层感知机将这些特征解码为每个三维点 $\mathbf{p}$ 的视角相关颜色 $C^i_{\mathbf{p},\,\mathbf{d}}$ 和截断有符号距离值 $\widehat{D}_\mathbf{p}$，如公式 (3) 和 (4) 所示。为了解码视角相关颜色，我们将经过位置编码的射线方向 $\mathbf{d}$ 与依赖于帧的潜在嵌入向量 ξ 拼接起来。

$$C^i_{\mathbf{p},\mathbf{d}} = MLP^C(\phi_\mathbf{p}, \Lambda(\mathbf{d}), \xi_i), \tag{3}$$

$$\widehat{D}_\mathbf{p} = MLP^D(\phi_\mathbf{p}), \tag{4}$$

其中，i 是输入视频中的帧编号，Λ 是频率位置编码函数。

神经渲染流程遵循 [2] 中的方法。使用已知的相机内参矩阵，我们为每个归一化图像平面中的图像像素 $(u,\ v)$ 沿其方向$\mathbf{d}$投射一条相机射线$\boldsymbol{r}$：

$$\mathbf{d}_{u,v} = \begin{bmatrix} R \mid t \end{bmatrix} \begin{bmatrix} \varrho_{u,v} \\ 1 \end{bmatrix} \tag{5}$$

$$\varrho_{u,v} = \begin{bmatrix} 1/f_x & 0 & -c_x/f_x \\ 0 & -1/f_y & c_y/f_y \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{6}$$

其中，$\begin{bmatrix} R \mid t \end{bmatrix}$ 是估计的相机位姿矩阵，f_x，f_y 是焦距，$(c_x,\ c_y)$ 是主点。

对于一个三维点$\mathbf{p}$，其用于渲染彩色图像的权重值 $\omega_\mathbf{p}$ 根据有符号距离值计算得出：

$$\omega_\mathbf{p} = \sigma(\frac{\widehat{D}_\mathbf{p}}{tr}) \cdot \sigma(-\frac{\widehat{D}_\mathbf{p}}{tr}), \tag{7}$$

其中，tr 是截断距离，σ 是 Sigmoid 函数。

射线$\mathbf{r}$ 在 i^{th} 帧中的最终渲染颜色 $\widehat{C}$，通过沿射线采样的点$\mathbf{p}$的辐射值加权求和计算得出：

$$\widehat{C}^i_\mathbf{r} = \frac{1}{\sum \omega_\mathbf{p}} \sum \omega_\mathbf{p} C^i_{\mathbf{p},\mathbf{d}} \tag{8}$$

3.2. 逐帧内参矩阵优化

与静态图像不同，视频帧容易受到相机运动的影响，从而产生如图 2 所示的运动模糊。彩色帧中的运动模糊通常建模为在曝光时间内对像素取平均，而深度帧中的运动模糊则更类似于最小值滤波器，即取曝光时间内的最小值。这种现象可能导致边界超出实际物体边界 [16]。我们的方法通过校正相机内参矩阵来处理深度帧中的运动模糊，这实际上改变了投影图像平面的尺度和平移。

在处理逐帧内参细化问题之前，我们首先采用 [2] 中的图像平面变形场和位姿优化技术，以纠正相机位姿误差，以及内参和相机镜头畸变可能导致的全局误差。我们修改了图像平面变形场，使用双层浅层多层感知机来缩短训练时间。

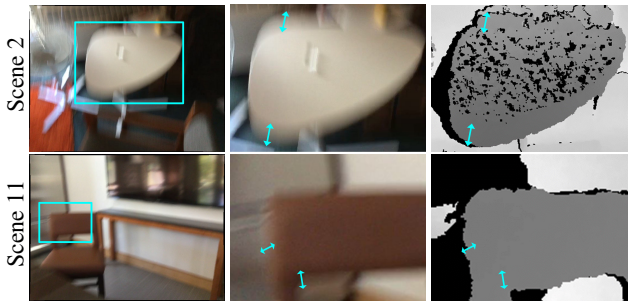


图 2：ScanNet V2 [17] 数据集中的样本展示了运动模糊带来的负面影响。RGB 帧（a、b）模糊且失真。与彩色帧不同，深度帧（c）呈现出扩展的物体边界，而非平均模糊 [16]。

