动画视频引导的正畸过程三维牙颌模型重建

梁亚倩[1)] 尤敬严[2)] 雷长松[2)] 范译茗[2)] 戴佳佳[2)] 范业莹[2)]
王少烽[3)] 白玉兴[3)] 刘永进[2)]

[1)](太原理工大学计算机科学与技术学院(大数据学院) 太原 030024)
[2)](清华大学计算机系 北京 100084)
[3)](首都医科大学附属北京口腔医院 北京 100081)

摘 要 隐形矫治方案设计的核心在于根据患者口腔内牙齿的初始状态预测正畸目标牙位,并规划牙齿在正畸过程中的完整移动序列。现有的自动排牙方法都依赖于真实的正畸中间过程三维牙颌模型来进行有监督训练,但在临床实践中,三维牙颌模型正畸移动序列数据集难以获取,严重制约智能正畸相关领域的发展。为解决数据缺乏的挑战,本文提出基于动画视频引导的正畸过程三维牙颌模型重建框架。具体来说,通过融合正畸前牙颌模型的三维几何特征与视频中上、下牙弓颌面图的二维视觉特征作为条件信息,引导扩散模型学习颌面图中牙齿位姿的分布,从而生成与视频对应的正畸中间过程三维牙颌模型。针对监督信号缺失问题,本文设计了一种基于可微分渲染的牙齿位姿约束方法,利用三维模型投影图像与视频帧图像之间的结构相似性来实现跨模态弱监督训练。实验结果表明,所提出方法可以有效捕捉帧间牙齿位姿的细微变化,并结合三维牙颌模型的几何先验信息重建出与动画视频一致的三维牙齿移动序列。

关键词 三维牙颌模型重建;正畸移动序列;条件扩散模型;跨模态信息融合;弱监督训练
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2026.01371

Animated Videos-Guided 3D Dental Model Reconstruction During Orthodontic Process

LIANG Ya-Qian[1)] YOU Jing-Yan[2)] LEI Chang-Song[2)] FAN Yi-Ming[2)] DAI Jia-Jia[2)]
FAN Ye-Ying[2)] WANG Shao-Feng[3)] BAI Yu-Xing[3)] LIU Yong-Jin[2)]

[1)](College of Computer Science and Technology (College of Data Science), Taiyuan University of Technology, Taiyuan 030024)
[2)](Department of Computer Science and Technology, Tsinghua University, Beijing 100084)
[3)](Beijing Stomatological Hospital, Capital Medical University, Beijing 100081)

Abstract The core of invisible orthodontic treatment planning is to predict a clinically feasible target tooth arrangement from the initial state of the patient's dentition and further design a smooth and biologically plausible tooth movement sequence that gradually transitions from malocclusion toward the planned outcome. In recent years, learning-based tooth arrangement methods have achieved encouraging progress. However, most existing pipelines still depend on dense supervision from real 3D dental model sequences. In routine clinical practice, such sequential

收稿日期:2025-06-24;在线发布日期:2026-03-05。本课题得到北京市自然科学基金-海淀原始创新联合基金(No.L222008)、北京市医院管理中心"扬帆"计划临床技术创新项目(No.ZLRK202330)、国家自然科学基金青年科学基金项目(No.62502337)资助。**梁亚倩**,博士,讲师,主要研究领域为计算机图形学、三维网格模型几何分析、深度学习。E-mail:liangyaqian@tyut.edu.cn。**尤敬严**,本科生,主要研究领域为深度学习。**雷长松**,博士研究生,主要研究领域为三维几何分析、深度学习。**范译茗**,本科生,主要研究领域为深度学习。**戴佳佳**,博士,主要研究领域为计算机图形学、三维网格模型几何分析、深度学习。**范业莹**(通信作者),博士,主要研究领域为三维模型处理、医学图像处理。E-mail:fyydemail@gmail.com。**王少烽**,博士,主要研究领域为口腔正畸学、人工智能辅助正畸诊疗。**白玉兴**,博士,教授,主要研究领域为口腔正畸学、人工智能辅助正畸诊疗。**刘永进**(通信作者),博士,教授,中国计算机学会(CCF)高级会员,主要研究领域为计算机图形学、可视媒体计算。E-mail:liuyongjin@tsinghua.edu.cn。

intraoral scans are rarely available due to the high cost of repeated acquisition, limited scanning frequency, inconsistent patient compliance, and other practical constraints. Consequently, the scarcity of paired and temporally ordered 3D dental models has become a critical bottleneck that hinders the deployment of intelligent orthodontic technologies and applications. To address this challenge of data scarcity, this paper proposes a an animated video-guided 3D dental model reconstruction framework for reconstructing dental model sequences throughout the orthodontic process. Rather than requiring real 3D dental model sequences as ground truth, the proposed framework leverages readily accessible orthodontic animation videos as a source of weak supervision, since such videos explicitly depict progressive tooth movements and provide rich temporal cues about how tooth poses evolve over time. The proposed method formulates 3D dental model reconstruction as conditional distribution learning over per-tooth rigid transformation matrices and optimizes it by introducing a diffusion probability model (DPM). The diffusion model operates in the space of tooth transformation matrices (e. g. , rotations and translations for each tooth) and learns to denoise from a Gaussian prior to plausible transformation sets conditioned on multimodal inputs. By integrating the tooth pose information contained in the top-down views of dental arches and the geometric prior information of the pre-orthodontic 3D dental model as conditions, the diffusion model is guided to generate 3D dental models movement sequences corresponding to video frames. To stabilize learning under weak supervision, the DPM is trained in a two-stage strategy. In the first stage, the model is trained to predict the post-orthodontic 3D dental models, which provides a strong initialization and equips the network with the fundamental ability to infer per-tooth rigid transformation matrices from images. In the second stage, the trained diffusion model is fine-tuned for intermediate reconstruction by conditioning on the combination of 2D feature encodings of the dental arch in the video and the 3D feature encodings of pre-orthodontic 3D models. This fine-tuning stage encourages the model to generate a temporally coherent sequence of intermediate states, where adjacent frames correspond to small pose updates, and the overall trajectory forms a continuous transition from the initial malocclusion to the predicted final arrangement. Because real 3D dental model sequences are not available as supervision, a differentiable rendering-based tooth pose constraint loss is further introduced to provide cross-modal weak supervision. Concretely, the generated 3D dental model is orthogonally projected under a differentiable renderer, producing 2D projections that can be compared directly with the corresponding animation frames. Then, the model is optimized by maximizing structural similarity between the rendered projections and the video frames, where the objective supplies frame-wise supervisory signals without requiring intermediate 3D labels. Experimental results show that the proposed method can effectively capture subtle inter-frame variations in tooth pose and reconstruct smooth, realistic tooth movement sequences consistent with animation videos, highlighting its potential to alleviate supervision bottlenecks and facilitate more scalable learning-based orthodontic planning.

Keywords 3D dental model reconstruction; orthodontic movement sequence; conditional diffusion model; cross modal information fusion; weakly supervised training

1 引 言

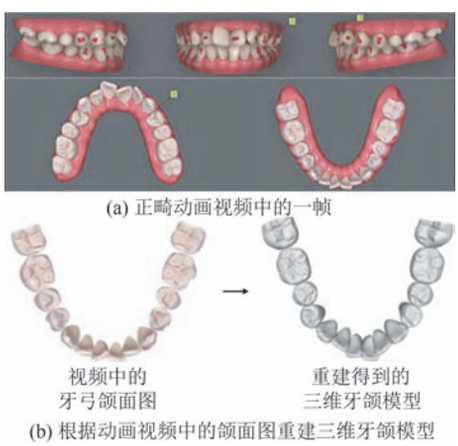
随着三维建模、计算机图形学等相关技术发展，计算机辅助几何设计已在许多自然科学和工程领域发挥了重要作用，包括医学领域的数字正畸。将深度学习方法、三维模型、数字正畸相结合，辅助正畸治疗已成为当前相关领域的热点研究问题。错颌畸

