DUSt3R: Geometric 3D Vision Made Easy

摘要

在多视图立体重建（MVS）的野外应用中，通常需要估计相机的内部和外部参数，这些参数对于在 3D 空间中三角化对应像素是必要的，而此过程往往繁琐且复杂。这项工作引入了 DUSt3R [3]，一种全新的密集且不受约束的立体 3D 重建范式，适用于任意图像集合，在无需相机校准或视角姿态先验信息的情况下运行；通过将成对重建问题视为点图回归问题，放松了传统投影相机模型的严格约束，并展示了该方法可以平滑地统一单目和双目重建案例。对于超过两张图片的情况，这项工作提出了一种简单而有效的全局对齐策略，将所有成对点图表达在一个共同的参考框架中。基于标准的 Transformer 编码器和解码器的网络架构允许利用强大的预训练模型，从而直接提供场景的 3D 模型及深度信息，并从中无缝恢复像素匹配、焦距以及相对和绝对相机位置。广泛的实验表明，DUSt3R 有效地统一了各种 3D 视觉任务，在单目和多视图深度估计及相对姿态估计方面设立了新的性能记录，使许多几何 3D 视觉任务变得更加简单。

关键词：三维重建；几何建模与处理

1 引言

不受约束的密集 3D 重建是从多张 RGB 图像中恢复特定场景的 3D 几何结构和相机参数的任务，这一直是计算机视觉领域的重要目标。这项工作专注于解决这一问题，旨在从未经校准和未定位的图像中实现端到端的 3D 重建。它不仅在地图绘制、导航、考古学、文化遗产保护及机器人技术等领域有着广泛应用，而且在所有 3D 视觉研究任务中占有特殊地位，几乎涵盖了所有的几何 3D 视觉任务。

传统上，现代 3D 重建方法依赖于一系列按顺序执行的组件，如特征点检测与匹配、鲁棒估计、从运动恢复结构（SfM）[1]、束调整（BA）以及密集多视图立体（MVS）[2]。尽管这种方法在某些情况下是可行的解决方案，但其复杂性高且各组件之间的缺乏交流导致了效率低下，每个步骤并非完美无缺，并可能为后续步骤引入噪声，增加了整体工作的复杂性和工程难度。此外，关键步骤在许多常见情况下容易失败，例如 SfM 阶段在处理视角较少、非朗伯表面物体或摄像机运动不足或过大的情况下表现不佳。

为了解决这些问题，这项工作提出了 DUSt3R，一种全新的密集且不受约束的立体 3D 重建方法。DUSt3R 的核心是一个可以从一对图像直接回归出一个密集且精确场景表示的网络，无需任何关于场景或相机的先验信息，甚至不需要相机的内部参数。该网络输出基于 3D 点图，这些点图同时封装了场景的几何结构、像素与场景点之间的关系，以及两个视角间的关系。通过这种方式，几乎所有场景参数都可以直接从输出中简便地恢复，因为网络联合处理

输入图像和结果 3D 点图，从而学习将 2D 模式与 3D 形状关联起来，并能够同时解决多个任务，实现内部“协作”。

训练方面，这项工作采用了完全监督的方法，使用简单的回归损失函数，并利用大型公共数据集进行训练。这些数据集的真实标注或是合成生成，或是通过 SfM 软件重构，或者是用专用传感器捕获。不同于集成任务特定模块的趋势，这项工作采取了一种全数据驱动策略，基于通用 Transformer 架构，在推理时不强制任何几何约束，却能受益于强大的预训练方案，使网络能够学习到强几何和形状先验。

为了融合来自多个图像对的预测，这项工作重新审视了针对点图的 BA 过程，实现了大规模 MVS。为此引入了一种全局对齐程序，该程序不涉及最小化重投影误差，而是在 3D 空间中直接优化相机姿态和场景几何，显示出快速且优秀的收敛性能。实验表明，这项工作的重建结果在各种未知传感器的实际场景中是准确且一致的，并且证明相同的架构可以无缝处理实际单目和多视图重建场景。

总结来说，这项工作的贡献包括：提出首个从未经校准和未定位图像中进行端到端 3D 重建的综合管道；为 MVS 应用引入了点图表示，简化了几何约束；引入了一种优化程序以在全球范围内对点图进行对齐，提取经典 SfM 和 MVS 流水线中的中间输出；并在一系列 3D 视觉任务上展示了优异性能，特别是在单目和多视图深度基准测试以及多视图相机姿态估计方面达到了最先进水平。