我们的逐帧内参优化方案在应用图像平面变形场以纠正深度相关的运动模糊后执行。具体而言，我们为每帧引入四个参数，其中两个用于缩放，另外两个用于平移。在将归一化图像坐标转换为世界坐标之前，先对其进行缩放和平移：

$$\widehat{\varrho}^{i}_{u,v} = \mathbf{s}^{i} \cdot (\varrho_{u,v} + \tau^{i}) \tag{9}$$

$$\mathbf{s}^{i} = \begin{bmatrix} s^{i}_{x} & 0 & 0 \\ 0 & s^{i}_{y} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ \tau^{i} = \begin{bmatrix} \tau^{i}_{x} \\ \tau^{i}_{y} \\ 0 \end{bmatrix} \tag{10}$$

其中，$\mathbf{s}^{i}$ 和 τ^{i} 是针对 i^{th} 帧的可训练参数，并在训练期间进行优化。因此，细化后的投射方向 $\widehat{\ }\mathbf{d}$ 由 $\widehat{\varrho}$ 计算得出：

$$\widehat{\mathbf{d}}^{i}_{u,v} = \begin{bmatrix} R \mid t \end{bmatrix} \begin{bmatrix} \widehat{\varrho}^{i}_{u,v} \\ 1 \end{bmatrix} \tag{11}$$

式（11）取代式（5），用于生成射线和采样点。这些参数可解释为对内参矩阵的主点和焦距进行校正。

图像平面变形场和逐帧内参细化方案都旨在提高射线方向的精度。图像平面变形场对所有帧进行全局细化，而逐帧内参细化则针对图像平面变形场未能处理的特定帧波动。

3.3. 优化

如图 1 所示，InFusionSurf 通过一个三阶段训练流程进行优化。在第一阶段，InFusionSurf 使用经典的实时算法 TSDF 融合 [8]学习几何先验。具体而言，InFusionSurf 使用 F=1 构建如公式 (1) 所示的密集网格，并对同一场景的深度帧运行 TSDF 融合。在优化过程中，我们随机采样网格单元，查询其中心点的有符号距离值，并尽量减小这些值与 TSDF 融合对应值之间的均方误差。此阶段仅训练 $V_{(feat)}$ 和 MLP^{D} 的参数。

优化阶段能够直接学习有符号距离值，绕过耗时的渲染过程，最终显著加快训练阶段。

在 InFusionSurf 的第二和第三阶段，我们采用与 [3] 类似的渐进式学习，依次细化低频和高频细节。所有参数，包括第一阶段未涉及的参数，都会在这些阶段进行优化。

在第二阶段，InFusionSurf 随机采样射线批次 r_b，并沿射线每隔 15.625 毫米采样点 S_c。我们的损失函数与 [2], 类似，使用估计的有符号距离值和渲染颜色（公式 (4) 和 (8)）：

$$\mathcal{L} = \lambda_{fs}\mathcal{L}_{fs} + \lambda_{sdf}\mathcal{L}_{sdf} + \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{reg}\mathcal{L}_{reg} \tag{12}$$

$\mathcal{L}_{fs}$ 和 $\mathcal{L}_{sdf}$ 分别表示截断区域外（S_{fs}）和截断区域内（S_{sdf}）点的损失分量：

$$\mathcal{L}_{fs} = \frac{1}{|r_b|}\sum_{r\in r_b}\frac{1}{|S_{fs}|}\sum_{\mathbf{p}\in S_{fs}}(\widehat{D}_{\mathbf{p}} - tr)^2, \tag{13}$$

$$\mathcal{L}_{sdf} = \frac{1}{|r_b|}\sum_{r\in r_b}\frac{1}{|S_{sdf}|}\sum_{\mathbf{p}\in S_{sdf}}(\widehat{D}_{\mathbf{p}} - D^{i}_{\mathbf{r}})^2 \tag{14}$$

其中，$D^{i}_{\mathbf{r}}$ 是深度帧中观测到的有符号距离值。

$\mathcal{L}_{rgb}$ 用于衡量渲染颜色与对应像素的观测颜色之间的差异：

$$\mathcal{L}_{rgb} = \frac{1}{|r_b|}\sum_{r\in r_b}(\widehat{C}^{i}_{\mathbf{r}} - C^{i}_{u,v})^2 \tag{15}$$