形指口腔内牙齿排列不齐或上下牙弓间咬合关系异常。根据世界卫生组织流行病学调查显示，错颌畸形已成为全球三大口腔疾病之一。它不仅会影响患者口内健康（造成如龋齿或牙周组织损伤等疾病）、降低面部美观，更会导致脊柱扭曲、消化疾病等严重的全身性疾病问题。正畸治疗是解决错颌畸形的主要手段，其核心机制是借助矫治装置施加持续、可控的矫治力，引导牙齿移动，最终实现牙列生理性复位及咬合功能重建。

在隐形矫治诊断中，一个关键环节是根据患者口腔内牙齿的初始状态设计正畸目标牙位以及牙齿在正畸过程中的完整移动序列，即牙齿如何从初始状态逐步移动到目标位。在当前临床实践中，通常先由技师根据正畸方案手动设计牙齿移动序列，之后与牙医进行多轮沟通与调整，最终共同确认牙齿移动方案[1]。这种人工设计的方式不仅耗时费力且存在主观偏差，对正畸医生和技师的专业技能、审美要求也极高。为突破这一技术瓶颈，自动排牙方法已成为口腔数字化诊疗领域的重点研究方向。这类方法不仅能够辅助正畸医生快速制定牙齿移动方案，评估牙齿运动的合理性，还可以为隐形牙套的生产提供数据支持。

起初，相关研究人员将牙齿移动过程建模为带约束的优化问题，通过构建包含医疗规则（如牙齿之间不能碰撞）的目标函数，利用群智能优化算法来求解全局最优移动路径[2-4]。然而，很多临床正畸规则（如尖牙优先远移原则、后牙垂直向控制策略等）无法建模为数学公式，导致传统优化方法生成的牙齿移动序列无法应用于临床正畸治疗中。随着人工智能技术的发展，近年来一些基于深度学习的自动排牙方法被提出。它们通过从现有病例中学习牙齿的移动模式，能够自动预测正畸目标牙位[5-8]或牙齿正畸移动过程[9-10]，为正畸治疗提供了新思路。尽管数据驱动的方法在理论上具有显著优势，但在实际应用中仍面临数据不足的挑战。现有的公开数据集仅包含正畸前后的静态三维牙颌模型，尚缺乏能覆盖完整治疗过程的三维牙颌模型移动序列数据。这一数据缺口导致预测正畸中间过程的神经网络难以获得有效监督，严重限制了该领域研究工作的发展。

受限于现实条件，患者在正畸治疗过程中无法多次接受三维口扫，导致难以直接从患者处采集整个正畸过程的动态三维数据。与之相比，正畸动画视频更容易获取及处理。技师在设计牙齿移动序列时，会为患者制作多视角牙齿移动模拟视频，该视频能够忠实反映满足医师要求的正畸过程中患者牙齿的完整移动过程。如图1所示，每帧图像通过五个正交视角来展示单步牙齿位移，其中第一行从左到右分别为左侧咬合图、正视图、右侧咬合图，第二行从左到右为上牙弓颌面图与下牙弓颌面图。为解决数据缺乏问题，本文提出利用患者正畸前三维牙颌模型以及较为容易获取的正畸动画视频共同作为先验信息，重建与动画视频对应的正畸中间过程的三维牙颌模型。尽管正畸动画视频完整地记录了牙齿从初始位置到正畸目标位的运动轨迹，但受限于其分辨率较低且缺乏真实纹理特征，加之单颌视角覆盖不足（仅包含四个有效视角）以及上颌对下颌的严重遮挡等问题，直接从视频帧中提取多视角信息进行三维重建难以获得精确的牙颌模型。为了有效地挖掘动画视频中的信息，本文提出构建基于扩散模型（Diffusion Probabilistic Model，DPM）[11]的正畸中间过程三维牙颌模型重建方法，通过融合牙弓颌面图（即俯视图）中包含的牙齿位姿信息以及正畸前三维牙颌模型的几何先验信息作为条件，引导扩散模型生成与视频帧对应的中间过程三维牙颌模型。



(a) 正畸动画视频中的一帧

(b) 根据动画视频中的颌面图重建三维牙颌模型

图1　本文方法的核心思路

具体来说，本文构建的扩散模型包含两个阶段的训练过程，第一阶段预测正畸后的三维牙颌模型，使扩散模型具有根据图片预测牙齿刚性变换矩阵的基本能力；第二阶段将训练好的扩散模型进行微调，利用视频中牙弓颌面图的特征编码与正畸前三维牙颌模型的特征编码共同作为提示条件，引导扩散模型重建与图片对应的正畸中间过程三维牙颌模型。本文方法的整体框架如图2所示，其中扩散模型预测的是每颗牙齿的刚性变换矩阵，因此正畸过程中牙齿的形态与大小都能够保持完全一致。

本文的创新点如下：(1)提出基于动画视频引导

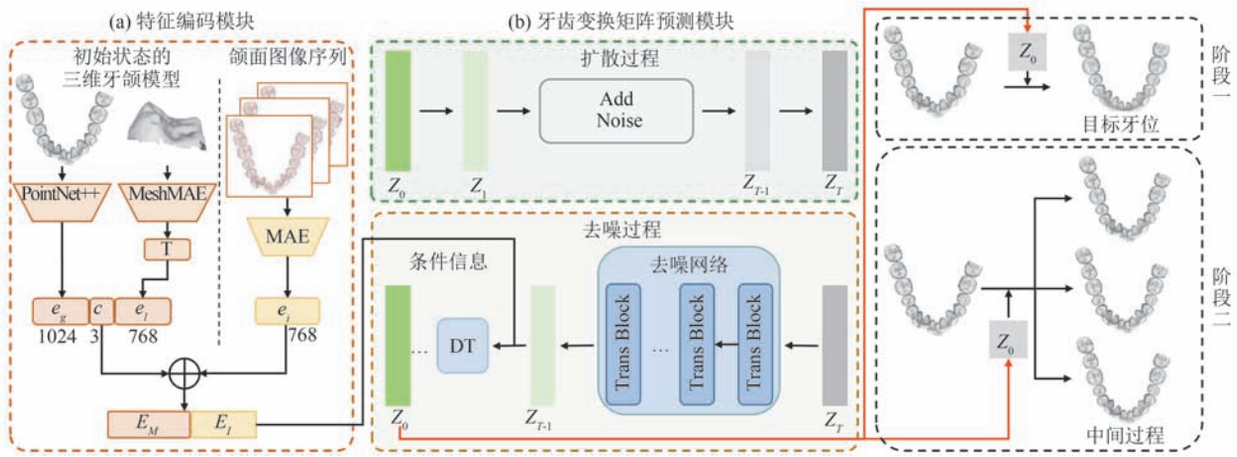


图2 本文方法的总体框架

的正畸过程三维牙颌模型重建框架。通过融合正畸动画视频中的牙齿位姿信息与三维牙颌模型的几何先验信息，实现正畸过程三维牙颌模型移动序列的高精度重建。(2)提出牙齿移动扩散模型。首先，模型通过逐步去噪学习牙齿从错颌畸形到正常咬合的变换矩阵的分布，从而具备根据图像预测牙齿刚性变换矩阵的基本能力。随后对训练完成的扩散模型进行微调，引入动画视频中牙弓颌面图和初始三维牙颌模型的特征编码作为条件提示，引导扩散模型学习目标图像中牙齿位姿的分布，并重建与之对应的三维牙颌模型。(3)提出一种弱监督训练策略。针对缺乏真实正畸过程三维牙颌模型数据的问题，设计基于可微分渲染的跨模态约束方法。通过比较三维牙颌模型渲染图像与视频帧中的牙弓颌面图，计算跨模态损失函数，从而解决监督信号缺失的训练难题。本文所提出方法为构建三维牙颌模型移动序列数据集提供了新的可行路径，不仅有助于弥补数据缺口，也为未来推动正畸智能化提供了重要技术参考。