2 相关工作

2.1 传统三维重建

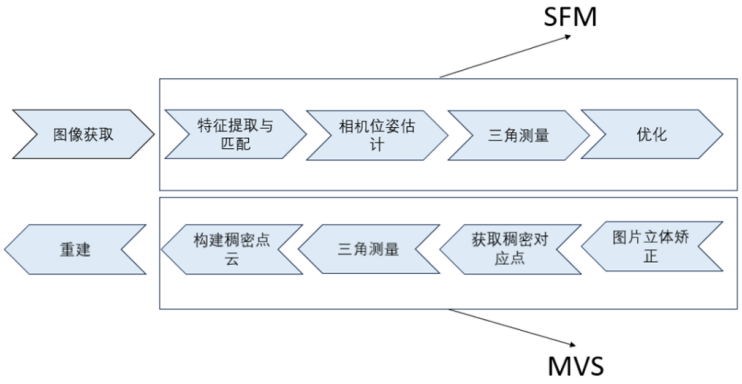

图 1. 传统三维重建过程示意图

从运动恢复结构（SfM）旨在通过一组图像同时重建稀疏 3D 地图并确定相机参数。传统流程始于多张图像之间通过特征点匹配获得的像素对应关系来确定几何关系，随后进行束调整以联合优化 3D 坐标和相机参数。近年来，SfM 流程经历了显著改进，特别是在其子过程中融入了基于学习的技术。这些改进包括高级特征描述、更精确的图像匹配、特征度量精炼以及神经束调整。尽管有这些进步，SfM 流程的顺序结构仍然存在，使得它容易受到各组件中的噪声和错误的影响。

多视图立体（MVS）的任务是密集地重建可见表面，这是通过多个视角之间的三角化实现的。在经典的 3D 重建工作中和视图合成的 MVS 中，所有相机参数都被假定为输入提供。

无论是完全手工设计的方法，还是基于场景优化的较新方法，亦或是基于学习的方法，都依赖于通过复杂的校准程序获得的相机参数估计值，这些程序可能是在数据采集期间或使用 SfM 方法进行野外重建时完成的。然而，在实际场景中，预估的相机参数不准确可能会对这些算法的正常运行造成不利影响。这项工作则提出直接预测可见表面的几何形状，而无需任何明确的相机参数知识。

2.2 直接从 RGB 到点云的三维重建

直接从 RGB 到点云。最近，一些方法旨在直接从一两个 RGB 图像预测 3D 几何结构。由于问题本质上是不确定的，除非引入额外假设，否则这些方法利用从大型数据集中学习强 3D 先验的神经网络来解决不确定性。这些方法可以分为两类。第一类利用类别级别的物体先验或扩散模型生成新的视角用于对象中心的重建。第二类工作与我们的方法最为接近，专注于一般场景。当从单个图像开始时，大量使用单目深度估计网络。深度图确实编码了一种形式的 3D 信息，并结合相机内参可以直接产生像素对齐的 3D 点云。例如，某些方法通过渲染特征增强的深度图来进行单图像的新视角合成，前提是已知所有相机参数。如果未知，可以通过利用视频帧中的时间一致性或由专门网络回归来恢复相机内参。然而，所有这些方法本质上受限于深度估计的质量，这在单目设置中显然是不确定的。为了解决这个问题，过去提出了多视图网络用于直接 3D 重建。它们主要基于构建一个可微分的 SfM 流程的想法，复制传统流程但进行端到端训练。然而，同样需要将真实的相机内参作为输入，输出通常是深度图和相对相机姿态。相比之下，我们的网络输出的是点图，即密集的 2D 3D 点场，它隐式处理相机姿态而不需要任何相机内参。

点图。使用一系列点图作为形状表示对于 MVS 来说似乎有些违反直觉，但在视觉定位任务中这种用法非常普遍，无论是在场景依赖的优化方法还是场景无关的推理方法中。类似地，视角建模是单目论文作品中的常见主题，其想法是将规范的 3D 形状存储在多个规范视图中以在图像空间中工作。这些方法通常利用显式的透视相机几何，通过渲染规范表示。