术语 $\mathcal{L}_{reg}$ 表示对帧相关的潜在嵌入向量、逐帧内参优化参数和图像平面形变参数进行的 L2 正则化。值得注意的是，逐帧内参优化中的缩放参数以 1 为中心进行正则化。

在第三阶段，InFusionSurf 通过将稠密特征网格划分为更高分辨率来关注细节，并将网格单元格大小减半为 gs。在 S_c 中识别出的表面周围会额外采样点 S_f。该阶段采用相同的损失函数（公式（12））， 同时利用 S_c 和 S_f 进行训练。

3.4. 实现细节

我们的网络基于 PyTorch 构建，并使用 ADAM [18]进行优化，初始设置为 5×10^{-4} 学习率（每 250K 次迭代按指数衰减至十分之一）、0.9 的 beta1 和 0.999 的 beta2。我们将特征网格设置为 F=12 和 gs=10cm，权重初始化为 0，截断有符号距离值初始化为 tr=5cm。颜色多层感知机和有符号距离多层感知机均包含 2 个隐藏层，每层 128 个节点；图像平面形变多层感知机则包含 2 层，每层 64 个节点。逐帧内参细化参数的初始值为：缩放为 1，平移为 0。损失项权重为 λ_{fs}=10、 λ_{sdf}=6 × 10^{3}、λ_{rgb}=0 5, λ_{reg}=0 1 3 7 65 。以及 . . 我们分别在每个训练阶段进行了 K、K 和 K 次迭代。我们使用了 TSDF 融合算法的 CUDA 实现，在 Tesla V100 GPU [19] 上平均每秒可处理 231.7 帧。

4. 实验

我们将 InFusionSurf 与先前的工作进行了对比研究，并对所提出框架的各个组成部分进行了消融研究，以展示逐帧内参细化和 TSDF 融合引导训练阶段的影响。更多实验和结果请参阅附录部分。

在评估每项研究时，我们以 1cm^3 的分辨率提取截断有符号距离值，并运行 March- ingCubes [20] 算法以生成三角网格。

4.1. 数据集

在定性研究中，我们使用 ScanNet V2 [17] 作为真实场景数据集。该数据集在图像采集过程中采用滚动快门方法，导致运动模糊、失真以及含噪的深度值，其中包括孔洞和缺失物体。

为了将定量结果与基线方法进行比较，我们使用了来自 [2] 的 10 个合成场景。它们使用室内三维模型，以照片级效果渲染彩色和深度帧，并提供真实相机轨迹。深度帧模拟了类 Kinect 噪声，包括孔洞和量化噪声。

4.2. 基线方法

我们将结果与 NeRF 风格的 RGB-D 三维表面重建方法 Neural RGB-D [2] 和 GO-Surf [5]进行比较。具体而言，我们分别报告 20K 和 75K 次迭代后的结果，以便分别与 GO-Surf 和 Neural RGB-D 进行比较，展示 InFusion-Surf 在效率和性能方面均更胜一筹。我们在 Tesla V100 GPU 上，使用各论文推荐的超参数训练了 GO-Surf 和 Neural RGB-D。

4.3. 结果与讨论

4.3.1. 定性结果

定性结果如图4所示。为了与 GO-Surf 进行比较，图中展示了我们经过 20K 次迭代后的重建结果。如图4所示，即使训练时间更短，InFusionSurf 在整个场景中仍呈现出更精细的细节和更少的错误表面。

经过 75K 次迭代后，我们的重建质量优于 Neural RGB-D，在某些情况下还重建出了 Neural RGB-D 遗漏的结构。与 Neural RGB-D 相比，我们的训练速度快了 7.3 至 8.7 倍。

4.3.2. 定量结果

我们从 Chamfer ℓ_1距离（C-ℓ_1）、交并比（IoU）、法线一致性（NC）和 F值等方面，将我们的方法与基线方法进行了比较。为了计算 C-ℓ_1、NC 和 F值，我们以 1cm^2 分辨率从输出网格中采样点云。F值的阈值设为 5cm。为计算交并比，我们将评估网格体素化，边长设为 10cm。如表1所示，经过 20K 次迭代后，我们的结果在 C-ℓ_1、交并比和 F值方面优于 GO-Surf，且训练时间更短（平均快 20%）。与 Neural RGB-D 相比，我们的方法在 C-ℓ_1、交并比和 F值方面也表现更优。同时，训练 75K 次迭代平均耗时 96 分钟，比 Neural RGB-D 快 7 倍。尽管 InFusionSurf 在三项主要指标上表现出色，但其法线一致性较低，表明重建结果包含相对不平整的表面。定量指标表明，我们的方法最适合快速重建复杂几何体，而非简单的平坦表面。