2 相关工作

2.1 三维牙颌模型正畸过程预测

三维牙颌模型正畸过程预测是数字化正畸的重要环节之一，旨在根据患者的初始牙颌模型和正畸目标牙位规划出完整的牙齿移动序列，从而辅助医生高效制定个性化的正畸治疗方案，同时直观展示治疗预期效果，增强患者的理解与信心。近年来，一些基于神经网络的正畸过程预测方法[5,12]被提出，它们通常采用自回归的方式生成牙齿移动序列。然而，这些方法直接使用具有大量参数的牙齿点云数据，导致牙齿移动步长有限且误差积累。在此基础上，Fan等[9]提出协同牙齿运动扩散模型，将正畸牙齿移动规划建模为一个扩散过程，通过图结构整合牙齿间的咬合关系，以增强对多牙齿移动分布的学习。Ma等[10]提出利用Transformer直接预测自适应长度的牙齿移动序列，以提升网络的灵活性。尽管上述方法在建模策略上各具创新，但它们普遍依赖大量真实的正畸过程三维牙颌模型数据作为监督信号来训练神经网络。然而，目前并没有公开的正畸过程三维牙颌模型数据集，上述方法的有效性和泛化性难以验证。

2.2 三维牙颌模型重建

随着三维建模技术的持续发展，如何重建三维牙颌模型在口腔正畸领域成为了研究热点。本节将介绍利用五个视角的口内照片进行三维牙颌模型重建的相关研究工作。

传统的多视图重建方法[13]和新兴的神经隐式场方法[14-17]虽然在通用的三维重建任务中表现优异，但由于正交视角下牙齿图像高度稀疏且几乎没有重叠区域，基于运动恢复结构(Structure from Motion，SFM)算法[18]难以估计正确的相机位姿，导致现有依赖精确相机位姿的三维重建方法无法实现高精度的三维牙颌模型重建。Johannes等[19]提出一种参数化的三维牙颌模型重建方法，利用牙齿轮廓信息来调整重建结果中的牙齿形态；尽管该方法能够生成解剖结构合理的牙齿模型，但其对预定义模板的依赖性较强，严重制约了个性化细节的表征能力。针对视角稀疏性问题，近期研究尝试利用生成模型来突破视角限制[20-25]。Xu等提出的TeethDreamer[26]融合了Zero123预训练模型[20]的先验信息，根据输入数据从固定视角合成新视角图像并基于Neus[14]进行三维

牙颌模型重建。但此类方法合成图像的几何一致性难以保证，导致非可见区域重建误差累积；此外，生成的三维牙颌模型是否符合医学要求仍需讨论。

2.3　扩散模型

扩散概率模型是一类基于马尔可夫链的生成模型，其核心过程包含两个部分：前向扩散过程通过逐步向数据注入高斯噪声来破坏原始分布，使数据分布在多步演化后逼近各向同性高斯噪声；逆向去噪过程则通过训练神经网络学习逐步恢复数据的转换规律，从而最终实现从随机噪声逐步生成目标数据。这种渐进式的生成方式使模型能够学习复杂的数据分布。具体而言，设x_0是一个具有未知分布的D维随机变量，DPM的前向扩散过程中通过高斯扰动来逐步破坏x_0的信息。对于任意时间，其转移概率分布[27]满足如下公式：

$$q_{0t}(x_t|x_0)=\mathcal{N}(x_t;\ \alpha_t x_0, \sigma_t^2 I) \tag{1}$$

其中，α_t和$\sigma_t>0$是t的可微函数。

在训练方面，Sohl-Dickstein等最早从非平衡热力学视角提出离散时间扩散链，并基于变分下界进行优化[28]。在此基础上，Ho等提出DDPM[11]，通过噪声重参数化将训练目标转化为更易优化的噪声预测形式，显著提升训练稳定性并为后续研究奠定了主流的训练范式。随后，LDM[29]将扩散迁移到潜空间，在保持生成质量的同时显著降低训练与采样成本，从而推动大规模条件生成方法的研究。围绕可控与交互式生成，InstructPix2Pix[30]将语言指令引入图像编辑；ControlNet[31]则将边缘、深度、分割等结构信息作为空间条件稳定注入预训练文本扩散模型，从而显著增强可控性与复用性。在模型结构上，DiT[32]以Transformer替代U-Net并展现出良好的可扩展性。

在应用层面，扩散模型已从二维图像生成扩展到更广泛的视觉与多模态场景，其在高保真图像生成[33]、超分辨率重建[34]、图像到图像翻译与编辑[35]等任务上持续刷新效果，并逐渐成为内容生成的重要技术路线。同时，扩散模型也深度了参与三维视觉建模，如点云/网格生成[36]、文本到三维生成[37]等方向，为三维内容生成与几何学习提供了新的统一框架与可扩展范式。此外，在视频建模领域，扩散模型通过在时空维度上进行一致性建模，推动文本到视频生成[38]、视频预测与视频编辑[39]等任务取得突破性进展，体现出其在复杂时序数据生成中的潜力。这些进展共同表明，扩散模型正在成为通用的生成式建模工具。

3　方　法

本文提出利用扩散模型来学习正畸动画视频帧中牙齿的位姿分布，从而实现正畸中间过程三维牙颌模型的重建。为保证在正畸过程中牙齿的大小以及形态完全一致，本文将正畸中间过程三维牙颌模型的重建问题转换为中间过程牙齿位姿估计问题。通过将预测牙齿的刚性变换参数应用到正畸前三维牙颌模型上，可以得到正畸后的三维牙颌模型以及正畸中间过程中的三维牙颌模型。

3.1　网络整体架构

本文方法的整体网络架构如图2所示，由特征编码模块和牙齿变换矩阵预测模块构成。在特征编码模块中，分别提取正畸前三维牙颌模型的几何特征编码以及牙弓颌面图像序列的二维特征编码。针对正畸前三维牙颌模型，首先采样离散点云并利用PointNet++[40]提取全局几何特征，随后采用MeshMAE[41]对单颗牙齿进行局部特征提取，这种多尺度特征提取策略可以有效捕捉三维牙颌模型的几何先验信息。针对牙弓颌面图像序列，利用MAE[42]自编码器进行二维特征编码。在牙齿变换矩阵预测模块中，采用基于Transformer架构的扩散模型框架，将特征编码作为条件信息，引导扩散模型逐步去噪来生成正畸后三维牙颌模型和正畸中间过程的三维牙颌模型。本文所提出的牙齿变换矩阵预测模块分为两个阶段：阶段一将正畸前牙颌模型的三维特征编码以及正畸后牙弓颌面图像的二维特征编码作为条件信息，生成正畸后三维牙颌模型；阶段二将正畸前牙颌模型的三维特征编码以及正畸中间过程牙弓颌面图像的二维特征编码作为条件信息，生成正畸中间过程的三维牙颌模型。

3.2　特征编码模块

3.2.1　三维牙颌模型特征提取

给定正畸前的三维牙颌模型网格$M_{pre}=\{m_{k_{pre}}\}$，其中，$m_{k_{pre}}$表示第$k(1\leqslant k\leqslant K)$个牙齿的独立网格模型，$K$指三维牙颌模型中牙齿的数量。为了充分挖掘三维牙颌模型蕴含的几何先验信息，本文提出从整体牙颌模型中提取全局几何表征，并从单颗牙齿中提取局部细节信息。

三维牙颌模型中通常包含大量的顶点和面片，直接进行全局特征提取计算代价较高。因此，本文提出在三维牙颌模型中采样离散点云数据作为输

入。具体来说，首先利用最远点采样法从整个三维牙颌模型中采样离散点云数据，其中每颗牙齿均匀保留256个顶点，最终形成维度为[14，256]的牙颌点云模型P_{pre}（当牙齿数量不足14时，用0补齐）。随后，利用常用的点云处理网络PointNet++作为全局特征提取网络来提取P_{pre}的全局特征编码e_g。