3 本文方法

在深入探讨我们方法的细节之前，下面介绍一些基本概念。

点图：在下文中，我们将一个密集的 2D 3D 点场表示为点图 $X \in \mathbb{R}^{W \times H \times 3}$。与之关联的是分辨率为 $W \times H$ 的对应 RGB 图像 I，X 形成了图像像素和 3D 场景点之间的一对一映射，即 $I_{i,j} \leftrightarrow X_{i,j}$，对于所有像素坐标 $(i,j) \in \{1 \dots W\} \times \{1 \dots H\}$。这里假设每条相机光线击中单个 3D 点，即忽略半透明表面的情况。

相机和场景：给定相机内参 $K \in \mathbb{R}^{3\times3}$，可以通过真实的深度图 $D \in \mathbb{R}^{W \times H}$ 直接获得观测场景的点图 X，公式如下：

$$X_{i,j} = K^{-1} D_{i,j} \begin{bmatrix} i \\ j \\ 1 \end{bmatrix} \tag{1}$$

这里，X 表达于相机坐标系中。接下来，我们将来自相机 n 并表达于相机 m 坐标系中的点图 X_n 记作 $X_{n,m}$：

$$X_{n,m} = P_m P_n^{-1} h(X_n) \tag{2}$$

其中 $P_m, P_n \in \mathbb{R}^{3\times4}$ 分别是图像 m 和 n 的世界到相机的姿态，而 $h:(x,y,z)\to(x,y,z,1)$ 是齐次映射。

3.1 本文方法概述

我们希望构建一个网络，该网络通过直接回归解决广义立体情况下的 3D 重建任务。为此，我们训练一个网络 f，它以两个 RGB 图像 $I_1, I_2 \in \mathbb{R}^{W\times H\times3}$ 作为输入，并输出两个对应的点图 $X_{1,1}, X_{2,1} \in \mathbb{R}^{W\times H\times3}$ 及其相关的置信度图 $C_{1,1}, C_{2,1} \in \mathbb{R}^{W\times H}$。注意，这两个点图均表达在同一个坐标系中，即 I_1 的坐标系，这与现有方法有根本性的不同，但提供了关键的优势（见第 1、2、3.3 和 3.4 节）。为了清晰起见且不失一般性，这里假设两个图像的分辨率均为 $W \times H$，但在实际应用中，它们的分辨率可以不同。

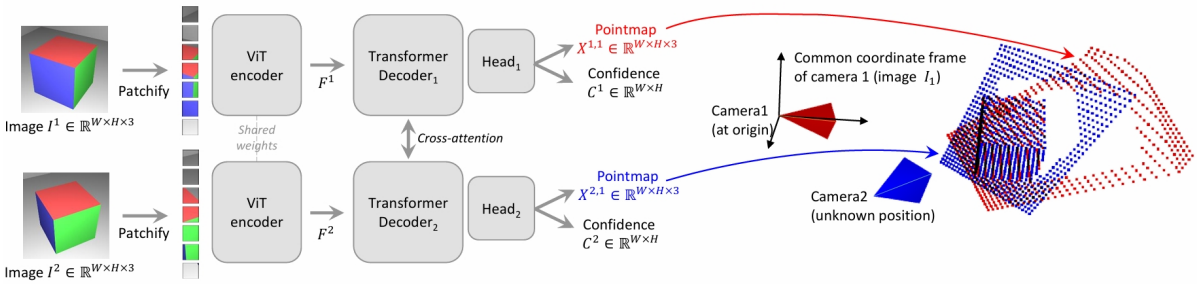

图 2. 框架图

网络架构：我们的网络 f 的架构受到 CroCo 的启发，使得我们可以充分利用 CroCo 的预训练成果。如图 2 所示，它由两个相同的分支组成（每个图像一个），每个分支包含一个图像编码器、解码器和回归头。两个输入图像首先通过共享权重的 ViT 编码器以 Siamese 方式编码，生成两个令牌表示 F_1 和 F_2：

$$F_1 = \text{Encoder}(I_1), \quad F_2 = \text{Encoder}(I_2).$$

网络随后在解码器中联合处理这两个令牌表示。类似于 CroCo 的架构，解码器是一个配备交叉注意力机制的通用 Transformer 网络。每个解码器块因此依次执行自注意力（每个视图中的令牌关注同一视图中的其他令牌），然后是交叉注意力（每个视图中的令牌关注另一视图中的所有令牌），最后通过多层感知机（MLP）处理令牌。重要的是，在解码过程中信息在两个分支之间不断共享。这是为了输出对齐正确的点图所必需的。具体来说，每个解码器块都会关注来自另一分支的令牌：

$$G_1^i = \text{DecoderBlock}_1^i(G_1^{i-1}, G_2^{i-1}),$$
$$G_2^i = \text{DecoderBlock}_2^i(G_2^{i-1}, G_1^{i-1}),$$