4.3.3. 消融研究

消融研究展示了逐帧内参细化和 TSDF 融合引导训练阶段的作用。为进行定性评估，我们比较了在相同迭代节点下，不进行逐帧内参细化或训练第一阶段——TSDF 融合先验学习（图3）时的结果。为了进行定量

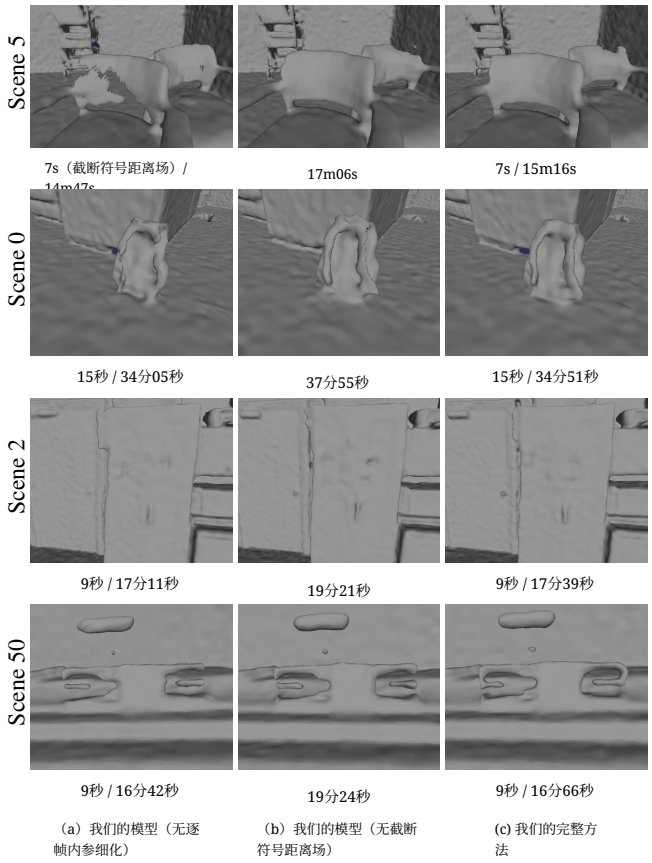


图3：消融研究。(a) 我们的方法不使用逐帧内参细化（PFIR）。(b) 我们的方法在训练第一阶段不使用 TSDF 融合先验学习（TSDF）。(c) 我们的方法同时采用这两种方法。子图下方的时间戳表示 TSDF 融合先验学习所用时间（如适用）和总训练时间。

评估 TSDF 融合的有效性，我们针对有无 TSDF 融合引导训练阶段进行了定量研究（表1）。

如图3所示，我们的逐帧内参细化方法修正了物体中存在错误的部分。我们的方法每帧仅增加四个额外参数，这些参数可在训练早期进行优化。该方法只需少量迭代即可修正重建误差，平均仅使训练时间增加 2.7%。

如图3所示，TSDF 融合先验学习阶段能够持续改善所有场景中的物体重建细节。此外，在相同迭代次数下，缺少 TSDF 融合先验学习会导致训练时间额外增加数分钟。表1所呈现的结果进一步支持了这一一致性规律：我们的方法在 C-ℓ_1、NC 和 F值指标上均表现出色，同时训练时间也有所缩短。训练过程的第一阶段利用先验知识，直接学习三维坐标的有符号距离值，而无需耗时的渲染过程，因此速度极快，尽管会存在一定的不准确性。后续训练阶段则可以专注于细化场景中的复杂细节，最终在缩短收敛所需时间的同时，提高重建质量。