相比于离散的点云数据，三维网格模型能够提供更为丰富的几何细节信息。因此，本文提出直接利用牙齿的三维网格模型$m_{k_{pre}}$作为输入来提取单颗牙齿的局部几何特征。然而，由于网格模型存在不规则的拓扑结构，难以直接使用标准的神经网络进行处理；此外，现有面向网格数据的深度学习方法大多仅适用于简单三维模型，难以有效处理细节丰富的牙齿网格模型。针对这一问题，本文提出引入MeshMAE对牙齿网格模型进行特征提取，并通过重新网格化操作，使MeshMAE中的Transformer架构能够高效处理牙齿网格数据。具体来说，首先将单颗牙齿模型简化至a个面得到基础网格，再对基础网格中的每个面片执行b次Loop细分操作，从而在保持整体形状外观一致的前提下构建牙齿模型的层次化结构；其次，将来自基础网格中同一个面的4^b个细分面划分到一个块中，通过聚合块中三角面片的几何特征（包含面积、三个内角角度、法向量以及法向量和三个顶点法向量的内积）形成块的特征矩阵；将所有块的特征矩阵有序拼接后，输入MeshMAE网络进行特征学习。针对有标签数据不足的问题，本文将引入基于掩码自编码机制的预训练策略。在预训练过程中，在送入编码器之前随机遮挡每个牙齿模型50%的块特征，经过编码器和解码器后网络预测被遮挡部分的顶点坐标来重建牙齿的几何结构；随后，利用预训练后的MeshMAE编码器提取单颗牙齿模型的局部几何特征e_{l_k}。同时，利用多层感知机（Multilayer Perceptron，MLP）对单颗牙齿的三维几何中心坐标进行位置编码，得到位置编码c_k。

最后，将三维牙颌模型的全局特征编码e_g、单颗牙齿的局部特征编码e_{l_k}与位置编码c_k按照牙齿序列进行拼接，构建三维几何特征编码：

$$E_{M_{pre}}=\mathrm{Stack}\{\mathrm{Concat}(e_g,c_k,e_{l_k})|k\in K\} \quad (2)$$

3.2.2 牙弓颌面图像特征提取

给定正畸动画视频，首先从每帧视频中截取上、下牙弓的颌面图像（如图1中第二行所示），随后采用预训练SAM（Segment Anything）模型[43]进行区域分割。根据亮度阈值与区域面积筛选、剔除牙龈等非牙齿部分，并将独立的牙齿图像进行合并，从而得到去除牙龈后的牙弓颌面图像；按时序组织牙弓颌面图像，构成颌面图像序列$\{I_n|n\in N\}$，其中N表示视频总帧数，即视频序列长度。随后，采用MAE对牙弓颌面图像进行特征提取。具体的，将预处理后的颌面图像I_n划分为16×16的像素块作为网络的输入，其中随机遮挡50%的像素块，将可见块的像素送入编码器进行隐空间特征编码；之后，解码器从隐向量中重建被遮挡部分的像素，从而恢复原始图片信息。预训练结束后，仅使用MAE的编码器来提取颌面图像的二维特征编码E_I。

3.3 牙齿变换矩阵预测模块

在完成牙颌模型的特征提取后，本文设计基于U-ViT架构[44]的牙齿移动扩散模型，主要包含以下两个训练阶段：阶段一预测正畸后的三维牙颌模型，使扩散模型具有根据颌面图像预测牙齿刚性变换矩阵的基本能力；阶段二对扩散模型进行微调，使其能够重建出与颌面图像对应的正畸中间过程三维牙颌模型。

3.3.1 正畸目标牙位预测

本文采用通用的三维点云配准方法，迭代最近点（Iterative Closest Point，ICP）方法[45]，对正畸前、后的三维牙颌模型进行配准以获得真实的牙齿刚性变换参数。在训练阶段，将每颗牙齿从初始状态到目标牙位的真实刚性变换矩阵表示为$\boldsymbol{z}_{0_k}=(\boldsymbol{t}_{0_k},\boldsymbol{r}_{0_k})$，其中，$\boldsymbol{t}_{0_k}\in\mathbb{R}^3$和$\boldsymbol{r}_{0_k}\in\mathbb{R}^3$分别指第$k$个牙齿的平移和旋转向量。将牙列中所有牙齿的变换参数拼接后得到的矩阵$\boldsymbol{z}_0\in\mathbb{R}^{|K|\times6}$作为扩散模型的输入。在扩散过程中，对$\boldsymbol{z}_0$实施$t$步加噪操作，逐步生成噪声扰动后的矩阵$\boldsymbol{z}_t$：

$$\boldsymbol{z}_t=\alpha_t\boldsymbol{z}_0+\sigma_t\delta \quad (3)$$

其中，$\delta\sim\mathcal{N}(\mathbf{0},\mathbf{I})$，而$\alpha_t$和$\sigma_t$如公式(1)中所定义。

在阶段一的训练过程中，将正畸前三维牙颌模型的几何特征编码$E_{M_{pre}}$以及正畸后牙弓颌面图像的二维特征编码$E_{I_{post}}$拼接后作为条件信息$E_1=[E_{M_{pre}},E_{I_{post}}]$，引导扩散模型预测牙齿从初始状态到目标牙位的刚性变换矩阵，此过程表示为$z_\theta(z_t,t,E_1)$，θ表示网络中的参数。网络的输出可以表示为$\bar{\boldsymbol{z}}_0=\{\bar{\boldsymbol{z}}_{0_k}|k\in K\}\in\mathbb{R}^{|K|\times6}$，其中$\bar{\boldsymbol{z}}_{0_k}=(\boldsymbol{t}_{0_k},\boldsymbol{r}_{0_k})$表示单颗牙齿$m_{k_{pre}}$的变换矩阵。最后，将变换矩阵$\bar{\boldsymbol{z}}_0$施加到初始三维牙颌模型$M_{pre}$上得到预测的三维牙颌模型正畸目标牙位$\bar{M}_{post}$，此过程可表示为

$$\bar{M}_{post}=aligner(M_{pre},\bar{\boldsymbol{z}}_0) \quad (4)$$

3.3.2　正畸中间过程的三维牙颌模型预测

在阶段二的训练过程中，由于缺乏正畸中间过程三维牙颌模型的真实数据，对应的牙齿变换矩阵的真实值也是未知的。因此，本文提出基于视频序列长度构建牙齿变换矩阵的伪真值，以作为扩散模型的输入。

具体来说，利用视频中的牙齿移动序列长度N对z_0（从初始状态到目标牙位的牙齿刚性变换参数）进行线性插值，构造正畸中间过程第$step$步的牙齿变换矩阵伪真值$\boldsymbol{z}_{0_k,step}=\left(\boldsymbol{t}_{0_k}\cdot\frac{step}{N},\boldsymbol{r}_{0_k}\cdot\frac{step}{N}\right)$。为克服线性插值导致的固有偏差，本文将对阶段一训练好的扩散模型进行微调，利用正畸前三维牙颌模型的特征编码$E_{M_{pre}}$与正畸中间过程牙弓颌面图像的二维特征编码$E_{I_{step}}$拼接后作为条件信息$E_2=[E_{M_{pre}},E_{I_{step}}]$，引导扩散模型预测与牙弓颌面图像中牙齿位姿预测对应的刚性变换矩阵$\bar{\boldsymbol{z}}_{0,step}$。

为克服缺乏中间过程三维牙颌模型真实值所导致的网络难训练问题，本文设计了一种基于可微分渲染的弱监督训练策略。具体而言，首先利用正交投影方法将预测的正畸中间过程三维牙颌模型$\bar{M}_{step}$投影为二维图像I_{step}^{syn}，并将其分辨率设置为与正畸动画帧中截取的牙弓颌面图像相同。当预测的三维牙颌模型$\bar{M}_{step}$与正畸中间过程三维牙颌模型位姿一致时，投影图像与动画视频帧中牙颌模型的形态应完全相同。基于这一特性，本文提出以预测结果的投影图像与动画视频帧之间的差异作为监督信号，通过最小化该差异，引导扩散模型逐步学习正畸中间过程中每一步的牙齿刚性变换矩阵$\bar{\boldsymbol{z}}_{0,step}$。最后，将预测得到的牙齿变换矩阵$\bar{\boldsymbol{z}}_{0,step}$施加到初始三维牙颌模型$M_{pre}$上，从而得到预测的正畸中间过程三维牙颌模型$\bar{M}_{step}$。

3.4　损失函数

网络的损失函数包含三个部分。

3.4.1　三维牙颌模型重建损失$\mathcal{L}_{recon}$

本文中采用倒角距离（Chamfer Distance，CD）来衡量预测的三维牙颌模型与真实三维牙颌模型之间的几何重建损失，其具体计算公式为

$$\mathcal{L}_{recon}=\frac{1}{|M_{gt}|}\sum_{x\in M_{gt}}\min_{y\in M_{pred}}\left\|x-y\right\|_2^2+\frac{1}{|M_{pred}|}\sum_{y\in M_{pred}}\min_{x\in M_{gt}}\left\|y-x\right\|_2^2 \tag{5}$$