对于 $i = 1,\dots,B$，其中 B 是解码器的块数，并且初始化为编码器令牌 $G_1^0 := F_1$ 和 $G_2^0 := F_2$。这里，$\text{DecoderBlock}_v^i(G_1, G_2)$ 表示分支 $v \in \{1,2\}$ 中的第 i 个块，G_1 和 G_2 是输入令牌，而 G_2 是来自另一分支的令牌。最终，在每个分支中一个单独的回归头接收解码器令牌集并输出点图及其相应的置信度图：

$$X_{1,1}, C_{1,1} = \text{Head}_1(G_1^0, \dots, G_1^B),$$
$$X_{2,1}, C_{2,1} = \text{Head}_2(G_2^0, \dots, G_2^B).$$

输出的点图 $X_{1,1}$ 和 $X_{2,1}$ 回归至一个未知的比例因子。需要注意的是，我们的通用架构从未显式地强制任何几何约束。因此，点图不一定对应于任何物理上合理的相机模型。相反，我们让网络从训练集中学习所有相关的先验知识，这些点图仅包含几何一致的数据。使用通用架构允许利用强大的预训练技术，最终超越现有任务特定架构所能达到的效果。

3.2 训练目标

我们的唯一训练目标基于 3D 空间中的回归。设地面真值点图为 $\bar{X}_{1,1}$ 和 $\bar{X}_{2,1}$，它们由公式 (1) 获得，并且每个视图有两组有效的像素集 $D_1, D_2 \subseteq \{1 \dots W\} \times \{1 \dots H\}$，这些是定义了地面真值的像素。对于视图 $v \in \{1, 2\}$ 中的有效像素 $i \in D_v$，回归损失简单地定义为欧几里得距离：

$$\ell_{\text{regr}}(v, i) = \frac{1}{z X_{v,1,i} - \frac{1}{\bar{z}} \bar{X}_{v,1,i}}. \tag{2}$$

为了处理预测与地面真值之间的尺度不确定性，我们通过缩放因子 $z = \text{norm}(X_{1,1}, X_{2,1})$ 和 $\bar{z} = \text{norm}(\bar{X}_{1,1}, \bar{X}_{2,1})$ 分别对预测和地面真值点图进行标准化，其中 $\text{norm}(X_1, X_2)$ 表示所有有效点到原点的平均距离：

$$\text{norm}(X_1, X_2) = \frac{1}{|D_1| + |D_2|} \sum_{v \in \{1,2\}} \sum_{i \in D_v} \|X_{v,i}\|. \tag{3}$$

置信度感知损失 (Confidence-aware Loss)

在现实中，与我们的假设相反，存在一些难以定义的 3D 点，例如天空或半透明物体上的点。更广泛地说，图像中的一些部分通常比其他部分更难预测。因此，我们联合学习为每个像素预测一个分数，该分数表示网络对该特定像素的信心。最终的训练目标是所有有效像素的置信度加权回归损失（来自公式 (2)）：

$$L_{\text{conf}} = \sum_{v \in \{1,2\}} \sum_{i \in D_v} C_{v,1,i} \ell_{\text{regr}}(v, i) - \alpha \log C_{v,1,i}, \tag{4}$$

其中 $C_{v,1,i}$ 是像素 i 的置信度得分，α 是控制正则化项的超参数。为了确保置信度严格为正，我们通常定义 $C_{v,1,i} = 1 + \exp(c_{v,1,i}) > 0$，其中 $c_{v,1,i} \in \mathbb{R}$。这使得网络被迫在较难的区域（如单视图覆盖的区域）进行外推。使用此目标训练网络 f 可以在没有显式监督的情况下估计置信度得分。输入图像对及其相应输出的例子见图 3 及补充材料中的图 1、2 和 5。

3.3 下游任务

输出点图的丰富属性允许我们相对容易地执行各种便捷操作

点匹配 (Point Matching)