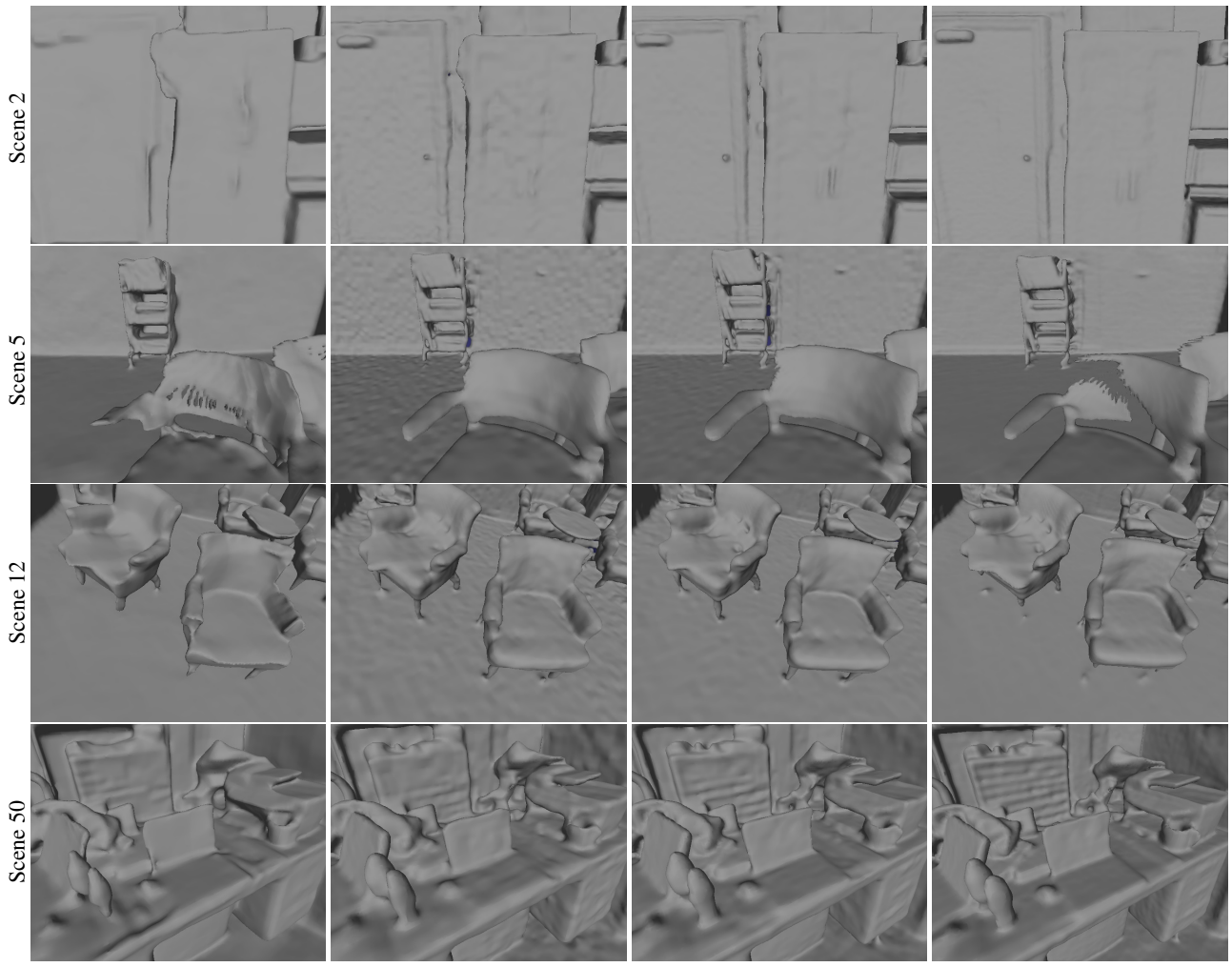


图 4：我们在不同时间点将本方法与 GO-Surf [5] 和 Neural RGB-D [2] 进行了比较。比较使用了 ScanNet V2 [17]中的场景 2、5、12 和 50。经过较短时间的训练后，InFusionSurf-20K (b) 恢复了 GO-Surf (a) 所忽略的高频细节，并生成了少得多的错误表面。在更长的训练时间下，InFusionSurf-75K (c) 实现了更高的质量，同时恢复了 Neural RGB-D (d) 中缺失的一些几何结构。

方法	C-ℓ_1 ↓	交并比 ↑	NC ↑	F值 ↑	Time
GO-Surf	0.042	0.723	**0.922**	0.918	22分57秒
Neural RGB-D	0.052	<u>0.757</u>	**0.922**	<u>0.938</u>	669分59秒
本文方法（20K）	**0.038**	0.737	0.904	0.929	18分24秒
本文方法（20K，不使用截断符号距离场）	0.042	0.750	0.902	0.926	21分05秒
我们的方法（75K）	<u>0.041</u>	**0.768**	<u>0.913</u>	**0.939**	95分57秒