其中，M_{pred}表示预测的三维牙颌模型，M_{gt}表示真实的三维牙颌模型。

3.4.2　牙齿位姿估计损失$\mathcal{L}_{MSE}$

本文中将直接计算预测的牙齿刚性变换矩阵与真实牙齿刚性变换矩阵之间的L2损失来衡量牙齿位姿估计的误差，其具体计算公式为：

$$\mathcal{L}_{MSE}=\|\boldsymbol{z}_0-\boldsymbol{z}_\theta(z_t,t,E)\|^2 \tag{6}$$

其中，$\boldsymbol{z}_\theta(z_t,t,E)$表示预测的牙齿变换矩阵，$\boldsymbol{z}_0$表示真实的变换矩阵。

3.4.3　牙颌模型渲染图像的结构相似性损失$\mathcal{L}_{SSIM}$

动画视频在渲染牙颌模型时设置了特殊的光照和材质参数，而本文方法仅使用PyTorch3D框架来实现三维牙颌模型到二维图片的几何投影，导致渲染图像与视频帧之间存在较大差距。为消除光照和材质差异带来的影响，本文采用结构相似性指数（Structure Similarity Index Measure，SSIM）来度量图像之间的相似性，其具体计算公式为

$$\mathcal{L}_{SSIM}=\frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)} \tag{7}$$

其中，x和y分别指预测牙颌模型的二维渲染图片以及视频中牙弓颌面图像的像素，μ_x和μ_y指像素均值，σ_x^2和σ_y^2指像素方差，σ_{xy}指两张图像的协方差。

在阶段一的训练中，网络的整体损失函数为

$$\mathcal{L}_1=\mathcal{L}_{recon}+\mathcal{L}_{MSE}+\mathcal{L}_{SSIM} \tag{8}$$

在阶段二的训练中，由于缺乏中间过程三维牙颌模型的真实值，仅利用$\mathcal{L}_{SSIM}$作为损失函数。

4　实　验

4.1　数据集

本文的实验数据来源于在北京口腔医院接受正畸治疗的患者。本文所开展的研究已通过该医院伦理委员会的伦理审核批准，所有患者（包括未成年患者的法定监护人）均签署了知情同意。由于隐形矫治技术的临床特点，大多数患者会经历多个治疗阶段，将每个治疗阶段中三维牙颌模型的初始状态和终止状态作为独立的成对数据。本文收集的数据集中共包含466组数据，每一组数据包含（1）正畸前、后的三维牙颌模型以及（2）与三维牙颌模型对应的正畸过程动画视频。其中，396组数据被分到训练集，其余分到测试集。

4.2　实验设置

本文在配备NVIDIA RTX A6000显卡和Intel Xeon Gold 5320处理器的工作站上开展实验，采用

PyTorch深度学习框架进行算法实现。

在数据预处理阶段，本文参考MeshMAE[41]和TADPM[8]中的设置，首先将每颗牙齿的三维网格模型将先简化至256个面片（即$a=256$），随后再执行3次细分操作（即$b=3$）。该设置在较大程度上保留模型原始几何形状的同时，可有效避免细分后模型带来的过高计算开销。为保持数据形式的一致性，本文从每个牙齿模型中均匀采样256个顶点，以构建三维牙颌模型的离散点云表示。对于截取的牙弓颌面图像，统一调整为固定像素分辨率（460×640），并将背景设置为黑色，同时进行像素值归一化与灰度处理；三维牙颌模型也同样进行归一化处理，以降低尺度差异对训练的影响。

在网络训练过程中，本文利用PyTorch3D框架对三维牙颌模型的预测结果进行二维投影渲染，并使用二分法调整摄像机高度：通过约束初始三维牙颌模型渲染图片中牙齿模型包围盒与颌面图像中牙齿模型包围盒的尺度一致，来确定摄像机高度。

在参数设置中，本文使各编码器配置与其原论文保持一致，其中PointNet++[40]的输出特征维度为1024，MeshMAE[41]的输出特征维度为768，MAE[42]的输出特征维度同样为768；在牙齿变换矩阵预测模块中，去噪网络采用12层Transformer，每层特征维度设置为800。训练阶段使用AdamW优化器更新参数，初始学习率设置为10^{-5}，权重衰减系数为0.05，并采用余弦退火算法调整学习率（周期长度为700个迭代步）。所有实验中批尺寸设置为2，训练共700个周期；整体模型参数量为226M。上述实验设计和超参数设置可以保证模型训练过程的合理性，同时使实验结果具有可重复性。

在推理阶段，本文采用DDIM（Denoising Diffusion Implicit Model）[46]采样器进行加速。生成单个牙颌模型变换矩阵的平均推理时间为6.92秒，处于临床应用可接受范围内。

4.3 比较实验

为验证本文方法的有效性，本节将针对正畸目标牙位的预测效果与正畸中间过程三维牙颌模型的重建效果，分别与当前主流方法进行对比实验分析。

4.3.1 正畸目标牙位预测效果

参考TADPM[8]，本文使用ADD（Average Distance of Model Point）、PA-ADD、CSA（Cosine Similarity Accuracy）和ME_{rot}来评估正畸目标牙位的预测效果。其中，ADD指预测的正畸后三维牙颌模型与真实值之间逐点距离的平均值，PA-ADD指经过模型刚性配准后的ADD指标，CSA指预测的变换矩阵与真实变换矩阵之间的误差，而ME_{rot}指三维牙颌模型旋转矩阵的平均误差。表1中列出了本文方法与TANet[5]、PSTN[6]、TAligNet[7]和TADPM[8]四种先进方法的正畸目标牙位预测结果的定量评估指标。

表1 正畸目标牙位预测效果比较

方法	ADD(↓)	PA-ADD(↓)	CSA(↑)	ME_{rot}(↓)
TANet	1.68	1.52	<u>0.92</u>	8.87
PSTN	1.89	1.73	0.89	8.83
TAligNet	1.80	1.69	0.91	9.06
TADPM	**1.52**	**1.32**	**0.93**	<u>8.79</u>
本文方法	<u>1.60</u>	<u>1.45</u>	0.86	**7.63**

实验结果表明，本文方法在指标上略低于TADPM，但相较于TANet、PSTN和TAligNet均取得了明显提升。值得强调的是，本文方法在阶段一的训练中创新性地引入了渲染图像的结构性相似损失，通过联合优化图像匹配和三维坐标回归双重目标，实现了二维颌面图像与三维牙颌模型的跨模态对齐。这种学习机制虽然没有提升目标牙位的预测效果，但为阶段二基于纯图像相似性约束的牙齿变换矩阵预测提供了重要基础。通过该机制，网络能够在缺乏显式三维几何监督的情况下，完成牙齿姿态变换的端到端学习。图3中展示了正畸目标牙位预测的可视化结果，面对复杂的输入病例，本文方法仍能够实现良好的目标牙位预测。

4.3.2 正畸中间过程三维牙颌模型重建效果

在正畸中间过程的三维牙颌模型重建效果评估中，本文方法将与传统多视角重建方法MVSNet[47]、神经辐射场重建方法Neus[14]、新视角合成方法TeethDreamer[26]进行对比。

由于动画视频中图像的视角覆盖不足，图像之间重叠区域过小，Neus和MVSNet均无法重建出有效的三维模型。而对于TeethDreamer方法，首先从正畸动画视频的每一帧中提取4个视角图像（包含正视图、双侧咬合图以及上牙弓或下牙弓颌面图）；之后以此作为条件输入，从固定视角合成上颌以及下颌模型的多视角渲染图像，再通过Neus网络根据这些多视角图片重建正畸中间过程的三维牙颌模型。尽管该方法能够借助预训练阶段引入的先验信息重建出三维牙颌模型（可视化效果如图4所示），但其重建结果与视频中的牙齿形态差距较大，且难以捕捉正畸移动序列中牙齿姿态的细微变化。