在两个图像的像素之间建立对应关系可以通过 3D 点图空间中的最近邻（NN）搜索轻松实现。为了最小化错误，我们通常保留互为最近邻的对应关系 $M_{1,2}$ 于图像 I_1 和 I_2 之间，即：

$$M_{1,2} = \{(a, b) \mid a = \text{NN}_{1,2}(b) \text{ and } b = \text{NN}_{2,1}(a)\}$$

其中，

$$\text{NN}_{n,m}(a) = \arg \min_{b \in \{0, \dots, WH\}} \|X_{n,1,b} - X_{m,1,a}\|.$$

恢复内参 (Recovering Intrinsics)

根据定义，点图 $X_{1,1}$ 表达在 I_1 的坐标系中。因此，通过求解一个简单的优化问题可以估计相机内参。在本工作中，我们假设主点大致位于中心且像素是正方形，因此只需估计焦距 f_1^*：

$$f_1^* = \arg\min_{f_1} \sum_{i=0}^{W-1} \sum_{j=0}^{H-1} C_{1,1,i,j} \left(\frac{i' - f_1 X_{1,1,i,j,0}}{X_{1,1,i,j,2}}, \frac{j' - f_1 X_{1,1,i,j,1}}{X_{1,1,i,j,2}} \right),$$

其中 $i' = i - \frac{W}{2}$ 和 $j' = j - \frac{H}{2}$。快速迭代求解器，例如基于 Weiszfeld 算法的求解器，可以在几次迭代中找到最优 f_1^*。对于第二个相机的焦距 f_2^*，最简单的方法是对图像对 (I_2, I_1) 进行推断，并用上述公式使用 $X_{2,2}$ 代替 $X_{1,1}$。

相对姿态估计 (Relative Pose Estimation)

相对姿态估计可以通过几种方式实现。一种方法是进行 2D 匹配并恢复内参，如上所述，然后估计极线矩阵并恢复相对姿态。另一种更直接的方法是使用 Procrustes 对齐来比较点图 $X_{1,1} \leftrightarrow X_{1,2}$（或等效地，$X_{2,2} \leftrightarrow X_{1,2}$），以获得缩放的相对姿态 $P^* = \sigma^*[R^*|t^*]$：

$$P^* = \arg\min_{\sigma,R,t} \sum_i C_{1,1,i} C_{1,2,i} \|\sigma(RX_{1,1,i} + t) - X_{1,2,i}\|^2,$$

这可以通过闭式解实现。然而，Procrustes 对齐对噪声和离群值敏感。一种更稳健的解决方案是依赖带有 PnP 的 RANSAC。

绝对姿态估计 (Absolute Pose Estimation)

绝对姿态估计，也称为视觉定位，同样可以通过几种不同方式实现。设 I_Q 为查询图像，I_B 为参考图像，已知其 2D-3D 对应关系。首先，可以从 $X_{Q,Q}$ 估计 I_Q 的内参，如前所述。然后，一种可能性是从 I_Q 和某个 I_B 之间的 2D 像素对应关系运行 PnP-RANSAC，从而为 I_Q 获得 2D-3D 对应关系。另一种解决方案是按照前述方法获得 I_Q 和 I_B 之间的相对姿态。然后，我们通过适当缩放将其转换为世界坐标，根据 $X_{B,B}$ 和 I_B 的地面真值点图之间的尺度关系。

4 复现细节

4.1 与已有开源代码对比

本次复现主要参考 github 上 DUSt3R 的开源代码，链接为：https://github.com/naver/dust3r。本次复现工作主要是希望将该三维重建技术应用到医学影像的牙齿重建上。用这个方法，可以实现从图像到三维模型端到端的转换。主要任务是输入一系列不加任何约束的图片，获得这些图片重建后的三维点云。下游任务主要有点的匹配、相机内参估计、相对姿态估计和绝对姿态估计/视觉定位。

不像传统的三维建模，这个模型是一个端到端的输出，直接从图片到点云。在传统的三维建模中，往往分为相机位姿估计（SFM）和稠密重建（MVS）两个部分，在相机位姿估计的部分尤其复杂，过程繁琐且不易出结果，对照片质量、拍摄条件要求极高，故不能大范围地使用，易用性差。归纳来说有以下三点劣势：