表1：合成场景数据集上的定量结果。与 GO-Surf [5] 相比，InFusionSurf 在训练时间更短的情况下，C-ℓ_1、交并比和 F值表现更佳。经过更多迭代后，其性能优于 Neural RGB-D [2]，且训练时间显著更短。20K 次迭代且不使用 TSDF 融合（TSDF）的性能表明，我们的 TSDF 融合引导训练阶段既提升了重建质量，也缩短了训练时间。

4.4. 局限性

我们方法的一个主要局限性在于依赖简单的变换矩阵来校正相机运动造成的运动模糊，有效处理均匀的帧失真，但无法处理物体运动或滚动快门效应等局部模糊。采用先进技术增强逐帧内参优化模块，或许能提高对这些尚未解决的失真的处理精度，但可能会牺牲处理速度。

未来的工作可以探索如何权衡精度与速度，以同时提升二者。

在评估方面，由于我们的重点是重建不透明表面，因此使用了主要由漫反射物体构成的 ScanNet 数据集。未来，将该框架扩展为能够处理透明或反射表面，或许是一个有前景的研究方向。

5. 结论

本文介绍了 InFusionSurf，这是一种 NeRF 风格 RGB-D 三维重建框架，通过利用逐帧内参优化和截断符号距离场融合，在对优化耗时影响极小的情况下提升重建质量。全面的对比评估表明，我们的方法能够准确重建具有高频细节的场景。

6. 致谢

本研究得到韩国政府（科学技术信息通信部、保健福祉部）资助的韩国再生医学基金（KFRM）项目（23B0104L1）以及 ROKIT Healthcare, Inc. 的支持。

7. 参考文献

[1]Ben Mildenhall、Pratul P Srinivasan、Matthew Tancik、Jonathan T Barron、Ravi Ramamoorthi 和 Ren Ng，《NeRF：将场景表示为用于视图合成的神经辐射场》，《*ACM* 通讯》，第 65 卷，106，2021 1，第 99 页。[2] Dejan Azinovi´c、Ricardo Martin-Brualla、Dan B Goldman、Matthias Nießner 和 Justus Thies，《神经 RGB-D 表面重建》，载于 *IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）论文集，2022 年6月，第 6290–6301 页。[3] Cheng Sun、Min Sun 和 Hwann-Tzong Chen，《直接体素网格优化：辐射场重建的超快收敛》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2022 年，第 5459–5469 页。[4] Thomas M¨uller、Alex Evans、Christoph Schied 和 Alexander Keller，《采用多分辨率哈希编码的即时神经图形基元》，*ACM* 图形学汇刊，第 41 卷，第 4 期，第 102:1–102:15 页，2022 年 7 月。[5] Jingwen Wang、Tymoteusz Bleja 和 Lourdes Agapito，《Go-Surf：用于快速、高保真 RGB-D 表面重建的神经特征网格优化》，载于*2022* 年三维视觉国际会议（*3DV*）。IEEE，2022 年，第 433–442 页。[6] Dogyoon Lee、Minhyeok Lee、Chajin Shin 和 Sangyoun Lee，《DP-NeRF：结合物理场景先验的去模糊神经辐射场》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2023 年，第 12386–12396 页。[7] P. Wang、L. Zhao、R. Ma 和 P. Liu，《BAD-NeRF：通过束调整实现去模糊神经辐射场》，载于*2023* 年 *IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*），美国加利福尼亚州洛斯阿拉米托斯，2023 年 6 月，第 4170–4179 页，IEEE 计算机学会。[8] Brian Curless 和 Marc Levoy，《一种利用距离图像构建复杂模型的体积法》，载于论文集