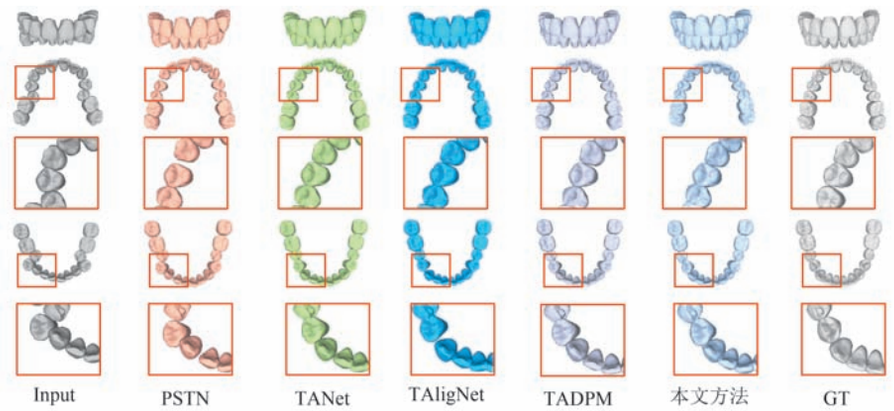


图3　正畸目标牙位预测效果的可视化比较

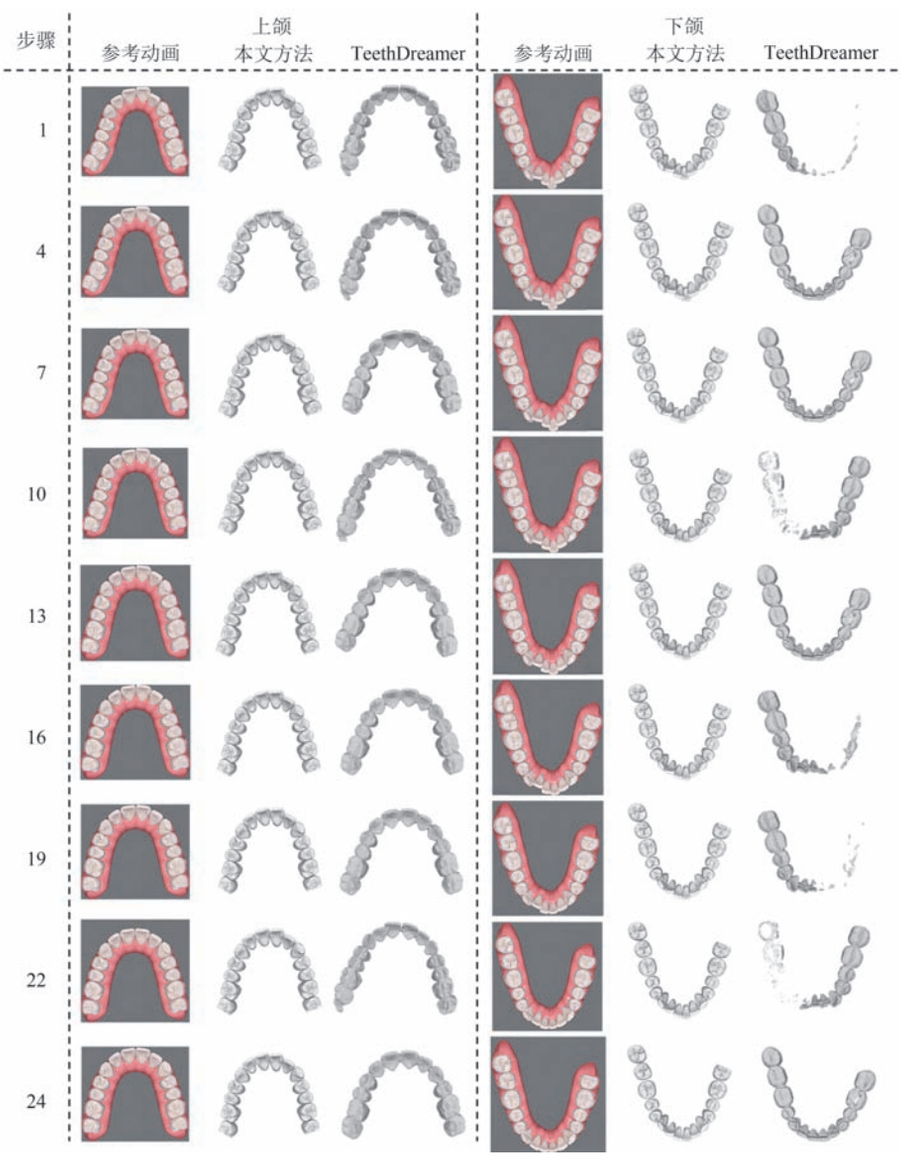


图4　正畸中间过程三维牙颌模型重建效果的可视化比较

根据TeethDreamer论文描述及其公开的实验结果，该方法主要在正常牙颌模型上表现出较好的重建效果。然而，当其应用于错颌畸形三维牙颌模型时，无论是使用论文作者公开的模型还是基于其方法进行复现，重建结果均不理想。本文分析认为主要原因有以下两方面：(1)对于错颌畸形三维牙颌

模型(存在缺牙、牙齿拥挤、牙弓形态异常),视频序列中部分牙齿存在严重遮挡问题,导致可用视角覆盖不足,限制了网络对三维几何结构的建模能力,造成几何一致性下降,并最终影响三维牙颌模型的整体重建精度。(2)TeethDreamer依赖4个固定视角的图像直接重建整体的三维牙颌模型,而在正畸移动序列中,不同帧的牙颌模型图像之间区别较小,使得网络难以精确重建正畸过程中牙齿逐步移动的细节信息。综上,TeethDreamer在错颌畸形场景下的泛化性明显不足,难以满足正畸任务中建模精细牙齿移动序列的需求。

相比于TeethDreamer,本文方法在建模思路上引入了更强的条件约束。具体来说,本文提出将正畸前牙颌模型的三维几何特征与视频帧中牙弓颌面投影的二维视觉特征共同作为扩散模型的条件输入,从而实现几何与视觉信息的联合建模。这种方式能够更全面地捕捉牙齿的位姿分布,尤其在牙齿存在遮挡的情况下,三维几何特征可以弥补视觉特征的缺失。其次,本文设计的$\mathcal{L}_{SSIM}$损失函数可以直接约束网络预测的中间过程三维牙颌模型渲染图像与视频帧之间的差异,从而为网络训练提供监督信号。从理论上来说,这些设计使本文的方法能够在建模过程中充分整合三维几何信息、视频运动信息和医学先验知识,因此在几何一致性和牙齿运动合理性上显著优于TeethDreamer。如图4所示,本文方法重建的牙齿形态与正畸动画视频中牙齿的形态特征高度吻合,不仅能够准确还原正畸过程中牙齿的位姿变化,还能有效保证正畸过程牙齿的大小与形态的一致性。

为进一步比较TeethDreamer和本文方法的重建效果,本节进行了定量比较。由于缺乏正畸中间过程的真实三维牙颌模型数据,本文提出一种基于图像一致性的定量评估方法。具体来说,将预测得到的三维牙颌模型渲染为图像,并与对应的正畸动画视频帧进行比较,通过计算常用的图像相似性评估指标来对重建质量进行定量分析,具体指标包括PSNR(Peak Signal-to-noise Ratio)、SSIM(Structural Similarity Index)、LPIPS(Learned Perceptual Image Patch Similarity)。实验结果如表2所示,其中GT表示真实正畸前三维牙颌模型的渲染图像与对应动画视频帧之间的基准误差,用于衡量PyTorch3D渲染结果与动画视频之间存在的固有差距。

表2 正畸中间过程三维牙颌模型重建效果比较

方法	PSNR(↑)	SSIM(↑)	LPIPS(↓)
GT	20.63	0.93	0.082
TeethDreamer	14.49	0.84	0.246
本文方法	**17.60**	**0.90**	**0.128**

从结果中可以看出,本文方法在所有指标中显著优于TeethDreamer,特别是在排除光照条件与材质属性影响的结构相似性指标SSIM中,其重建结果与GT参考值是相当的。上述实验结果表明本文方法在缺乏中间状态真实三维数据约束的情况下,依然能够准确重建出正畸过程中的三维牙颌模型,具备较好的泛化能力。