1. 过程冗余：每个子问题之间缺乏沟通，关键步骤容易中断

2. 要求严格：MVS 算法的好坏取决于输入图像的质量和相机参数

3. 效果欠佳：每个子问题都不能完美解决，且会给下一步添加噪音

而在 DUSt3R 中，这些问题迎刃而解，因为它对照片的输入没有任何约束，直接跳过繁琐的相机位姿估计步骤做重建，非常地简单、灵活、通用。优势如下：

1. 简单：无需相机位姿、内参等校准信息，适用于不同条件下的三维重建

2. 灵活：将重建问题视为点云回归，统一单目和双目重建情况

3. 通用：可处理多种下游任务，恢复相机位姿、内参及深度信息

4.2 实验环境搭建

环境配置 (Environment Configuration)

安装 Python 3.11 及其依赖库，并通过以下命令创建并激活 Conda 虚拟环境：

```
conda create -n dust3r python=3.11 cmake=3.14.0
conda activate dust3r
```

接着，安装 PyTorch 和其他必要的依赖项：

```
conda install pytorch torchvision pytorch-cuda=12.1 -c pytorch -c nvidia
pip install -r requirements.txt
pip install -r requirements_optional.txt
```

为了加速 RoPE 的运行时，需要编译了 CUDA 内核：

```
cd croco/models/curope/
python setup.py build_ext --inplace
cd ../../../
```

数据准备 (Data Preparation)

下载并解压所需的预训练模型和数据集。使用 Hugging Face Hub 自动下载模型，并放置在 `checkpoints/` 目录下。

```
mkdir -p checkpoints/
wget https://download.europe.naverlabs.com/ComputerVision/DUSt3R/
  DUSt3R_ViTLarge_BaseDecoder_512_dpt.pth -P checkpoints/
```

对于训练数据集，从各自官方资源下载，并按照 `datasets_preprocess` 目录下的脚本进行预处理。具体步骤如下：

1. 下载 CO3Dv2 数据集子集并进行预处理：

```
    mkdir -p data/co3d_subset
    cd data/co3d_subset
    git clone https://github.com/facebookresearch/co3d
    cd co3d
```

```
    python3 ./co3d/download_dataset.py --download_folder ../ --
        single_sequence_subset
    rm ../*.zip
    cd ../../..
    python3 datasets_preprocess/preprocess_co3d.py --co3d_dir
        data/co3d_subset --output_dir data/co3d_subset_processed
        --single_sequence_subset
```

2. 下载预训练的 CroCo v2 检查点：

```
    mkdir -p checkpoints/
    wget https://download.europe.naverlabs.com/ComputerVision/
        CroCo/CroCo_V2_ViTLarge_BaseDecoder.pth -P checkpoints/
```

Docker 支持 (Docker Support)

使用 Docker 来简化环境设置，这确保了实验环境完全一致，避免因环境差异导致的结果不可复现问题。具体步骤如下：

1. 安装 Docker 和 NVIDIA Docker Toolkit。

2. 构建 Docker 镜像并运行容器：

```
    cd docker
    bash run.sh --with-cuda --model_name="
        DUSt3R_ViTLarge_BaseDecoder_512_dpt"
```

4.3 创新点

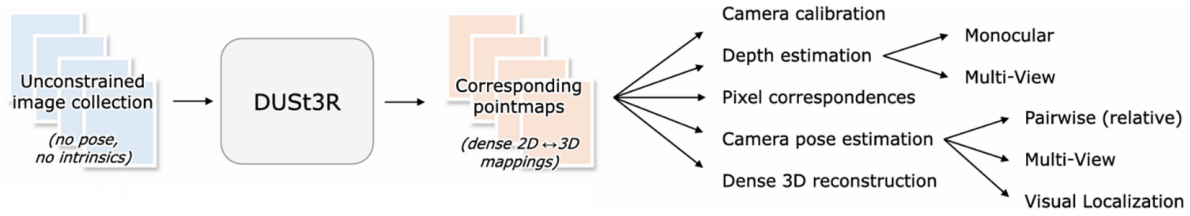

图 3. 工作流示意图

这个框架结构比较简单，由两个 ViT 编码器（共享权重，实际上是一个编码器），两个 transformer 解码器和两个回归头组成。输入两帧 RGB 图像，分割成 patch 后进行编码，然后回归出点云图和置信度图。值得注意的是第一帧图像的输出结果是根据自己本身的坐标系建立的，而第二帧图像的输出结果则是建立在第一帧图像的坐标系上。换而言之，两帧图像的输出都是基于第一帧坐标系，以此保持重建的一致性。