第*23*届计算机图形学与交互技术年会，美国纽约，1996年，SIGGRAPH '96，第303–312页，美国计算机协会。[9] Michael Zollh¨ofer、Patrick Stotko、Andreas G¨orlitz、Christian Theobalt、Matthias Nießner、Reinhard Klein 和 Andreas Kolb，《RGB-D 相机三维重建的最新进展》，《计算机图形学论坛》，第37卷，652，2018年5月，625页。[10] Richard A. Newcombe、Shahram Izadi、Otmar Hilliges、David Molyneaux、David Kim、Andrew J. Davison、Pushmeet Kohi、Jamie Shotton、Steve Hodges 和 Andrew Fitzgibbon，《KinectFusion：实时稠密表面映射与跟踪》，载于*2011*年第*10*届*IEEE*混合与增强现实国际研讨会，2011年，第127–136页。[11] Angela Dai、Matthias Nießner、Michael Zollh¨ofer、Shahram Izadi 和 Christian Theobalt，《BundleFusion：通过即时表面重新整合实现实时全局一致的三维重建》，《*ACM* 图形学汇刊》（*ToG*），第36卷，第4期，第1页，2017年。[12] Kangle Deng、Andrew Liu、Jun-Yan Zhu 和 Deva Ramanan，《深度监督NeRF：以更少视图实现更快的自由视角训练》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2022年，第12882–12891页。[13] T. Neff、P. Stadlbauer、M. Parger、A. Kurz、J. H. Mueller、C. R. A. Chaitanya、A. Kaplanyan 和 M. Steinberger，《DON-eRF：利用深度预言网络实现紧凑神经辐射场实时渲染》，《计算机图形学论坛》，第40卷，第4期，第45–59页，2021年7月。[14] Syed Waqas Zamir、Aditya Arora、Salman Khan、Munawar Hayat、Fahad Shahbaz Khan、Ming-Hsuan Yang 和 Ling Shao，《多阶段渐进式图像复原》，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2021年，第14821–14831页。[15] Orest Kupyn、Tetiana Martyniuk、Junru Wu 和 Zhangyang Wang，《DeblurGAN-v2：更快、更优的去模糊方法（速度提升数个数量级）》，载于*IEEE/CVF* 国际计算机视觉会议论文集，2019年，第8878–8887页。[16] Siddharth Tourani、Sudhanshu Mittal、Akhil Nagariya、Visesh Chari 和 Madhava Krishna，《深度相机的卷帘快门与运动模糊去除》，载于*2016 IEEE*机器人与自动化国际会议（*ICRA*），2016年，第5098–5105页。[17] Angela Dai、Angel X. Chang、Manolis Savva、Maciej Halber、Thomas Funkhouser 和 Matthias Nießner，《ScanNet：带有丰富标注的室内场景三维重建》，载于*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，2017年。[18] Diederik P. Kingma 和 Jimmy Ba，《Adam：一种随机优化方法》，载于第三届国际学习表征会议（*ICLR 2015*），美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集，Yoshua Bengio 和 Yann LeCun 编，2015年。[19] Andy Zeng、Shuran Song、Matthias Nießner、Matthew Fisher、Jianxiong Xiao 和 Thomas Funkhouser，《3DMatch：从RGB-D重建中学习局部几何描述子》，载于*CVPR*，2017年。[20] William E. Lorensen 和 Harvey E. Cline，《行进立方体：一种高分辨率三维表面构建算法》，载于第*14*届计算机图形学与交互技术年会论文集，美国纽约，1987年，SIGGRAPH '87，第163–169页，美国计算机协会。

8. 附录

本节中，我们将介绍一系列实验，以展示逐帧内参优化模块的有效性。此外，我们还与其他基于 RGB 的三维重建方法进行了对比研究，以进一步证明我们方法的优势。

8.1. 逐帧内参细化的有效性

我们在合成数据集 [1] 上，通过模拟内参矩阵中的误差，对逐帧内参细化的有效性进行了全面研究。为此，我们特意将焦距初始化为 570，而非其真实值 554.26。此外，我们还在每一帧的焦距和主点上引入了服从正态分布 $\mathcal{N}(0, 10^2)$ 的随机扰动。随后，我们针对不同情形对模型进行优化，分别省略逐帧内参细化、图像平面变形场，或同时省略二者（表 2）。表 2 中的定量结果表明，图像平面变形场和逐帧内参细化对提升重建质量具有互补作用。

PFIR	IDPF	C-ℓ_1 ↓	交并比 ↑	NC ↑	F值 ↑
		0.081	0.364	0.845	0.514
	✓	0.067	0.486	0.861	0.645
✓		0.060	0.579	0.883	0.795
✓	✓	**0.051**	**0.656**	**0.888**	**0.863**