4.4 消融实验

本文方法在TADPM框架的基础上进行了多维度改进,通过整合正畸动画视频中的颌面图像作为条件信息,并引入图片的结构相似性损失来约束预测结果中的牙齿位姿,使神经网络能够预测与视频一致的正畸中间过程三维牙颌模型。为验证所提出策略的有效性,本节设计消融实验来分析阶段一的损失函数中引入$\mathcal{L}_{SSIM}$以及其不同设置所带来的影响,实验结果如表3所示。在表格中,基准组(No.1)指在阶段一的损失函数中不引入$\mathcal{L}_{SSIM}$,之后网络不进行微调、直接预测正畸中间过程的三维牙颌模型;实验组(No.2~5)在阶段一的损失函数中引入$\mathcal{L}_{SSIM}$,阶段二微调网络后再预测正畸中间过程的三维牙颌模型。其中,No.2和No.3分别测试了不同权重的$\mathcal{L}_{SSIM}$对实验结果的影响。No.2为本文方法的设置。

表3 消融实验结果

No.	$\mathcal{L}$	颌面图像背景颜色	PSNR(↑)	SSIM(↑)	LPIPS(↓)
1	$\mathcal{L}_{recon}+\mathcal{L}_{MSE}$	黑色	17.50	0.88	0.162
2	$\mathcal{L}_{SSIM}+\mathcal{L}_{recon}+\mathcal{L}_{MSE}$	黑色	**17.60**	**0.90**	**0.128**
3	$0.03\times\mathcal{L}_{SSIM}+\mathcal{L}_{recon}+\mathcal{L}_{MSE}$	黑色	17.60	0.88	0.160
4	$\mathcal{L}_{SSIM}+\mathcal{L}_{recon}+\mathcal{L}_{MSE}$	白色	17.60	0.84	0.153
5	$\mathcal{L}_{SSIM_{ss}}+\mathcal{L}_{recon}+\mathcal{L}_{MSE}$	黑色	17.40	0.87	0.168

实验结果表明，在阶段一的损失函数中引入$\mathcal{L}_{SSIM}$后虽然会降低目标位预测效果（表1中ADD指标下降8%），但表3中的实验组No. 2~4相较于No. 1均有所提升，表明引入$\mathcal{L}_{SSIM}$后可以提升在阶段二中正畸过程三维牙颌模型的预测效果，验证了本文方法对于中间过程位姿估计的有效性。在No. 3中仅将$\mathcal{L}_{SSIM}$的权重设置为0.03，其结果显示当$\mathcal{L}_{SSIM}$的权重过低时并不能有效改善阶段二的效果。通过对比分析，本文最终采用1:1:1的损失函数权重设置，以平衡阶段一和阶段二中的目标牙位预测与正畸过程三维牙颌模型预测两者的效果。

此外，本文在计算$\mathcal{L}_{SSIM}$损失函数时，统一将颌面图像以及预测结果渲染图像的背景处理为黑色。因此，本节在No. 4中分析颌面图像背景颜色对结果的影响，以及在No. 5中分析去除$\mathcal{L}_{SSIM}$中的亮度分量是否可以降低不同渲染方式带来的影响。在No. 4中，分析了颌面图像背景颜色对实验结果的影响。当背景设置为白色时，牙齿颜色与背景颜色较为相近，导致图片的结构相似性损失$\mathcal{L}_{SSIM}$无法有效地约束牙齿位姿，进而影响正畸中间过程三维牙颌模型的重建效果。而No. 5中的结果则显示了当背景颜色设置为黑色时，保留$\mathcal{L}_{SSIM}$中的亮度分量能够有效增强牙齿边界特征的表征能力。

图5展示了不同实验设置下模型预测结果的可视化对比结果，从红框中可以看出本文方法的设置能够更准确地捕捉颌面图像中的牙齿位姿信息，并且重建与之匹配的正畸过程三维牙颌模型。

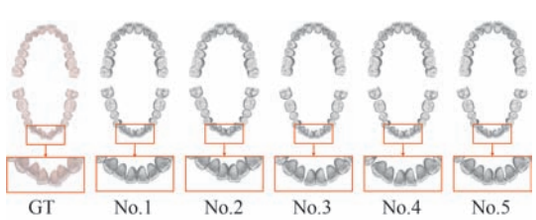


图5　消融实验结果的可视化比较

4.5　讨　论

4.5.1　生成结果的一致性讨论

条件扩散模型在生成过程中通常依赖随机噪声的逐步去噪，因此不同的噪声条件会导致生成结果存在一定的差异。本文采用在相关研究[8]中已被广泛验证的余弦（cosine）噪声调度策略，以保证生成结果的稳定性。为进一步评估生成结果的一致性，本文在相同输入条件下改变随机噪声初始化，多次生成正畸过程的三维牙颌模型。图6展示了在不同随机数种子下预测结果的渲染图像，结果表明生成的三维牙颌模型在整体形态与牙齿相对位置上均保持了较高一致性。

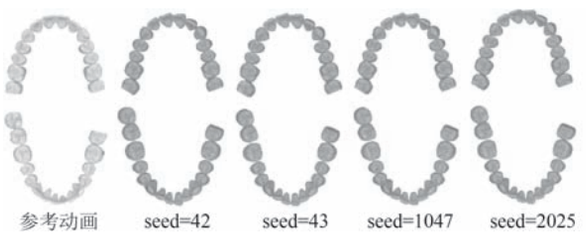


图6　一致性实验的可视化结果

此外，表4中计算了预测结果渲染图像与视频帧图像的相似性指标，其中结果的方差值（SD）非常小，表明随机噪声的变化仅对局部细节产生轻微影响，而不会改变牙齿的全局排布趋势。综上，本文方法生成的三维牙颌模型能够在不同噪声条件下保持良好的几何一致性与稳定性。

表4　一致性实验的定量结果

随机数种子	42	43	1047	2025	AVG	SD
PSNR(↑)	17.65	17.60	17.69	17.58	17.63	0.0497
SSIM (↑)	0.91	0.90	0.90	0.91	0.905	0.0058
LPIPS(↓)	0.123	0.128	0.122	0.125	0.125	0.0026

4.5.2　生成结果的医学合理性讨论

本文所使用的数据集中，每组样本的正畸目标位以及正畸过程动画视频均由临床正畸医生和技师共同设计。因此，数据集本身已在一定程度上纳入了牙齿碰撞、牙齿生理移动范围等正畸医学相关因素。当网络预测结果的渲染图像逐渐逼近视频帧时，重建得到的三维牙颌模型也能够在一定程度上避免牙齿碰撞，并保证每一次牙齿的移动处于合理生理范围内。同时，本文提出将视频帧与初始状态三维牙颌模型的特征编码共同作为条件信息，引导扩散模型学习视频帧中牙齿位姿的分布，可以有效捕捉牙齿之间的相对位置关系，使重建得到的正畸过程三维牙颌模型移动序列符合正畸医学要求。

然而，由于动画视频的分辨率较低，且网络训练过程中缺乏真实三维牙颌模型作为监督信号，导致当前重建得到的三维牙颌模型移动序列尚不足以直接应用于临床隐形牙套的生产。未来的研究中将进一步引入医学先验知识，结合临床正畸医生的评估来构建显示约束，从而严格确保预测结果中牙齿无碰撞，提高预测结果的临床可用性与可靠性。

6 结 论

本文提出一种动画视频引导的正畸过程三维牙颌模型重建方法，通过融合颌面投影图像以及正畸前三维牙颌模型的多模态特征信息作为条件，引导扩散模型生成正畸过程以及正畸后牙齿的变换矩阵，从而构建三维牙颌模型正畸过程的完整移动序列。针对真实三维牙颌模型移动序列数据缺失的挑战，本文创新性地构建了基于可微分渲染的视觉一致性损失函数，建立起二维投影与三维几何空间的弱监督关系。通过学习视频帧中牙齿位姿的分析，本文方法可以有效捕捉牙齿之间的相对位置关系，从而在一定程度上避免碰撞的发生。未来工作将进一步引入医学先验知识以及临床正畸医生的评估来构建显示约束，从而提升预测结果的合理性，为智能化正畸方案生成提供数据支撑。同时，本文方法也为基于口内视频重建三维牙颌模型，实现远程正畸诊疗奠定了坚实的技术基础。