其中置信度图指的是每个点预测的可信程度，举个例子，像透明的地方例如天空，远处的景观这样的事物预测准确度往往比较低，比较容易出错，这时候置信度就会下降。这里的

置信度用于计算损失时作为考量标准，置信度低的点，惩罚比较小，损失比较小，使得结果更加鲁棒。

Datasets	Type	N Pairs
Habitat [103]	Indoor / Synthetic	1000k
CO3Dv2 [93]	Object-centric	941k
ScanNet++ [165]	Indoor / Real	224k
ArkitScenes [25]	Indoor / Real	2040k
Static Thing 3D [68]	Object / Synthetic	337k
MegaDepth [55]	Outdoor / Real	1761k
BlendedMVS [161]	Outdoor / Synthetic	1062k
Waymo [121]	Outdoor / Real	1100k

图 4. 数据集

这个模型有两个成功的关键之处，一是模型结构，二是数据集的选取。在模型结构上，它两帧图片共用一个编码器，共享权重提取了特征，而在解码部分使用了两个解码器，分别做任务，这其实是一个隐式的 matching 过程。第一帧图片在做解码的时候，参考了第二帧图片的信息输出，侧重点在于获取更多信息使得点云输出更准确；而第二帧图片解码时也参考了第一帧的信息，但侧重点更多在于如何根据第一帧图片的坐标系来准确输出。在数据集的选取上，这个工作获取了千万量级图片对的数据集，主要是室内、室外的建筑场景，难度跨越比较大，这使得模型在各种场景下都能推理出比较好的结果。

除了双目重建外，这个模型在单目重建和多视图重建中也同样适用。在单目重建中，只需要输入两帧一模一样的照片也可以获得比较好的重建效果，由此统一了单目与双目重建的任务，为后续研究提供了一个可行的方向。而多视图重建任务中，首先将图片两两分组，在各自的局部坐标系下重建，完成后再做全局对齐到第一帧坐标系下。

以上便是这个模型的主要任务。而下游任务包含了点的匹配、相机内参估计、相对姿态估计和绝对姿态估计/视觉定位等。正如 1.1 所述，使用这个模型可以达到端到端的效果，省去了相机内参估计等中间结果。尽管如此，我们可以根据点云和图片的关系反推出这些中间结果，也就是下游任务。