表 2：在定量数据集上对逐帧内参细化（PFIR）和图像平面变形场（IPDF）进行的消融研究。我们的 PFIR 和 IPDF 方案对重建质量产生了互补影响。

8.2. 内参细化后的深度帧可视化

为定性评估逐帧内参优化模块的细化效果，我们使用 ScanNet V2 数据集中的深度帧进行了实验。对于每一帧，我们使用原始内参矩阵将所有像素转换为图像坐标，应用缩放和平移，然后使用相同的内参矩阵将其重新投影为像素。如图 5 所示，结果证明了逐帧内参优化模块能够有效修正深度帧中的物体边界。通过补偿相机运动造成的误差，我们的方法实现了更准确的重建。

8.3. 与基于 RGB 的方法比较

此外，除了与基于 RGB-D 的三维重建方法进行对比实验外，我们还将分析扩展至两种主要的基于 RGB 的方法：Voxurf [2] 和 Neuralangelo [3]。我们使用输入的 RGB 帧通过 COLMAP [4] 恢复相机参数，然后应用这些方法从 ScanNet V2 [5]重建室内场景。如图 6 所示，基于 RGB 的模型难以准确重建场景。我们认为，这一局限性源于其固有设计：该设计依赖上下文信息来施加多视图约束。尤其是在稀疏视图视频采集场景中，仅依赖颜色输入可能无法提供足够的信息来进行恰当的重建。

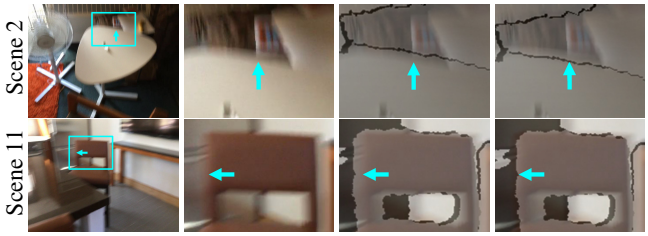


图 5：逐帧内参优化模块的可视化。(a)、(b) 带有相机运动模糊的彩色帧。(c) 叠加的深度帧显示物体边界被错误地延伸。(d) 我们的逐帧内参优化模块使深度帧与实际边界对齐。

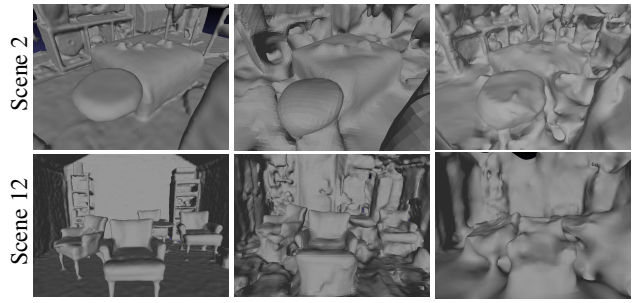


图 6：ScanNet V2 [5] 室内场景的对比结果。此图对比了我们的方法 (a) 与基于 RGB 的方法 (b) 和 (c) 在重建室内场景方面的效果。基于 RGB 的方法似乎难以准确重建。

9. 参考文献

[1] Dejan Azinovi´c、Ricardo Martin-Brualla、Dan B Goldman、Matthias Nießner 和 Justus Thies，“Neural RGB-D 表面重建”，载于*IEEE/CVF* 计算机视觉与模式识别会议（*CVPR*）论文集，2022年6月，第 6290–6301 页。[2] Tong Wu、Jiaqi Wang、Xingang Pan、Xudong Xu、Christian Theobalt、Ziwei Liu 和 Dahua Lin，“Voxurf：基于体素的高效精准神经表面重建”，载于国际学习表征会议（*ICLR*），2023年。[3] Zhaoshuo Li、Thomas M¨uller、Alex Evans、Russell H Taylor、Mathias Unberath、Ming-Yu Liu 和 Chen-Hsuan Lin，“Neuralangelo：高保真神经表面重建”，载于*IEEE* 计算机视觉与模式识别会议（*CVPR*），2023年。[4] Johannes Lutz Sch¨onberger 和 Jan-Michael Frahm，“重新审视运动恢复结构”，载于计算机视觉与模式识别会议（*CVPR*），2016年。[5] Angela Dai、Angel X. Chang、Manolis Savva、Maciej Halber、Thomas Funkhouser 和 Matthias Nießner，“ScanNet：室内场景的丰富标注三维重建”，载于计算机视觉与模式识别会议（*CVPR*）论文集，*IEEE*，2017年。