参 考 文 献

[1] Ke Y, Zhu Y, Zhu M. A comparison of treatment effectiveness between clear aligner and fixed appliance therapies. BMC Oral Health, 2019, 19: 1-10

[2] Li Z, Yang G. Research on simulation and optimization method for tooth movement in virtual orthodontics//Proceedings of International Conference on Computer Science, Environment, Ecoinformatics, and Education. Wuhan, China, 2011: 270-275

[3] Ma T, Lyu J, Yang Q, et al. Orthodontic overcorrection scheme generation based on improved multkliparticle swarm optimization. Journal of Healthcare Engineering, 2021(1): 3624515

[4] Li Z, Liu T, Li H A, et al. Orthodontic path planning method based on optimized artificial bee colony algorithm//Proceedings of the 5th International Conference on Intelligent Computing and Signal Processing. Suzhou, China, 2020: 012017

[5] Wei G, Cui Z, Liu Y, et al. Tanet: Towards fully automatic tooth arrangement//Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020: 481-497

[6] Li X, Bi L, Kim J, et al. Malocclusion treatment planning via pointnet based spatial transformation network//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Lima, Peru, 2020: 105-114

[7] Lingchen Y, Zefeng S H I, Yiqian W, et al. iorthopredictor: model-guided deep prediction of teeth alignment. ACM Transactions on Graphics, 2020, 39(6): 216

[8] Lei C, Xia M, Wang S, et al. Automatic tooth arrangement with joint features of point and mesh representations via diffusion probabilistic models. Computer Aided Geometric Design, 2024, 111: 102293

[9] Fan Y, Wei G, Wang C, et al. Collaborative tooth motion diffusion model in digital orthodontics//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 1679-1687

[10] Ma J, Lou J, Jiang B, et al. Neural orthodontic staging: predicting teeth movements with a transformer. IEEE Transactions on Visualization and Computer Graphics, 2024, 31(9): 6253-6267

[11] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851

[12] Wang C, Wei G, Wei G, et al. Tooth alignment network based on landmark constraints and hierarchical graph structure. IEEE Transactions on Visualization and Computer Graphics, 2022, 30(2): 1457-1469

[13] Schönberger J L, Zheng E, Frahm J M, et al. Pixelwise view selection for unstructured multi-view stereo// Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 501-518

[14] Wang P, Liu L, Liu Y, et al. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction// Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual, 2021: 27171-27183

[15] Wang Y, Han Q, Habermann M, et al. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction// Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 3295-3306

[16] Li Z, Müller T, Evans A, et al. Neuralangelo: High-fidelity neural surface reconstruction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 8456-8465

[17] Darmon F, Bascle B, Devaux J C, et al. Improving neural implicit surfaces geometry with patch warping//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 6260-6269

[18] Schonberger J L, Frahm J M. Structure-from-motion revisited//Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4104-4113

[19] Chen Y, Gao S, Tu P, et al. Automatic 3d teeth reconstruction from five intra-oral photos using parametric teeth model. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(8): 4780-4791

[20] Liu R, Wu R, Van Hoorick B, et al. Zero-1-to-3: Zero-shot one image to 3d object//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 9298-9309

[21] Shi R, Chen H, Zhang Z, et al. Zero123++: a single image to consistent multi-view diffusion base model. arXiv preprint arXiv: 2310.15110, 2023

[22] Long X, Guo Y C, Lin C, et al. Wonder3d: Single image to 3d using cross-domain diffusion//Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 9970-9980

[23] Liu Y, Lin C, Zeng Z, et al. Syncdreamer: Generating multiview-consistent images from a single-view image. arXiv preprint arXiv:2309.03453, 2023

[24] Liu M, Xu C, Jin H, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. Advances in Neural Information Processing Systems, 2023, 36: 22226-22246

[25] Liu M, Shi R, Chen L, et al. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2024: 10072-10083

[26] Xu C, Liu Z, Liu Y, et al. Teethdreamer: 3D teeth reconstruction from five intra-oral photographs//Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention. Marrakesh, Morocco, 2024: 712-721

[27] Song Y, Sohl-Dickstein J, Kingma D P, et al. Score-based generative modeling through stochastic differential equations//Proceedings of the 9th International Conference on Learning Representations. Vienna, Austria, 2021: n.p

[28] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics//Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: 2256-2265

[29] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 10684-10695

[30] Brooks T, Holynski A, Efros A A. Instructpix2pix: Learning to follow image editing instructions//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 18392-18402

[31] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 3836-3847

[32] Peebles W, Xie S. Scalable diffusion models with transformers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 4195-4205

[33] Zhang J, Huang Q, Liu J, et al. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 23464-23473

[34] Jeong J, Han S, Kim J, et al. Latent space super-resolution for higher-resolution image generation with diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 2355-2365

[35] Xia M, Zhou Y, Yi R, et al. A diffusion model translator for efficient image-to-image translation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 10272-10283

[36] Zhu D, Di Y, Gavranovic S, et al. Sealion: Semantic part-aware latent point diffusion models for 3d generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 11789-11798

[37] Yang Y, Shao J, Li X, et al. Prometheus: 3d-aware latent diffusion models for feed-forward text-to-3d scene generation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 2857-2869

[38] Kara O, Singh K K, Liu F, et al. Shotadapter: Text-to-multi-shot video generation with diffusion model//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 28405-28415

[39] Pallotta E, Azar S M, Li S, et al. Syncvp: Joint diffusion for synchronous multi-modal video prediction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2025: 13787-13797

[40] Qi C R, Yi L, Su H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems, 2017, 30: 5105-5114

[41] Liang Y, Zhao S, Yu B, et al. Meshmae: Masked autoencoders for 3d mesh data analysis//Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 37-54

[42] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 16000-16009

[43] Kirillov A, Mintun E, Ravi N, et al. Segment anything//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France, 2023: 4015-4026

[44] Bao F, Nie S, Xue K, et al. All are worth words: A vit backbone for diffusion models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 22669-22679

[45] Besl P J, McKay N D. A method for registration of 3d shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(2): 239-256

[46] Song J, Meng C, Ermon S. Denoising diffusion implicit models//Proceedings of the 9th International Conference on Learning Representations (ICLR 2021). Vienna, Austria, 2021: n.p

[47] Yao Y, Luo Z, Li S, et al. Mvsnet: Depth inference for unstructured multi-view stereo//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 767-783

LIANG Ya-Qian, Ph. D. Her research interests include computer graphics, 3D mesh geometric analysis and deep learning.

YOU Jing-Yan, B. E. candidate. His main research interests include deep learning.

LEI Chang-Song, Ph. D. candidate. His main research interests include 3D geometric analysis and deep learning.

FAN Yi-Ming, B. E. candidate. Her main research interests include deep learning.

DAI Jia-Jia, Ph. D. Her research interests include computer graphics, 3D mesh geometric analysis and deep learning.

FAN Ye-Ying, Ph. D. Her research interests include 3D model processing and medical image processing.

WANG Shao-Feng, Ph. D. His main research interests include orthodontics and AI assisted orthodontic diagnosis and treatment.

BAI Yu-Xing, Ph. D. professor. His main research interests include orthodontics and AI assisted orthodontic diagnosis and treatment.

LIU Yong-Jin, Ph. D. professor. His main research interests include computer graphics and visual media computing.

Background

This paper belongs to a topic in the field of artificial intelligence-assisted orthodontics. Although numerous studies have explored the use of deep learning methods to predict tooth alignment and orthodontic processes, there is currently no publicly available dataset containing orthodontic movement sequences of 3D dental models. In clinical practice, obtaining such 3D orthodontic sequences is also highly challenging, which significantly hinders progress in intelligent orthodontics research.

To address this limitation, this paper proposes an animated video-guided method for reconstructing 3D dental models during the orthodontic process. By leveraging a diffusion model to learn the distribution of tooth poses from occlusal views, the proposed method generates 3D dental models corresponding to each frame of the orthodontic animation video. This paper offers not only a potential source of synthetic training data for related tasks but also a novel direction for data generation in the field, thereby facilitating further development of intelligent orthodontic technologies. Furthermore, this paper lays a solid technical foundation for remote diagnosis and treatment by enabling 3D dental model reconstruction from intraoral video data.