5 实验结果分析

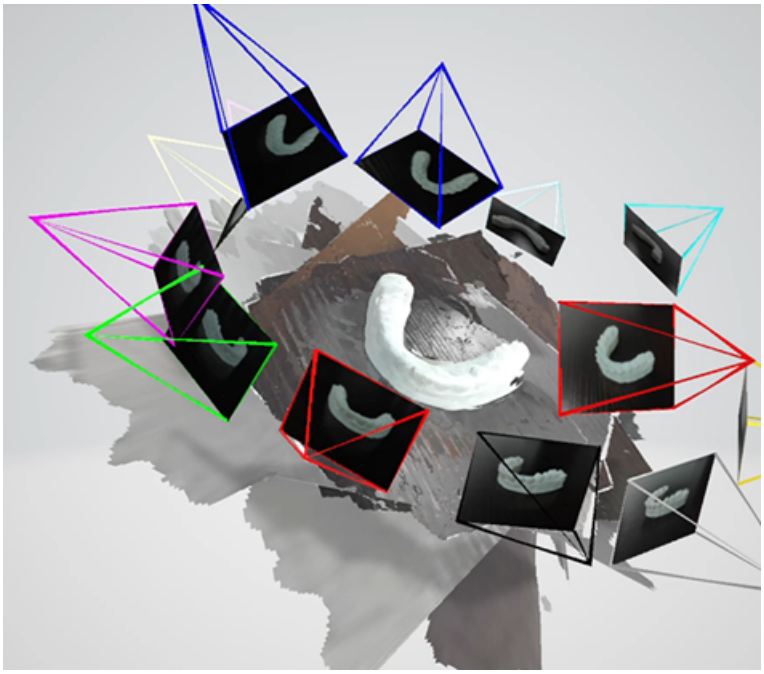

图 5. 实验结果示意 1

上图是输入 10 张不同视角的牙模图像建成的三维模型，除了模型外还有相机位姿估计的结果。

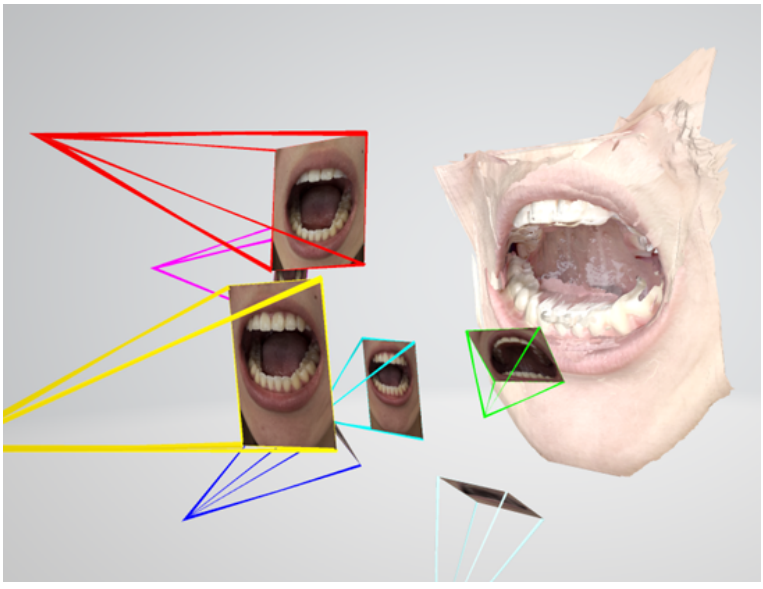

图 6. 实验结果示意 2

上图是输入 8 张不同视角的张嘴图像建成的三维模型，除了模型外还有相机位姿估计的结果。具体的输出可看附件。由以上结果可以看出，建成的模型结果不算特别精准，远远没达到落地应用的标准，还有非常大的改进空间。针对这个结果，上牙的重建效果比下牙要好，在改进上：

1. 针对输入：拍摄时尽量保持静止，光线充足，最好戴上开口器。

2. 针对模型：可以参考后续的研究，例如 MASt3R [4]，对输入的图片进行遮掩并增设回归头，使得效果提升 30%；亦或是 Spann3R [5]，任务预测每帧的点云图，从而无需进行全局对齐，提升推理速度。

3. 针对训练过程：更换训练数据集，由于本人的研究方向是牙齿重建，与该模型的数据集（室内外建筑）有较大差距，对重建结果影响较大，后续研究中我会尽可能收集与人体，尤其是头部、牙齿相关的数据重新训练模型。

6 总结与展望

DUSt3R 是一种全新的密集且不受约束的立体 3D 重建方法，适用于任意图像集合，无需相机校准或视角姿态先验信息。该方法通过将成对重建问题视为点图回归问题，简化了几何约束，并统一了单目和双目重建案例。对于多张图片的情况，DUSt3R 提出了一种全局对齐策略，将所有成对点图表达在一个共同的参考框架中。

在未来，泛化能力是我提升的重点，通过更多样化的训练数据提升模型的适应性，以此应用到我的研究领域——牙齿上；除此以外，我也会继续探索是否有更精巧的模型架构设计，来降低模型训练及推理成本，提升模型输出的质量。

参考文献

[1] David Crandall, Andrew Owens, Noah Snavely, and Daniel Huttenlocher. SfM with MRFs: Discrete-continuous opti mization for large-scale structure from motion. PAMI, 2013.

[2] Hainan Cui, Xiang Gao, Shuhan Shen, and Zhanyi Hu. Hsfm: Hybrid structure-from-motion. In Proceedings of the IEEE conference on computer vision and pattern recog nition, 2017.

[3] Wang S, Leroy V, Cabon Y, et al. Dust3r: Geometric 3d vision made easy[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 20697-20709.

[4] Zhang J, Herrmann C, Hur J, et al. Monst3r: A simple approach for estimating geometry in the presence of motion[J]. arXiv preprint arXiv:2410.03825, 2024.

[5] Leroy V, Cabon Y, Revaud J. Grounding Image Matching in 3D with MASt3R[J]. arXiv preprint arXiv:2406.09756, 2024.